Meaning and compositionality as statistical induction of categories and constraints

by

Lauren A. Schmidt

B.S., Symbolic Systems, Stanford University, 2001
M.S., Computer Science, Stanford University, 2002

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2009

Author ....
Department of Brain and Cognitive Sciences
June 29, 2009

Certified by ....
Joshua Tenenbaum
Associate Professor of Cognitive Science
Thesis Supervisor

Accepted by ....
Earl Miller
Professor of Neuroscience
Chairman, Committee for Graduate Students

Meaning and compositionality as statistical induction of categories and constraints

by

Lauren A. Schmidt

Submitted to the Department of Brain and Cognitive Sciences
on June 29, 2009, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

What do words and phrases mean? How do we infer their meaning in a given context? How do we know which sets of words have sensible meanings when combined, as opposed to being nonsense? As language learners and speakers, we can solve these problems starting at a young age, but as scientists, our understanding of these processes is limited.

This thesis seeks to address these questions using a computational approach. Bayesian modeling provides a method of combining categories and logical constraints with probabilistic inference, yielding word and phrase meanings that involve graded category memberships and are governed by probabilistically inferred structures. The Bayesian approach also allows an investigation to separately identify the prior beliefs a language user brings to a particular situation involving meaning-based inference (e.g., learning a word meaning or identifying which objects an adjective applies to within a given context), and to identify what the language user can infer from the context. This approach therefore provides the foundation also for investigations of how different prior beliefs affect what a language user infers in a given situation, and how prior beliefs can develop over time.

Using a computational approach, I address the following questions: (1) How do people generalize about a word's meaning from limited evidence? (2) How do people understand and use phrases, particularly when some of the words in those phrases depend on context for interpretation? (3) How do people know and learn which combinations of predicates and noun phrases can sensibly be combined and which are nonsensical? I show how each of these topics involves the probabilistic induction of categories, and I examine the constraints on inference in each domain. I also explore which of these constraints may themselves be learned.

Thesis Supervisor: Joshua Tenenbaum
Title: Associate Professor of Cognitive Science

Acknowledgements

Josh Tenenbaum has been an advisor to me since before I got to MIT. I started working with him at Stanford, where he got me excited about the possibilities of computational modeling and helped me transform my questions about word learning from vague ideas to a coherent project. That's been my experience throughout grad school as well: Josh is wonderful at taking interests and questions that are overly vague and helping to formulate them into excellent scientific approaches. He has taught me a great deal about how to do good scientific work as well as how to do computational modeling in particular. In addition, he has inspired me and kept me going at times when I thought I could not make it through a particular project, or through grad school itself. Josh has been, in turns, teacher, mentor, and cheerleader — as well as a leader of wonderful lab retreats where he lead us all in playing games, singing songs, and going on possibly overly ambitious hikes. He has been a friend as well as an advisor, and I value his enthusiasm and his ability to have fun (and his encouragement to do so) as well as the academic role he has played in my career. His lab has been a wonderful place to learn and also to have a good time in great company — I will miss the lab and the yearly retreats.

Ted Gibson has also been a great source of advice and help over the years. Ted's lab provided a wonderful second home within the department, and Ted has helped me both with research-related questions and with negotiating grad school in general. Ted has also encouraged me greatly in thinking about teaching as a career. At this point in time, I'm not sure if that's what my future holds, but I very much appreciate his encouragement about both my teaching and my research, and I am glad to know that he thinks highly of my teaching skills. In general, Ted has always made me realize and feel proud of my accomplishments. It is no exaggeration to say that I would not have made it through grad school without both him and Josh helping to put my achievements in perspective from time to time, as well as advising me on how best to make further progress.

Molly Potter has offered insightful advice about my research throughout my grad school career. She asks excellent questions, and in particular, she has made me think a great deal about the broader implications of my work, and how to frame it within the context of other work in the field of cognitive science. My work is far better for it. I would also like to thank Andrew Shtulman, Jesse Snedeker, Neal Pearlmutter, and Mike Collins for valuable insights about my work over the years.

Lera Boroditsky was my undergraduate advisor and an advisor to me during the beginning of my grad school career. Lera helped me turn my passion for psychology, language, and cognitive science into a coherent program of research. Until I met her, I didn't think I was going to do research, but she showed me how to have fun doing so — and how to do so well. She also gave me invaluable advice on presenting my work in an understandable and interesting format — and always provided a wonderful example of how to do so. She greatly helped my development as a cognitive scientist and encouraged me to go to grad school and to ask interesting questions.

Dave Barner has been a great help with the research reported in Chapter 3. Dave has been tireless in proposing research approaches and discussing the implications of our work. He has also done wonderful work collecting data from both children and adults. He has additionally been extremely helpful in writing up this work. And, in general, he has been a friend and offered a lot of encouragement to me with all of my research.

A number of other people have provided invaluable support to me while I've been at MIT. Noah Goodman did a great deal of work towards developing the research in Chapter 3, and was terrific at answering questions and helping to present the work formally and rigorously. Much of the model section in Chapter 3 and the related section of Appendix A was written with the help of Noah (all mistakes, however, are mine). Thanks to Noah also for general advice and support about both the adjectives project and grad school more broadly.

Charles Kemp helped tremendously with the work presented in Chapter 4, and is largely responsible for helping to work out the details of the M constraint model — and to write about the model coherently. The equations in the section of Appendix A related to the M constraint were formulated with his extensive guidance and help (again, however, all mistakes are mine). Charles is an incredibly patient and good teacher, and I learned a lot about Bayesian modeling from working with him. In addition, he has been an excellent friend over the years. I have many fond memories of debates and discussions with him, and traveling around San Diego with him and visiting him in Pittsburgh were highlights of the past few years.

Liz Baraff Bonawitz gave me extremely helpful and thorough feedback on Chapter 3 of this thesis. She was also my first local friend when I moved to Boston, and our friendship has grown over the years. I am so grateful to her for being such a good friend, as well as for offering lots of great advice on negotiating academia over the years. I am very much looking forward to continuing to spend time together in California. Also in Berkeley is Tom Griffiths, who has offered his friendship and mentorship ever since we were officemates at Stanford, and Tania Lombrozo, a friend who has offered a number of useful research insights and also helped me figure out how to make it through grad school while retaining my sanity. I can't wait to live closer to them again.

The other members and affiliates of the Cocosci lab, past and present, have also given me lots of good feedback and offered friendship over the years. In particular, I would like to thank Pat Shafto, Mike Frank (another long-time friend who has given so much good feedback over the years), Keith Bonawitz, Vikash Mansinghka, Dan Roy, Virginia Savova, Ed Vul, Tim O'Donnell, and Steve Piantadosi. I would also like to thank Mara Breen, Ev Fedorenko, Michelle Greene, Barbara Hidalgo-Sotelo, Retsina Meyer, and John Kraemer for making the department such a friendly and fun place to be. Frequent departmental visitor Celeste Kidd has also been a terrific friend and source of ideas and support.

I've saved Amy Perfors for last out of the MIT-affiliated people I wish to thank, because she has been very much a part of my life outside of MIT as well as in the department. I was lucky enough to live with Amy as well as share an office with her for most of my time at MIT, and I could not have asked for a better roommate, officemate, or friend. I certainly would not have made it through grad school without Amy's support and friendship, but my life would also have been much poorer without it.

I was fortunate enough to have another fabulous roommate after Amy moved away: Christina Jenkins. I would like to thank Chris for her friendship and also her patience with me during the process of writing my dissertation. I'm very excited that Chris is also moving to California. But I will be sad to leave behind a number of other people who have been terrific friends during my time on the East Coast. In particular, I would like to thank Hilary Mason, Michael Stewart, and Miler Lee for wonderful Thanksgiving celebrations — and thank you to all of them for friendship, support, and visits at other times of the year as well. Thanks to Danforth Nicholas for a lot of support and smiles over the years, and for talking with me about various psychology and neuroscience topics, often helping to remind me, when I was frustrated, why I was interested in the field in the first place. Thanks also to Rosa Carson, Darth May, and Aaron Mandel for their friendship.

On the west coast, I would like to thank Lee Abuabara for her friendship, empathy, and help in figuring out how to get through the difficult parts of grad school. We made it! I would also like to thank the household of HALFLAB (Lisa Eckstein, NeilFred Piccioto, Louise McFarlane, and Brandon Nutter) for letting me visit so frequently and occupy their guest room for so long. All of them are terrific friends, whom I've missed a great deal (when I haven't been staying with them). In particular, I would like to thank Lisa, not just for being a great friend, but also for getting me to do more writing. That experience has been valuable not only in writing fiction, but also in writing academic papers; it turns out lots of the lessons I learned transfer, and it is in part thanks to Lisa that I have managed to write a dissertation as well as a novel in the past several years.

I would also like to thank Brandon for his endless support over the years, and for selflessly encouraging me to pursue and finish grad school even though it meant being far away. Brandon has offered so many keen insights and ideas over the years; he is, in many ways, the best thinker I have ever known, and I have valued his input — about my work and everything else — tremendously. He has also provided so much emotional support. He has known what to say to keep me going through the toughest of times, and to make me laugh at all times. I couldn't have made it without him.

I am also very lucky to have wonderful and supportive parents, Jill and Dietger Schmidt. They have always provided inspiration, comfort, and absolute belief in my capability to do whatever I set my mind to. They have also, along with my brother Alex, provided an excellent home in Washington to come home to whenever I have the time. My mother has provided many creative insights over the years that have influenced how I think about and how I present my work. My father got me interested in computer science long ago, and always has great questions and insights about technical matters, as well as a broad range of other topics. More than that, though, my family members provide friendship, support, conversation, and a sense of home and love that I will always be grateful for, and always be glad to return to, wherever I am living.

A few miscellaneous thanks remain. First, I would like to thank Jan Butterfield for making me think about the real-world applications of my work and of science in general. I am sorry I acted like the question was an affront at one point long ago; I do believe that pure science is important, but I have come to believe that the world is a better place if we also consider both the consequences and the possible applications of our work. I would like to thank the NaNoWriMo community in Boston and the Bay Area (and online) for providing incentive to write and environments in which to do so. I would like to thank my many friends on Web-based communities — especially LiveJournal, Brunchma, and WomBaT — who helped me pilot various parts of my research. I would also like to thank the community members of BeaversOnCrack, a productivity group that helped me get a lot done while doing research and while writing my thesis. I am very lucky to have a large network of very supportive friends and family, and I would also like to say thank you to everyone who has been a friend to me over the past six years, even if I did not mention you by name here. I am blessed and inspired to have such good people in my life.

Contents

Chapter 1

Meaning and meaningfulness

How do we infer the meanings of words and phrases? How do we know if word combinations are meaningful at all? In this thesis, I address such topics of meaning and meaningfulness, presenting a computational approach to address specific instances of these questions. In particular, I seek to take substantial steps forward in addressing the following three broad issues of meaning:

- How do we learn the meaning of a new word based on limited evidence?
- How do we compose the meanings of individual words in order to interpret a larger phrase?
- How do we know which word combinations are sensible and which are nonsense?

Within each of these broad topics, I ask two kinds of questions. First, how do we judge meaning or sensibleness within a specific context? Second, how do our judgments change over time? I present modeling accounts that allow investigation of both of these questions within a unified framework.

Dichotomies of cognitive science

When cognitive scientists investigate meaning and sensibility, the following questions often arise: First, do we make judgments about meaning and sensibility through statistical methods, or through logical constraints and categorization? Second, when solving learning and reasoning about meaning and compositionality, do people rely on nature (innate constraints and beliefs) or nurture (experience with the world)? Here I discuss each of these dichotomies further, and how this thesis seeks a unified account of both divisions.

Logic vs. statistics

Cognitive science brings together many fields, each with different approaches to understanding word meaning, word learning, and compositionality. Classically, linguistics and philosophy address semantics as a matter of formal sets and logic; many words and combinations of words refer to sets of items in the world (noun phrases) or to predicates with truth values (e.g., adjectives), and truth is determined by reasoning about logical entailment (see Chierchia & McConnell-Ginet, 1990 for an overview). Psychology also often addresses word-meaning and -learning in terms of categories and logical constraints, but the field frequently deals with empirical statistics as well: How good of a member of category Y is object X (Rosch & Mervis, 1975)? How many examples does a child get of positive vs. negative evidence about word meaning (Chouinard & Clark, 2003)? Computational linguistics and natural language processing – as well as many connectionist modeling approaches by psychologists – often seek to understand word meanings based on the contexts where those words have appeared, and to similarly predict which words can combine together based on previously seen combinations (e.g., Manning & Schütze, 1999; Justeson & Katz, 1995).

This thesis seeks to bring together the best of all these cognitive science approaches. I will look at meaning and compositionality as the probabilistic induction of categories and constraints on categories. This combination predicts both people's graded judgments about word and phrase meaning, and also their ability to infer complex structures from limited and/or noisy data.

Prior constraints vs. learning from the data

Rarely is anything that humans do a matter of purely nature or purely nurture. Still, cognitive science is full of debates as to whether specific cognitive structures or learning constraints are innate or learned. For instance, people have a tendency to generalize most nouns to items of the same shape. Must such a shape bias be innate, or could it be learned (see, e.g., Kemp, Perfors, & Tenenbaum, 2007)? Likewise, it is often unclear what learners are able to infer from a given context and what cognitive tools and assumptions they come pre-equipped with.

Throughout this thesis, I examine what knowledge and constraints language users bring to a given problem, and what they learn based on a given set of evidence. I examine how different constraints and parameters make a difference to people's judgments and to their developmental trajectories. Through Bayesian modeling, I show how we can pick out how different cognitive tools are influenced by different types of evidence. We can therefore transform overly simplistic nature/nurture types of distinctions into a more useful question of how prior knowledge and new evidence are combined to form new representations, which can then serve as a basis for future inference.

Bridging dichotomies

In my thesis work, I seek to bridge these artificial dichotomies. I show how Bayesian modeling can provide probabilistic categorization judgments that match people's intuitions about the meaning of words and phrases, presenting a more complete picture than either simple statistical methods or strict categorization. Bayesian models also allow us to combine prior beliefs and constraints with evidence presented by the data. They provide insight into how much a rational learner can conclude based on experience (or the data in a given situation), and with what degree of confidence – all without reducing the question to the overly simple nature vs. nurture duality.

A Bayesian approach to meaning

All of the problems addressed in this thesis involve problems of induction – guessing the true underlying state of the world from limited observations. The Bayesian modeling framework allows us to ask what a rational agent (one who reasons optimally from a given set of information) would infer about underlying structures given a set of observed data. This section explains how the Bayesian approach works.

A Bayesian model specifies two components that go into such reasoning. First, the rational agent's prior beliefs and knowledge – what are the hypotheses that the agent considers to be possible explanations of the situation at hand, and how strongly does the agent favor each of these possibilities *a priori*? Specific constraints or biases are often incorporated into these prior beliefs; Occam's razor, for instance, can be incorporated as a higher initial probability for simpler hypotheses (so long as "simplicity" is formalized in some quantifiable way). This component of the Bayesian framework assigns a *prior probability* to every hypothesis, h.

The second component of the Bayesian framework specifies the *likelihood* of a set of data, D, being observed given a particular hypothesis. Specifying a Bayesian model entails specifying how data are generated from the underlying true hypothesis.

These components are combined in Bayes' rule:

$$p(h|D) = \frac{p(D|h)p(h)}{p(D)} \tag{1.1}$$

This equation states that the *posterior probability* (the probability of a particular hypothesis given the observed data) is proportional to the product of the prior probability and the likelihood. This value is divided by the overall probability of the data set, capturing the idea that a data set that is very likely to arise across lots of different hypotheses is not very informative about a given hypothesis. When comparing how likely multiple hypotheses are based on the same set of data, we can ignore the denominator, $P(D)$ — the probability of the data will be the same in all cases if the data never changes. In such cases, Bayes' rule is then simply formulated as:

$$p(h|D) \propto p(D|h)p(h) \tag{1.2}$$

An illustrative example

To make all this more concrete, imagine that you are playing a game with your friend, Amy. There are two boxes of books behind Amy, one containing 10 science fiction books and nothing else, and one containing 5 science fiction books and 5 chick lit books. Amy is going to select a box of books and pull a book from it. You then have a chance to guess what proportion of sci-fi the box contains. Or you can opt to ask to see more books before you guess. If you are right when you do guess, you get to gloat about your fine reasoning skills, but if you are wrong, Amy will mock you mercilessly. Therefore, you do your best to figure out what box she has based on the evidence you see.

Amy selects a box. Before she draws any books from it, you have two hypotheses about which box it might be. If you believe that she will select a box at random, then your prior probability distribution over the *hypothesis space*, or set of possible hypotheses, is the following:

$$\begin{aligned} P(S) &= \frac{1}{2} \\ P(M) &= \frac{1}{2} \end{aligned}$$

Where the hypotheses are S, the box contains only sci-fi, and M, the box contains a mix of sci-fi and chick lit. However, perhaps you know that the sci-fi books are substantially heavier than the (light, fluffy) chick lit, and you believe that Amy is a silly rugby player who likes to show off her muscles by dragging heavy things around. If you believe it is twice as probable that she will select the heavier of the two boxes, your set of prior probabilities is:

$$P(S) = \frac{2}{3}$$
$$P(M) = \frac{1}{3}$$

Now Amy draws a book from the box. It is a book by Sheri Tepper (science fiction). This book could have been drawn from either of the boxes. However, if Amy has drawn the book at random from the boxes, then the boxes each have different probabilities of producing a science fiction book. S, the box containing only sci-fi, will always produce a science fiction book. M will produce a sci-fi book during half the random draws. Therefore, the likelihood of the data you have seen, D, is the following given each hypothesis:

$$P(D|S) = 1$$
$$P(D|M) = \frac{1}{2}$$

On the other hand, perhaps you think that Amy is not selecting randomly, and instead is probably making this game as hard for you as possible. Then you believe that she would select a sci-fi book first no matter which box was in front of her, because that is maximally confusing:

$$P(D|S) = 1$$
$$P(D|M) = 1$$

Let's say that the first set of priors and the first set of likelihoods represent your beliefs – you think that Amy does not care about the weight of the box and truly selects the box at random, and you think she is playing a fair game and selecting the book at random. Then your posterior beliefs are:

$$\begin{aligned} P(S|D) &\propto P(D|S)P(S) = 1 * \tfrac{1}{2} \\ &= \tfrac{1}{2} \\ P(M|D) &\propto P(D|M)P(M) = \tfrac{1}{2} * \tfrac{1}{2} \\ &= \tfrac{1}{4} \end{aligned}$$

In other words, once you have seen a single sci-fi book, the probability that Amy has S, the sci-fi only box, is twice as high as the probability that she has M, the mixed box. This example illustrates how initial beliefs are reflected in the prior probability, how two different models (the books are randomly selected vs. the books are selected by a devious opponent) affect the likelihood, and how prior probabilities and likelihoods give the evidence can be combined to yield posterior probabilities.

More complicated generative models

We have now seen how to model a simple inference problem using a Bayesian model. The potential for applying these models ranges far beyond simple guessing games. For instance, suppose we were guessing about who was friends with whom on Facebook out of a set of 100 people based on the messages we saw on different Facebook members' walls. We could use Bayesian modeling to describe a hypothesis space of all possible graphs of connections between the 100 people. And we could create a more complex model of how data is generated from the hypothesis space, as well. For instance, perhaps we realize that some people are not very talkative and are less likely to write messages on the walls of their friends than other people are – so a lack of messages between two talkative people is more informative about their friendship status than a lack of messages between two quiet people. As long as we can describe how all the parameters influence the prior probability and the likelihood, we can continue to invert the generative model and figure out the most likely underlying state of the world.

Advantages of the Bayesian framework

The Bayesian framework provides a clear method for specifying initial beliefs about a situation, and incorporating new information. Both the priors and the likelihood can be varied independently, and the effect of either factor on the posterior probability is easily observed. Bayesian modeling makes it easy to address questions of what can be induced from a data set given a set of initial constraints. This makes Bayesian modeling a powerful tool for modeling cognitive phenomena relatively transparently. With other modeling methods, such as connectionist networks with hidden layers or recurrent loops (e.g., Dilkina, McClelland, & Boroditsky, 2007, Li & MacWhinney, 1996; Elman, 1990), it is not always clear how specific factors in the network setup correspond to prior beliefs, or how the factors in the network setup and the input data are connected to the probabilities of the hypotheses end up in their final state.

Additionally, many models of cognition require both positive and negative feedback in order to make inferences. Negative evidence consists of examples of things that are *not* generated by the hypothesis. For instance, say your friend Brandon is thinking of a category of numbers, and you're trying to guess what the category is. If he tells you that the category contains 1, 7, and 13, he has given you only positive evidence. If Brandon also tells you that the category does not contain the number 9, he has just given you a piece of negative evidence. Some connectionist networks and similarity based models require negative evidence in order to distinguish between overlapping hypotheses such as PRIME NUMBERS and ODD NUMBERS; without a negative example like "9 is not in the category", the probability of the PRIME NUMBER hypothesis will never decrease so long as all examples seen are members of both categories (see further discussion of the need for negative evidence in Chapter 2). However, a Bayesian model can take into account expectations about how numbers are sampled from the category. Therefore, a Bayesian learner can become more and more certain that Brandon is thinking or PRIME NUMBERS and not ODD NUMBERS as more positive examples that fit both categories become available (because it would otherwise be a suspicious coincidence that all the examples were prime — see Griffiths & Tenenbaum, 2007)[1]. Many problems of induction that people face in language acquisition and meaning inference are problems of learning from positive evidence only. The Bayesian framework allows us to model inference in these situations.

The Bayesian approach also provides an ideal framework for combining statistics and logic. Logic and sets play multiple roles in Bayesian reasoning: we can consider particular hypotheses, each of which is strictly true or false (e.g., "Amy has the mixed box in front of her"); we can reason about categories which contain no uncertainty (e.g., 1, 7, and 13 are prime numbers by definition); and we can apply logical constraints or biases in our reasoning (e.g., "simpler hypotheses are better"). The approach also incorporates probabilistic reasoning, however. We can take two logical possibilities ("the sci-fi book came from the sci-fi only box"; "the sci-fi book came from the mixed box") and quantify our uncertainty about them. We can reason about uncertainty over sets to yield graded category membership judgments. And we can turn a preference for simplicity into a quantitative factor affecting our prior beliefs. Thus, given a particular set of evidence, we can reason backwards about the most probable underlying category or set of rules, but we can quantify our uncertainty about it statistically.

The computational modeling approach to cognitive science

The Bayesian approach addresses questions at the *computational level*, as defined by Marr (1982). This level of description allows us to define the nature of the problem being solved, and to examine how people solve it at a high level – what beliefs and reasoning constraints do they bring to the problem? How does new evidence change their beliefs? This level of description is in contrast to the *algorithmic level* and the *implementational level.*

[1] Of course, if you think Brandon is trying to trick you, you may not have the same beliefs about how likely the examples are to have been generated from PRIME NUMBERS rather than ODD NUMBERS — but as with the example of Amy and the boxes, the model can accommodate different sampling assumptions, and can learn from positive evidence as reasonable.

An algorithm is a specific approach to solving a problem that specifies the particular representations used and steps taken; for instance, solving the problem $x = 2{+}4{+}1$ can be solved using an algorithm that adds the numbers on the left first, or the right first. The implementational level describes how things are implemented in the brain of the human reasoner (or in the circuitry or other physical implementation of a non-biological problem solver).

The computational models of cognition presented by this thesis could be implemented in many ways at both the physical level and the algorithmic level. Both a human and a computer could represent prior beliefs and reason based on new evidence to arrive at the same answer, according to our models, but they could do so in different ways at both the implementational and the algorithmic level. At the algorithmic level, it is likely that human reasoners are not calculating exact probabilities for all possible hypotheses when reasoning in most situations; they are probably using heuristics and simplified representations. This thesis does not commit to any particular process for heuristically approximating the computational models described within, nor to any given neural realization of the models. Instead, the thesis offers a method for understanding how these problems of meaning can be solved and what particular constraints and biases humans bring to the table when solving them. For these purposes, the computational level offers the ideal approach.

Overview

In each of the following chapters, I apply the Bayesian approach to answering questions of meaning and semantic compositionality in a way that combines the logic of rules and sets with the gradedness of statistics.

Chapter 2 of my thesis work addresses how people learn the meanings of words (specifically, nouns) based on limited evidence and how our inference process differs in a novel vs. a familiar domain. I also examine how our hypothesis space about possible word meanings develops.

In Chapter 3, I address the issue of compositionality in meaning. Specifically, I investigate how we interpret noun phrases that involve gradable adjectives such as *tall*, when the meaning of such an adjective depends on what it is modifying. I discuss how to use this framework for modeling compositional meanings more broadly.

In Chapter 4, I delve further into compositionality to ask which combinations of predicates and objects yield sensible combinations. I examine the development of judgments about sense and nonsense, and examine whether the learning constraint that allows us to make such judgments can itself be learned.

In Chapter 5, I discuss the overall implications of this work.

Chapter 2

Learning word meanings

How do we learn words when we're first encountering novel objects and categories? What possible word meanings do we entertain, and how do we evaluate those hypotheses based on the evidence we see? Xu and Tenenbaum (2007) (henceforth XT) presented a model of word learning that showed how people generalize about new word meanings from a given set of evidence. This model accounted for a number of phenomena that other word learning models could not. However, the experiments were run with familiar objects that already had several known category labels. It is possible that people in these experiments were considering new meanings for words based on a hypothesis space that was strongly influenced by their existing familiarity with the category and their English labels (for instance, perhaps they were only considering categories that already were named in English as possible meanings for the new words).

This chapter explores how people learn words in an unfamiliar domain. What hypotheses about word meanings do people initially consider when they lack knowledge of what categories exist in the domain, and know no words for any of the objects? How do they learn from limited evidence when generalizing about word meanings for unfamiliar objects? I address these questions and explore whether the Bayesian model presented by XT predicts word learning behavior with unfamiliar objects. Addition-

The work in this chapter was carried out in collaboration with Joshua Tenenbaum.

ally, this chapter explores which factors affect prior beliefs about word meanings: does real-world experience alone cause the formation of a hypothesis space, or can language have an effect as well?

The problem of induction in word learning

Suppose you are a space explorer, setting foot on a new world for the first time. The world is inhabited by a sentient race, and one of the natives is showing you around and telling you the names of different items. Your guide leans over and scoops up something wriggly and apparently 5-mouthed, then hands it to you, exclaiming "Fep!" You gingerly accept the object, and jot down a few observations in your notebook while trying to avoid its vicious fangs. Then you put it in a specimen jar and continue onward carefully, looking around for other feps that might latch onto the ankles of an unwary traveler.

You have only had a single fep labeled for you at this time, but chances are that your guesses about what feps are like as a category are pretty good, even if you have some uncertainty about how to extend the word. Children are faced with this kind of problem all the time when learning a first language here on Earth, and they, like the hypothetical space explorer, learn words quickly – often inferring word meanings from only one or a few positive examples (e.g., Carey & Bartlett, 1978). They also become rapidly more certain about the meaning of a word the more examples they see, as XT showed.

The problem of inducing a whole category from a small set of examples is more difficult even than the space explorer thought experiment may initially reveal. As Quine (1960) famously put it: upon hearing the word *gavagai* applied to a running rabbit, how does one know what the word refers to? It could indicate the specific rabbit, all rabbits, all animals, the rabbit's ears, the action of the rabbit, this rabbit only on this particular day, related thoughts like "Food!" or "Let's hunt!" or any of a number of other possibilities. A number of solutions or partial solutions to this predicament have been proposed in the past. We first outline the major approaches explored prior to XT, and then describe XT's Bayesian approach to word learning.

Previous word learning accounts

One suggestion as to how people learn the correct meaning of a word is that they rule out other possible meanings through *hypothesis elimination* (e.g., Pinker, 1989; Siskind, 1996). This approach combines observations of how words are used across multiple situations with observations of how words are *not* used to deduce the correct meaning. As a simple example, if a child first hears the word *car* in a situation containing both a car and a house, this is consistent with both *car*-CAR and *car*-HOUSE meaning hypotheses I will indicate words with italics and categories with capital letters.. If she now sees a context where the word *car* is used in the presence of a car and a bird, but no house – or a context where a house is clearly labeled *house* – then this does not support the *car*-HOUSE hypothesis. This hypothesis elimination paves the way for the learner to correctly learn *car*-CAR and *house*-HOUSE meanings.

These proposals help explain how children can choose between two non-overlapping categories when learning a word meaning. However, they come up short when trying to explain how children differentiate *car*-CAR and *car*-VEHICLE. Every use of *car* when a car is present will co-occur with an example of both categories. This proposal does not solve the question of how children choose between multiple consistent overlapping hypotheses. Intuitively, a learner's confidence should increase about the correct level of generalization the more examples she sees – the more uses of *car* our learner encounters, the less likely she is to think that it extends to all vehicles, having never seen *car* paired with tractors or buses. Hypothesis elimination fails to account for this phenomenon. Berwick (1986) proposed the *Subset Principle* as a constraint which address this issue by restricting learners' generalizations to the smallest consistent hypothesis out of multiple overlapping choices. However, XT showed that in situations with very few examples, word learners show graded generalization beyond the smallest possible hypothesis, inconsistent which the Subset Principle.

A second proposal for how people learn words is *associative learning.* Associative learning is a mechanism whereby people track the number of times they hear each word in contexts with various items or properties present. This approach involves comparing past uses of a word to other candidate items or categories that the word may label to produce generalizations about word meanings. This can be implemented either by measuring the similarity between past examples and new candidates (e.g., Landau, Smith, & Jones, 1988; D. Roy & Pentland, 2002) or by altering the strengths of connections in a connectionist network between words and various features of the items the refer to (e.g., Colunga & Smith, 2005; Regier, 2005). Additionally, statistical approaches to cross-situational learning often fall into this category (e.g., Yu & Ballard, 2007; Yu & Smith, 2007). This approach shares the limitation of hypothesis elimination: any time the learner hears *blicket* refer to a Dalmatian, the associative strength for *blicket*-DALMATIAN is increased, but so are *blicket*-DOG, *blicket*-ANIMAL, *blicket*-MAMMAL, and so forth. Without negative evidence (e.g., "*blicket* does not mean ANIMAL"), there is no way to decrease the association strengths of the incorrect connections.

Some approaches attempt to address this problem by implementing forms of implicit negative evidence within an associationist model (e.g., Regier, 1996). If a model never receives input pairing *blicket* with any dog other than a Dalmatian, the *blicket*-DOG connection may weaken over time relative to the *blicket*-DALMATIAN connection, for example. However, this proposal does not address how learners can perform fast mapping – learning from as little as a single example.

Other proposed solutions for handling overlapping hypotheses involve biases and constraints limiting the possible word meanings that a learner considers. Markman (1989) proposes that people have bias to label each object with only a single word (the *mutual exclusivity* constraint), and that this bias works together with a preference to map words onto basic-level categories such as DOG rather than other levels of possible generalization (e.g., subordinate categories like DALMATIAN or superordinate categories like ANIMAL)[1]. This combination of constraints would allow children to choose between multiple consistent hypotheses and quickly learning of some of

[1] Markman also proposed a bias to select taxonomic categories rather than other sets of objects; the idea that people have a bias or constraint to view the world hierarchically is discussed further in Chapter 4.

the most common labels for objects. A number of other authors propose an idea related to the basic-level bias — that a shape bias causes children to preferentially map words onto categories of similarly shaped items, at least in the case where the referents are simple, solid objects (Landau, Smith, & Jones, 1988; Imai, Gentner, & Uchida, 1994; Soja, Carey, & Spelke, 1991). Some models show how such a shape-bias could be learned from the evidence (e.g., Kemp et al., 2007; Colunga & Smith, 2005; Samuelson, 2002; L. Smith, 2005). Alternatively, some modeling approaches have addressed the problem of selecting the right meaning by implementing mutual exclusivity in the form of competition between hypotheses (e.g., MacWhinney, 1989; Merriman, 1999; Regier, 2005). However, all of these biases, constraints, and competitions models can only explain how words at a specific level of generalization are learned. People learn to use *animal*, *Dalmatian*, *dog*, and many other words with overlapping references, rather than only learning a single label for each object.

Frank, Goodman, and Tenenbaum (in press) offer a probabilistic model of cross-situational learning which incorporates inference of speaker intention. This allows the model to predict fast word inference in cases like that described by Carey and Bartlett (1978), as well as other child word learning phenomena like mutual exclusivity. The probabilistic cross-situational approach is effective for learning basic-level words, but like many of the other approaches, it does not handle the issue of multiple labels for objects.

None of the major word learning accounts discussed so far can explain all of the word learning phenomena evidenced by children. Children generalize from as little as one example, they rule out overlapping hypotheses based on small numbers of positive examples, and they also learn multiple labels that apply to the same word. The following section describes a model that explains how children can accomplish all of these tasks while learning words.

A Bayesian approach to word learning

XT propose a third learning framework to overcome these limitations. They present a Bayesian model for learning object category names, where even a single piece of evidence provides the learner with probabilistic guesses about the word meaning, and the learner's confidence in various hypotheses changes with more evidence. This model combines the learner's prior beliefs about word meanings with the evidence seen so far. Following the standard Bayesian framework outlined in Chapter 1, the posterior probability of that a word refers to a hypothesized category h, given the data D, is proportional to the product of two factors, the prior probability of the hypothesis and the likelihood of the data given the evidence:

$$P(h|D) \propto P(D|h)P(h) \tag{2.1}$$

The prior probability, $P(h)$, describes which hypotheses are under consideration as possible meanings, and quantifies the probability of each hypothesis *a priori.* XT's implementation included constraints proposed by Markman (1989) for learning object category names – that words are hypothesized to refer to whole objects (rather than, e.g., rabbit ears), and to refer to taxonomic categories (rather than, e.g., things of a particular size). The prior probability also incorporates a preference for more distinctive categories; for example, DOG would have a higher prior probability than ALL DOGS BUT FIDO or DOGS AND WOLVES AND COYOTES BUT NOT JACKALS. The prior probability can also include a preference for basic-level categories (e.g. DOG rather than DALMATIAN or ANIMAL).

The likelihood, $P(D|h)$, describes the probability that the observed data would have been produced if the hypothesis were true. The model includes an assumption that the observed data have been sampled randomly from the underlying hypothesized category. If a learner has seen *blicket* used to label ten Dalmatians, then it is more likely that this evidence was sampled from the category DALMATIAN rather than a larger category like DOG or ANIMAL; if blicket meant ANIMAL, then seeing it applied only to ten Dalmatians would be quite a coincidence. Even if the learner has seen only one use of *blicket*, there are fewer Dalmatians in the world than there are animals, so it's more likely that the category containing just Dalmatians would give rise to this particular example than the category of all animals. This idea is known as the *size principle.* This principle can alternately be seen as a consequence of a goal of informativeness on the part of the speaker (Frank, Goodman, Lai, & Tenenbaum, 2009). The likelihood is:

$$P(D|h) = \left[\frac{1}{|h|}\right]^n \tag{2.2}$$

where n is the number of examples in the dataset D, and $|h|$ is the size of the hypothesized category. Thus, for two overlapping hypotheses that are both consistent with the data, the likelihood of the smaller hypothesis will be higher.

In deciding whether word C also applies to new object y based on the evidence so far, the word learner takes the average of all hypotheses that contain y, weighted by the posterior probability of each hypothesis:

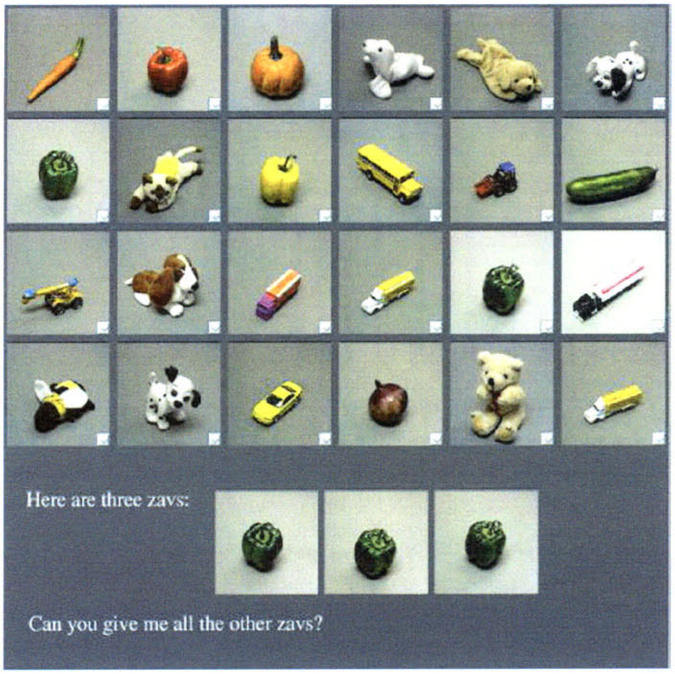
$$P(y \in C) = \sum_{h \supset y,D} p(h|D) \tag{2.3}$$

This model explains how a learner generalizes about word meaning from as little data as a single example, and how the learner's confidence in overlapping hypotheses changes differentially based on evidence.

Evaluating the model in a familiar domain

To test the predictions of the Bayesian model, adults and children were asked to generalize about the meaning of a novel word. An example of an experimental trial is shown in Figure 2-1. In each trial, participants saw either one or three examples of the novel word. In the three example trials, the size of the taxonomic category that the examples were drawn from varied; for example, the participant might see three green peppers, three peppers, or three vegetables all labeled with the new word. Participants saw a large grid of items and were asked whether the new word applied to each one. The items in the grid were drawn from three superordinate categories (animals, vegetables, and vehicles), and included multiple objects drawn from different sizes of category within the taxonomy.

XT represented the participants' hypothesis space by finding out how similar they



Here are three zavs:

Can you give me all the other zavs?

Figure 2-1: A trial from the XT word learning study.

thought each pair of stimuli was and creating a tree by clustering items based on their similarity (see Figure 2-2). The height of node h in the tree, height(h), corresponds to the average dissimilarity of the objects within the category. XT take this value to approximate category size, and they calculated the likelihood as follows:

$$p(D|h) \propto \left[\frac{1}{\text{height}(h) + \epsilon}\right]^n \tag{2.4}$$

if all the items seen in the data are within the category h; 0 otherwise. ϵ is a parameter that allows for some uncertainty about the size of the category, weakening the size principle somewhat accordingly. Xu and Tenenbaum used a value of $\epsilon = 0.05$ for their model simulations, but predicted that larger values may be appropriate when word learners were less able to judge the size of the categories in the hypothesis space.

The prior probability was calculated as follows:

$$P(h) \propto \beta(\text{height}(\text{parent}(h)) - \text{height}(h)) \tag{2.5}$$

The distinctiveness bias was implemented as the length of the branch connecting the node h to its parent (the distance between the heights of the two nodes). β indicates the strength of the basic-level bias (this parameter was set to 1 at non-basic level nodes).

XT found that the model was very good at predicting people's generalizations of word meanings. Both adults and children showed graded generalizations from a single example. Their certainty about the correct word meaning increased with more evidence, consistent with the size principle. Consistent with the prior bias for distinctive categories, word learners did not simply generalize to the smallest possible category that contained the example items. Interestingly, adults showed a strong basic-level bias while children in children the bias was less pronounced. The model also correctly predicted that adults and children were able to generalize multiple word meanings to include the same object at different levels of description (e.g., *Dalmatian*, *dog*, *animal*).

The XT model solves many previous issues, showing how people can choose between overlapping hypotheses and learn from little evidence, while still being able to multiple words for an object. But it leaves some questions unanswered, particularly when considering the word learning problem faced by a learner in a foreign domain – such as a space explorer, or, far more commonly, a child explorer learning about our own world. The children participants in XT were old enough to be familiar with the categories used in the experiment.

Does a word learner in a similarly complex but unfamiliar domain learn words in the same way? Does she still form a structured hypothesis space given novel objects? Within this space, does she show a distinctiveness bias? A preference for words to label categories at the basic level? Does she incorporate new evidence about word meanings in a manner consistent with the size principle? The following experiment investigates how well XT's Bayesian model predicts word learning with novel objects and categories.

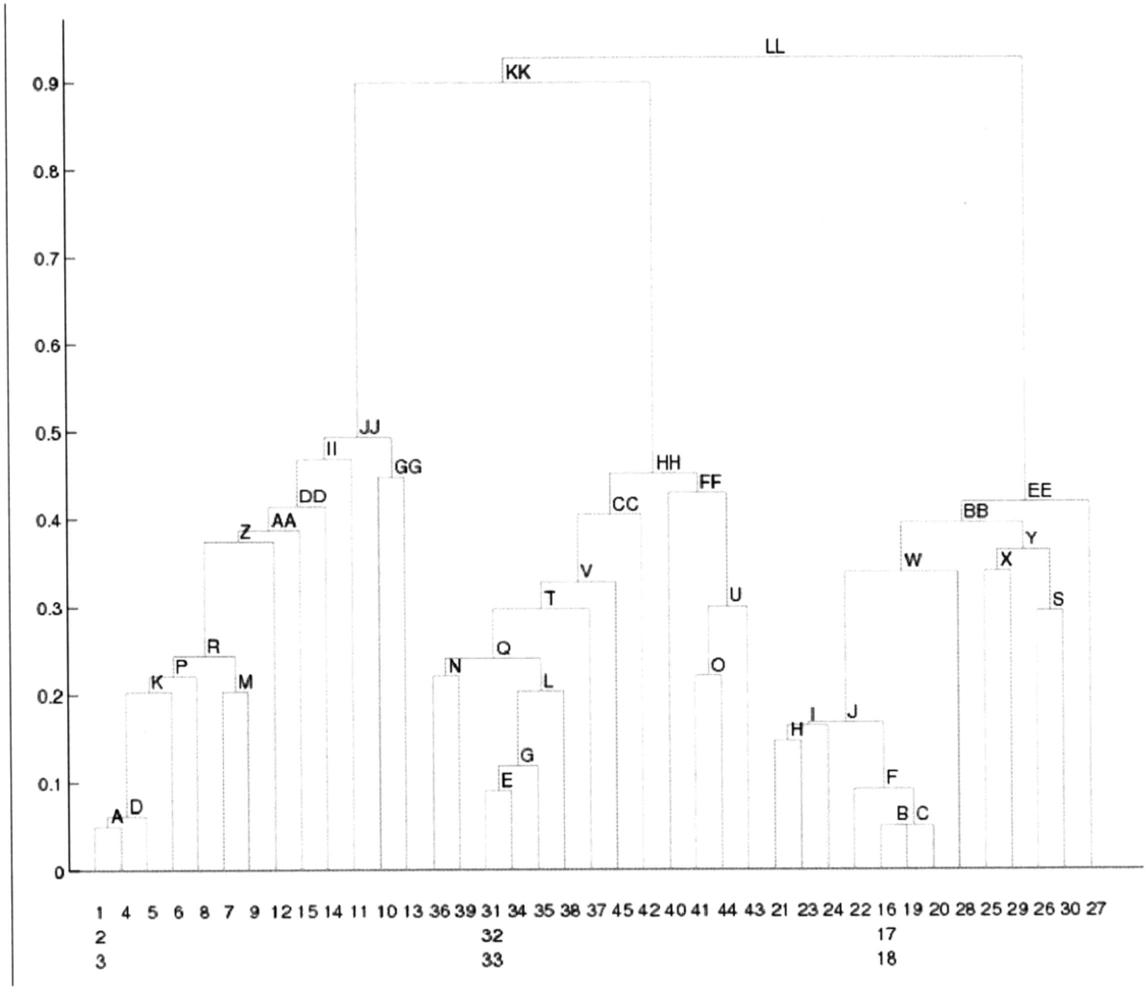


Figure 2-2: The taxonomy created from the adult similarity ratings in the XT study (picture taken from Xu & Tenenbaum (2007)). The superordinate nodes are vegetable (EE), vehicle (HH), animal (JJ); basic-level nodes are pepper (J), truck (T), dog (R); subordinate nodes are green pepper (F), yellow truck (G), and Dalmatian (D).

Experiment 1: Word learning in an unfamiliar domain

This experiment replicated the XT adult word-learning paradigm using novel objects with only minor modifications. As in the original study, the experiment consisted of two phases, a word learning phase and a similarity judgment phase. In each trial of the word learning phase, participants saw one or more examples of a new word; the data that varied in terms of the number of examples (one or three) and the size of the category represented in the three-example trials (subordinate, basic-level, or superordinate categories). Participants were shown a number of other objects and asked which ones they thought the word applied to. Instead of being offered a binary choice for each object, as in XT, subjects were asked to rate their confidence that the word applied to each object in the test set. In the similarity judgment phase, participants rated pairwise similarities of objects based on their perceptual features. These judgments were used to infer participants' representation of the relationships between the objects.

This experiment investigated whether the XT model predicts word learning behavior in a domain containing novel objects and categories. If the model is a still a good predictor of word learning in these circumstances, then the following predictions should hold:

- The amount of evidence a word learner sees will affect her degree of generalization, in accordance with the size principle. Word learners will be more confident about the meaning of a new word after seeing three examples than after seeing a single example. In particular, when a word learner sees three subordinate examples (equivalent to seeing three Dalmatians in the original experiment), she will be more confident that the word means DALMATIAN than when she only sees a single example, and will restrict her generalization to objects that are not in the same subordinate level category accordingly.

- The type of evidence that a word learner sees will affect her patterns of generalization. Word learners will generalize to different extents depending on whether they see three items drawn from the same subordinate, basic-level, or superordinate category.

- Word learners will show a preference for generalizing to more distinctive categories. Instead of generalizing only to the smallest possible subset of items that matches the data, word learners will show graded generalization beyond the examples given and will tend to extend word meanings to distinctive categories like subordinate, basic-level, or superordinate level categories.

- Word learners may also show a basic-level bias. XT found such a bias in adults but not children. If the preference for the basic level is one that adults generalize beyond existing categories to learning situations in unfamiliar domains, then a basic-level bias is predicted here as well. If, however, adults behave more like children when they are faced with similarly unfamiliar categories, then a basic-level bias may not appear.

After outlining the experimental methods, I will break the results into three parts. In the first part, I will examine the outcome of the first three of the above predictions with regard to the word learning data. In the second part, I will look at participants' similarity data and discuss differences between their hypothesis space and that of the adult subjects in XT. In the third part, I will discuss the details of the fit of the Bayesian model to the word learning data. The modeling results will include a discussion of the final prediction, regarding a basic-level bias.

Method

Participants

Subjects were 36 students from Stanford University, participating for pay or course credit. All subjects participated in the word learning task followed by the similarity judgment task. All were native speakers of English.

Materials

The stimuli were 45 pictures of novel objects (see Figure 2-3 and Figure 2-4). The objects were generated using 3D modeling software called Biowin. The Biowin program allows for randomization and mutation of the parameters used to generate object models.

The objects were designed so as to mimic the taxonomic structure of the familiar objects. The novel taxonomy, like the familiar object taxonomy of XT, consisted of three superordinate categories of objects. I will refer to the three novel superordinate categories as plants, tools, and shells for convenience, based on the Earth categories that were used as partial inspiration in creating them. However, it is important to note that these novel categories are not intended to map directly onto any familiar Earth categories, and these words were never used within the context of the experiment.

The stimuli were constructed so that, for one item in each of the three superordinate categories, the stimuli set contained potential matches at the same subordinate category, four additional members of the same basic-level category, and seven additional members of the same superordinate category. The Biowin parameters used to generate items of the same subordinate category varied a small amount[2]. The parameters varied more for items of the same basic level, and to an even greater extent for items within the same superordinate category.

The 45 objects were divided into 21 stimuli that people saw as examples of the novel words (the *training set*), and 24 stimuli that they could choose from when generalizing about the meaning of the novel words (the *test set*).

[2] Parameters were varied through a combination of random mutation and hand-tweaking. The Biowin program did not randomly mutate entire shapes, but only characteristics such as number of body segments or degree of rotation; this meant that variation at the basic and especially superordinate levels required my supervision. In all cases, I attempted to mimic the type and degree of variation found in Earth taxonomies, without mimicking too closely any known categories.

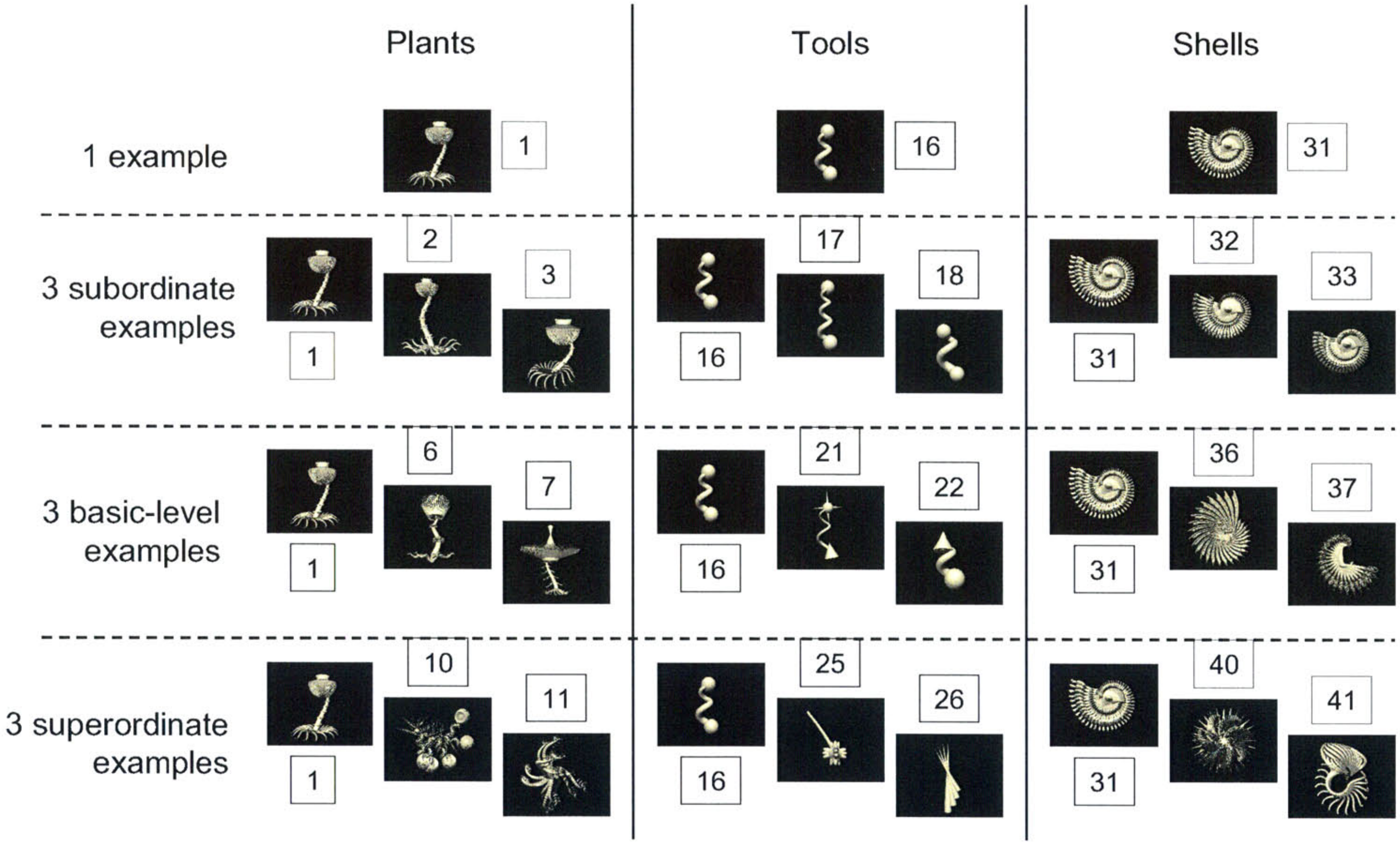


Figure 2-3: The sets of objects used as examples of novel word meanings in Experiments 1 and 2.

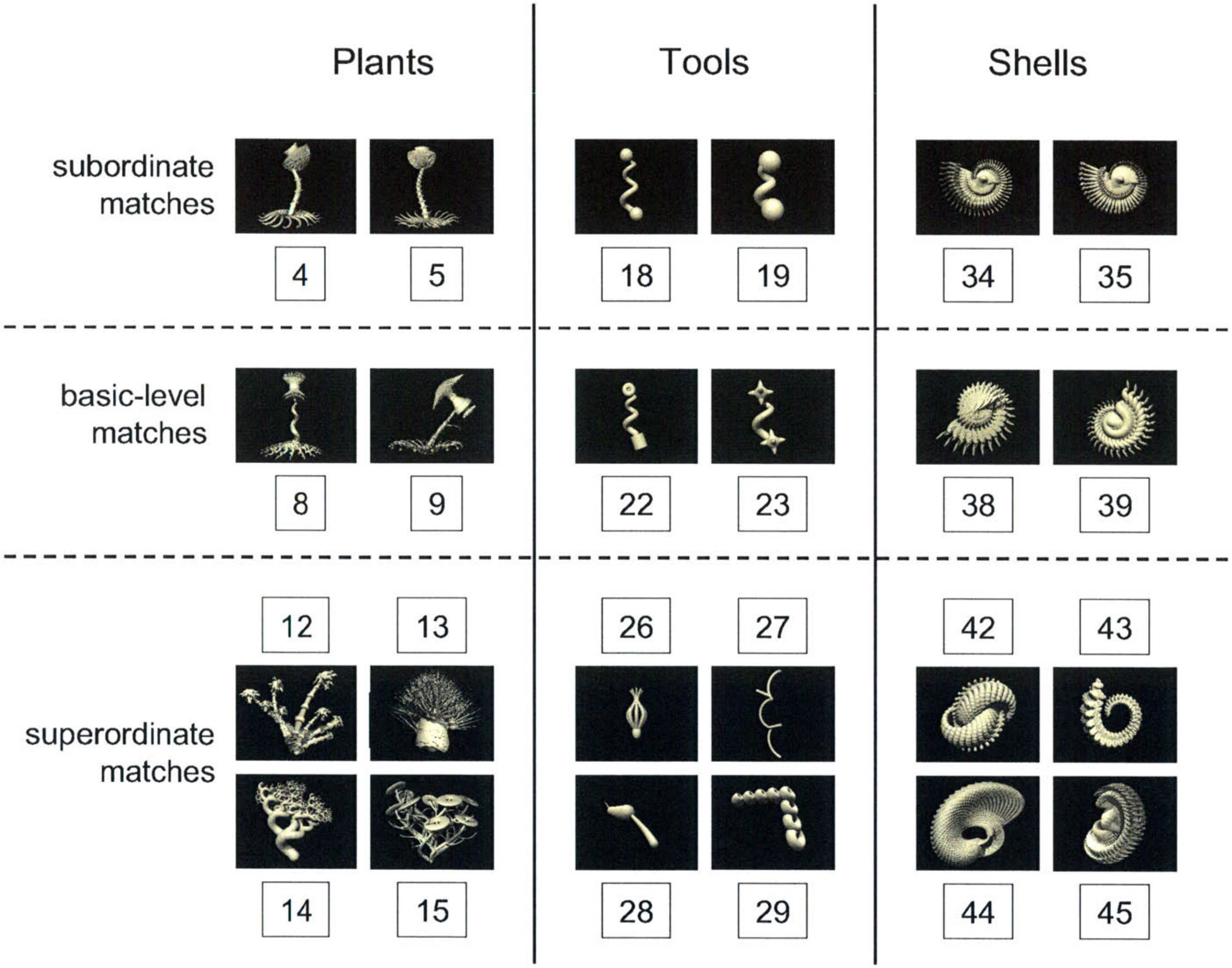


Figure 2-4: The objects used as candidates for word generalization in Experiments 1 and 2.

The items in the training set were shown as examples in 12 different trials (see Figure 2-3). The first three trials showed one example each (one plant, one tool, and one shell), each labeled with a novel term. The remaining nine trials showed three items each. The trials differed in terms of the object type and category specificity of the examples: plant, tool, shell x subordinate, basic-level, superordinate. In each of the three-item trials, one of the training items was the same one that was shown in the one-item trial from the same superordinate level category. The trials occurred in a pseudo-random order, with the content of the training sets counterbalanced across participants for object type and category specificity.

The test set was the same across all trials (see Figure 2-4). The test set was constructed so that, given any one training item, there were two test items from the same subordinate-level category, two test items from the same basic-level category, four test items from the same superordinate-level category, and 16 distractor items (the objects from the two other classes of items). All of the setup of the training and test sets match that of the original XT experiment; only the objects are different.

Procedure

The experiment was divided into two portions, the word-learning and the similarity-judgment phases. Both portions took place at a computer. In the word-learning task, participants were told that they were going to be exploring the planet Gazoob, where their native guide would be teaching them the Gazoobian words for some objects found on the planet. On each trial, the participants saw a novel, monosyllabic word labeling either one or three example items. Above the example item(s), they saw the test set. For each item in the test set, participants indicated whether they thought the word applied to that item.

The second portion of the experiment was the similarity-judgment phase. In this phase, subjects saw pairs of items and were asked to rate how similar the two items were on a scale of 0 (not at all similar) to 9 (extremely similar). Judgments were collected for all possible pairs of 45 objects, but each participant rated only a subset of the possible pairs. Each participant provided similarity judgments for all possible within-class pairs – that is, each pair of plants, each pair of tools, and each pair of shells – along with a fraction of the possible remaining cross-class pairs, selected pseudo-randomly. The order of the trials and the order of the two objects in each trial were randomized across participants.

Preceding these trials, participants were given a short warm-up session containing practice trials, each showing a pair of items randomly selected from the same stimuli. During these trials, the participants became familiar with the range of similarities they would be asked to rate, and they were encouraged to spread their ratings out so that they made use of the entire 0-9 scale in rating object-pair similarities. They were also encouraged to develop a consistent method for using the scale during this period.

I used the same model as XT, letting both β and ϵ vary as free parameters. In other words, the degree to which the model had a basic-level bias and the degree to which the size principle was in effect in the model were determined by the best fit to the data.

Word learning data

In analyzing the results, I thresholded subjects' ratings to yield a binary result (the word does or does not apply to the object in question). I averaged the responses across participants, and for most analyses, I also collapsed results across the three superordinate categories. In other words, most analyses addressed questions of the type, "Given N examples of taxonomic specificity S, how broadly did word learners generalize?" Figure 2-5 shows the main results overall, and broken down by superordinate category type.

As discussed further in the following section, one of the shell objects in the test set which was designed to be a superordinate match was perceived as being a basic-level match by participants. This item was therefore grouped with the basic-level matches for all of the following analysis. I return now to the initial predictions and discuss the outcomes broadly:

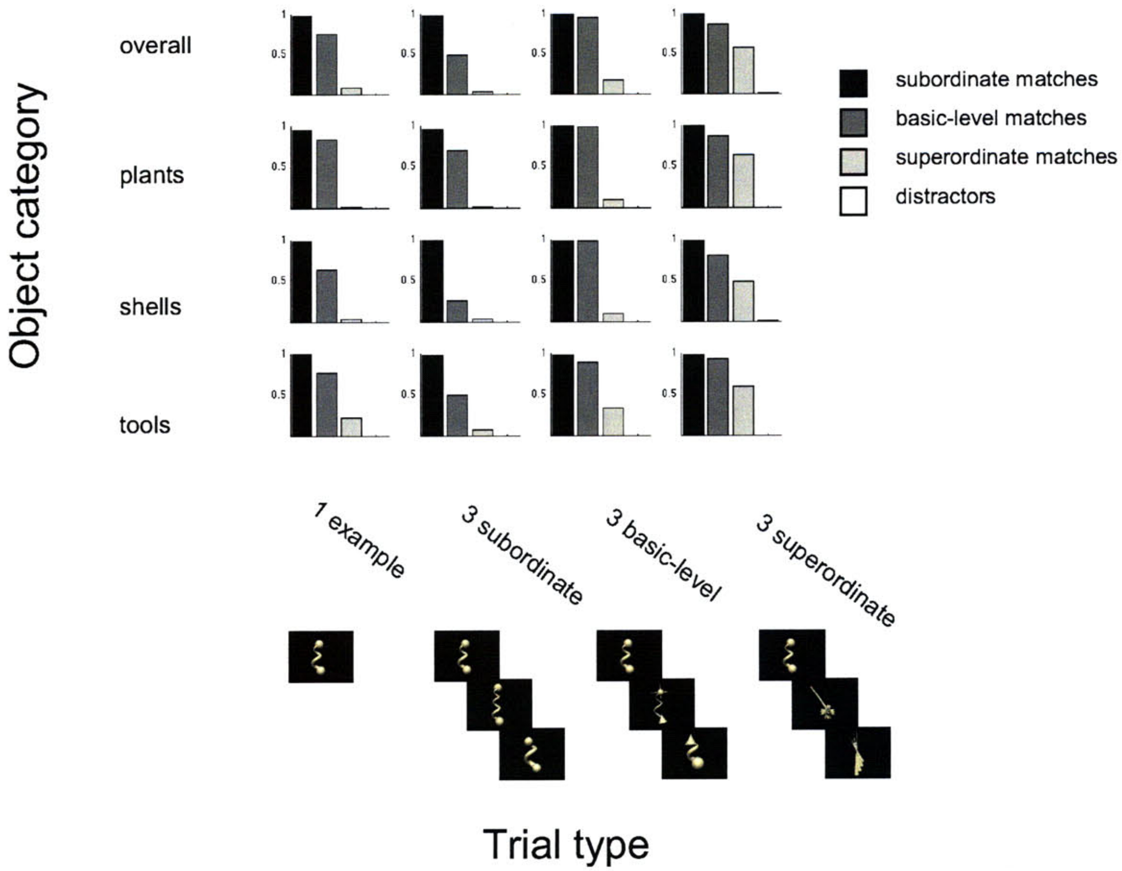


Figure 2-5: The word learning results from Experiment 1. The overall results are shown at the top, and breakdown by superordinate category shown below. Generalization results are collapsed across trial type, indicated at the bottom of the figure.

> *The amount of evidence a word learner sees will affect her degree of generalization, in accordance with the size principle.*

Participants did show different degrees of confidence about word meanings based on how many examples were shown. In particular, participants showed the predicted generalization restriction from the one example trials to the three subordinate example trials. Generalization from one example at the basic level (75%) was significantly greater than generalization from three subordinate examples (49%) ($t(70) = 3.80$, $p = 0.00015$), averaging across all classes. Breaking down the items by classes, we find that the difference in generalization was significant or marginally significant for all of the object types: the plants (83% vs. 69%, $\chi^2 = 5.95$, $p = 0.051$), the tools (64% vs. 26%, $\chi^2 = 11.41$, $p = 0.0033$), and the shells (77% vs. 50%, $\chi^2 = 8.52$,

$p = 0.0364$)[3].

The type of evidence that a word learner sees will affect her patterns of generalization.

As predicted, participants showed different degrees of generalization depending on the specificity of examples they saw in the three-example trials. In the analysis, I looked for generalization to be essentially at ceiling for all of the test items at the same or greater degrees of specificity as the example items, and then to drop sharply for all test items that were members of larger (less specific) categories.

Subjects showed the predicted large drop from subordinate to basic level generalization in the 3-subordinate trials (98% vs. 49%, $t(70) = 10.25$, $p < 0.0001$) and a large drop from basic to superordinate level generalization in the 3-basic-level trials (95% vs. 17%, $t(70) = 27.57$, $p < 0.0001$). Adults in the original XT experiment generalized to most or all of the superordinate matches in the 3-superordinate trials. This did not hold true with the novel objects; subject chose only 57% of the superordinate objects, significantly less than all the objects($p < 0.0001$). However, though they were not at ceiling for superordinate generalization in 3-superordinate trials, subjects' generalization patterns were significantly different from 3-basic-level trials in the expected direction. Subjects generalized to significantly more superordinate level items when they saw three superordinate examples vs. three basic-level examples (57% vs. 17%, $t(70) = 7.59$, $p < 0.0001$).

Word learners will show a preference for generalizing to more distinctive categories.

Participants showed graded generalization as predicted, generalizing beyond the bounds of the smallest consistent subset, and also including other members of distinctive categories consistent with the evidence. Subjects who saw a single example item generalized to almost all subordinate matches (98%) – beyond the bounds of

[3] If the shells are analyzed according to designed taxonomic level rather than perceived level, this result is not significant: $p = 0.066$. Breakdowns by object type not reported henceforth unless their outcome is different from the overall outcome

the single-item category that is the smallest possible word extension. The degree of generalization dropped significantly between subordinate and basic -level matches (75%) ($t(70)$ = 4.55, p < 0.0001). Participants also generalized significantly more to matches at the basic level than the superordinate level (8.3%) ($t(70)$ = 12.06, p < 0.0001). The difference between the basic and superordinate levels of generalization (66%) is significantly greater than the difference between the subordinate and basic levels of generalization (23%) ($t(70)$ = 6.48, p < 0.0001). This pattern of graded generalization based on one item replicates that found in the XT study.

Overall, the Bayesian word learning model presented by XT makes predictions that are substantially borne out with novel objects and categories, as well as in a familiar domain. With the exception of the generalization gradient on the three superordinate training item trials, which is discussed more in the modeling section, the general predictions of the model matched participants' word learning data.

Similarity data

The similarity data collected were used to form the hypothesis space of the Bayesian word learning model. Using hierarchical clustering (Duda & Hart, 1973), we generated a tree-structured arrangement of the objects used as stimuli based on the similarity ratings between the objects (see Figure 2-6). Each node in the tree is a category in the hypothesis space of the Bayesian word learning model. The distinctiveness of a category is proportional to the length of the branch separating the category node from the one above it.

Participants recognized the intended taxonomic structure with only a single exception. Almost all of the categories designed to be distinctive within the novel object taxonomy were rated as coherent categories by the participants; that is, if a category at one of the three major levels of specificity (subordinate, basic-level, or superordinate) was designed to contain five items, participants grouped those five items together in a single node in the tree which contained no other items (see the labeled nodes in Figure 2-6). The exception was a single shell object designed to be a superordinate match, which was perceived by participants to be a basic-level match (object 45).

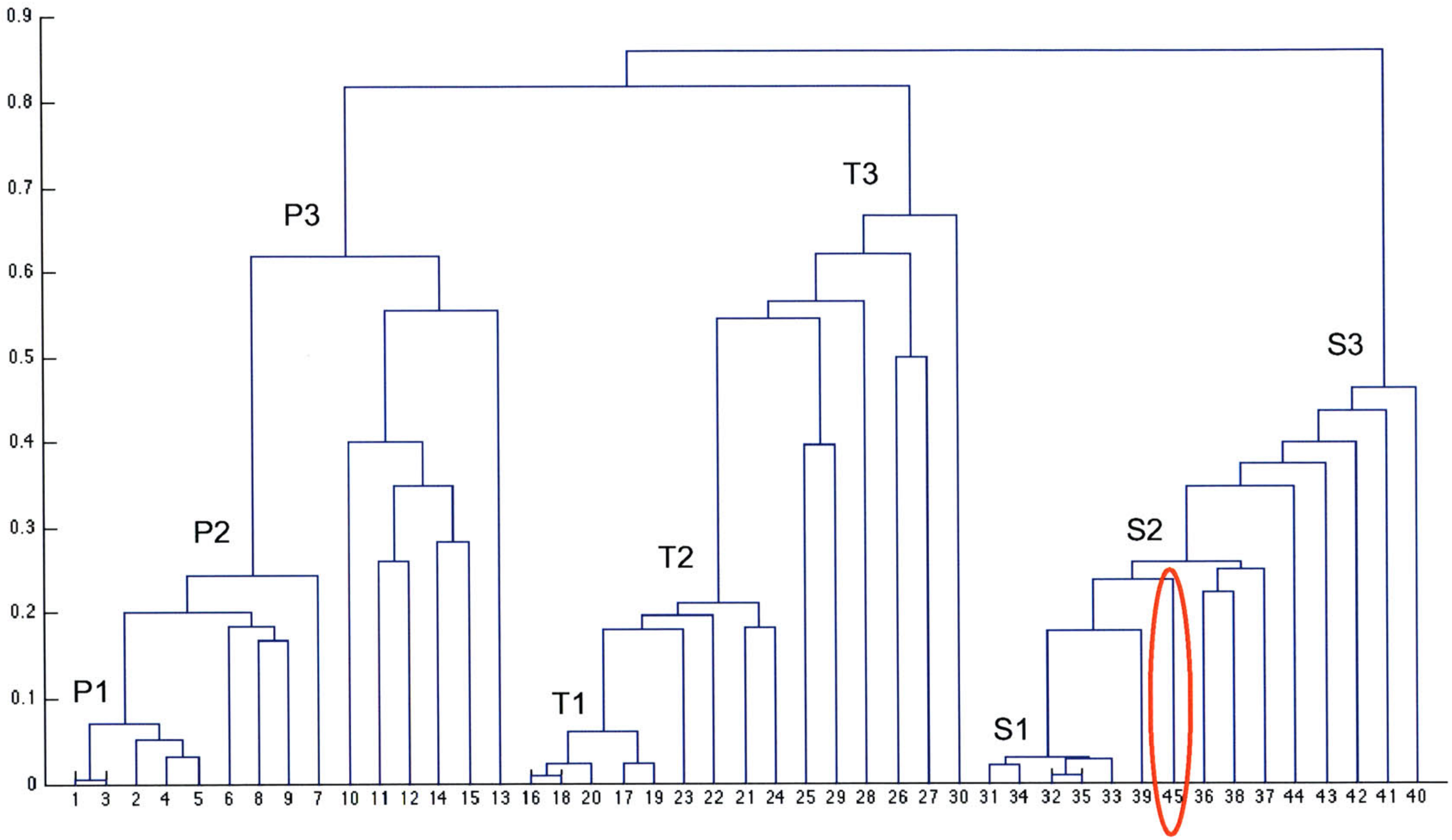


Figure 2-6: The taxonomic structure created from similarity ratings in Experiment 1. Nodes corresponding to the subordinate (1), basic-level (2), and superordinate (3) categories are labeled (immediately above the node) with the corresponding number and a letter indicating the category type (P=plant, T=tool, S=shell). The circled object is Object 45, a shell designed to be a superordinate category match, but perceived as a basic-level match.

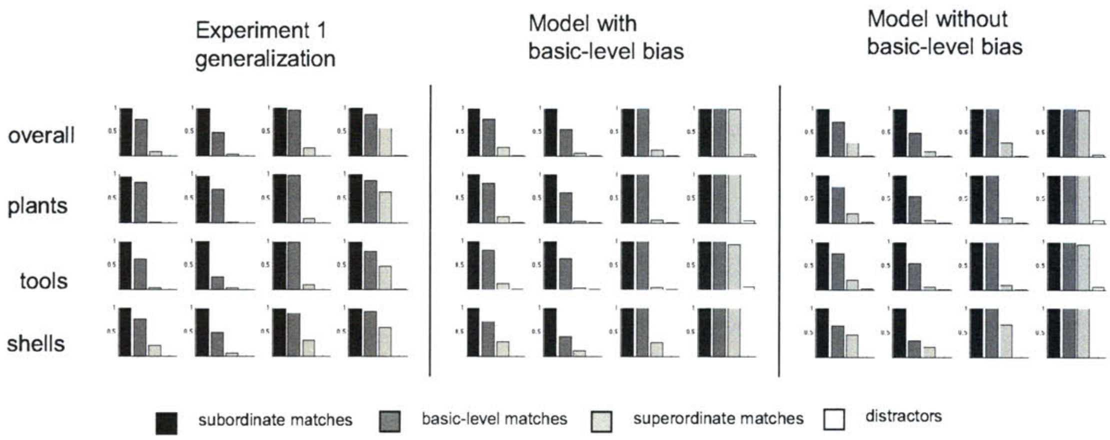


Figure 2-7: The Experiment 1 word learning data reproduced next to the model results, with and without basic-level bias.

The three superordinate categories of items were not as distinct from one another in this experiment as in the taxonomy generated by XT's adult subjects (cf. Figure 2-2); the unfamiliar objects were rated more similarly across superordinate categories on average than the familiar objects were. In both XT's experiment with adult subjects and this experiment, the subordinate category nodes were very distinct from the basic-level matches (separated by longer branches than within-category branches). The basic-level nodes were also distinct from all the superordinate matches, as in XT. However, in the shell category, the basic-level category was not nearly as distinctive as in the plant and tool categories. The superordinate category itself was distinctive, but the structure within it was less clear.

Modeling results

I compared the predictions of the XT model both with and without a basic-level bias to the empirical word learning results. Figure 2-7 shows the comparison of the participants' generalizations to the predictions of both models. Both models predicted learner's generalization patterns well; there was a slight improvement in model fit given a small basic-level bias (r=0.9749 with basic-level bias vs. r=0.9685 without basic-level bias). The optimal basic-level bias was very weak; $\beta = 1.1$ as opposed to $\beta = 10$ in XT (a lack of basic-level bias is equivalent to $\beta = 1.0$).

The model showed a weakened effect of the size principle as compared to XT's results with familiar objects. With a basic-level bias, the optimal value of ϵ, the parameter indicating uncertainty over category size, was 0.25; without a basic-level bias, $\epsilon = 0.35$ (cf. XT, $\epsilon = 0.05$).

The main difference between the model predictions and the participants' generalizations occurs in the 3-superordinate example trials. In these cases, the model predicts generalization to all matches within the same superordinate category, but participants chose far fewer than 100% of the superordinate matches, as discussed in the section on the word learning data. It is possible that the participants had no problem recognizing the superordinate categories but did not generalize from three examples to that category with high confidence. However, given the novelty of the objects, it seems more likely that the participants were uncertain of the taxonomic structure – especially the relationships superordinate matches, which have the least in common perceptually with one another and with the training examples. The similarity ratings lend support for this idea; the heights of the superordinate category nodes indicate a high degree of dissimilarity between the superordinate objects. The relatively weak size principle here is also consistent with uncertainty over the structure of the taxonomy, which would result in uncertainty about the sizes of categories within the taxonomy.

Uncertainty at the superordinate level in particular is perhaps not surprising; learning superordinate categories frequently involves understanding of properties such as the behavior or function of items within the category. Perceptual features such as shape and texture can vary widely within a single superordinate category on Earth – and did so in the Gazoobian objects designed to mimic Earth taxonomies, as well. Participants in this experiment lacked any information about the stimuli except for their appearance, and thus had limited data from which to infer their relationships.

Child learners may also initially have limited evidence from which to learn superordinate categories; labels for these categories are relatively hard to learn (Markman, 1989). Therefore, uncertainty over taxonomic structures may be common in early language learners as well as adult learners in a novel environment. In learning environments with unfamiliar objects and categories, it therefore may not ultimately make sense to model the learners' hypothesis space with a single tree, as done here. Instead, models that allow for modeling uncertainty about the tree structure itself may better predict a word learner's generalizations. See the General discussion section for further exploration of this option.

The lack of a strong basic-level bias here is a particularly interesting result; XT found that adults exhibited a basic-level bias, but children did not. There are many ways in which Rosch (1978) originally pinpointed the basic level as being more salient than other levels: such cues include common attributes, similarity in shapes, and similarity in the motor movements used by someone interacting with them. Subordinate level category members also share these features, of course, but categories at the basic level are more distinctive from one another according to these cues than any other level of category. Many of these cues, being related to perceptual features of the objects, are already captured by the distinctiveness bias; these features make the basic level matches stand out even before a learner has any knowledge about their actions, functions, or labels. The one measure by which Rosch found the basic level to be by far the most distinctive, however, was the labels that people used in spontaneous naming tasks. People asked to name random objects provided basic-level labels 98.6% of the time — far more than the subordinate or superordinate labels (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976).

These data lead to an interesting possibility about the developmental trajectory of the basic-level bias. Perhaps learners start out with only a distinctiveness bias predisposing them to generalize new word meanings to other objects in the same basic level category. However, as they — and everyone else — tend to use the basic-level label far more frequently, that category level becomes more salient, leading to a feedback loop that culminates in the development of a strong basic-level bias for familiar objects. If this is the case, then we would predict that both young children and word learners in an unfamiliar domain, having not yet developed this language and attention feedback loop, would show a weak or non-existent basic-level bias. If such a feedback loop does cause basic-level salience, it is dubious that a basic-level bias would be a very powerful force in guiding new word learning. The possibility of language affecting prior expectations about word meanings is explored further in the following experiment.

Overall, the Bayesian word learning model presented by XT predicts most of the generalization patterns seen by human word learners in a novel domain. A strong basic-level bias proves unnecessary to explain word learning with novel objects. Uncertain generalization at the superordinate level, and a weakened size principle, both suggest that modeling uncertainty over tree structures may be a more accurate way to model the cognitive realities of word learners with unfamiliar objects. However, the current Bayesian model provides a strong fit to the data of Experiment 1. It continues to account for word learning phenomena which cause problems for other word learning approaches, allowing for graded generalization from even a single example, as well as selection between overlapping hypotheses in accordance with the evidence.

Experiment 2: Constructing a hypothesis space

How do people form their initial hypotheses about what words might mean? Thus far, I have assumed that the structure of the hypothesis space is inferred based solely on pre-linguistic knowledge of the objects – their perceptual features, in the case of the Gazoobian objects. Yet as people learn words that pick out particular categories, they will potentially find these categories more salient. Does this alter their hypothesis space for future word learning, and possibly for other kinds of category learning?

There are two very distinct ways in which people might formulate their hypotheses about word meaning during the course of learning about an unfamiliar set of objects. These are illustrated in Figure 2-8:

- The non-Whorfian account (so called for reasons discussed below): people assume that the perceptual features of the objects that they see stem from the underlying taxonomic structure[4], which may itself be governed by parameters

[4] and potentially other knowledge about the objects as well, but we limit our discussion to perceptual features for the purposes of discussing the Gazoobian word learning experiments.

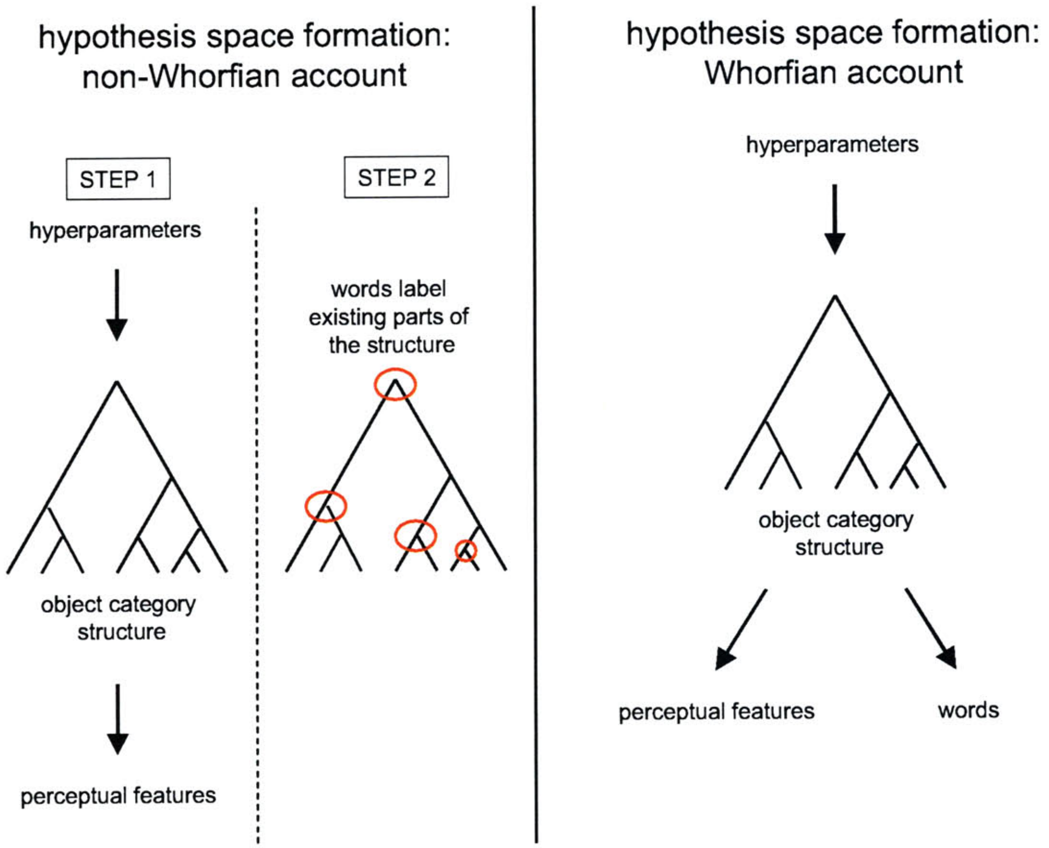

hypothesis space formation:
non-Whorfian account
STEP 1
hyperparameters
object category structure
perceptual features
STEP 2
words label existing parts of the structure
hypothesis space formation:
Whorfian account
hyperparameters
object category structure
perceptual features
words

Figure 2-8: Two possible accounts of how the hypothesis space is formed. In the first account, the hypothesis space is inferred based on non-linguistic features such as perceptual features. Words can then label existing categories without affecting the hypothesis space. In the second account, words themselves can also be used to infer the underlying representation of object categories, changing the hypothesis space.

specifying, for instance, the clumpiness of objects, or the number of distinctive categories in the taxonomy. People therefore work backwards from the perceptual features to infer the underlying relationships between the objects. The taxonomic structure that they infer then serves as a hypothesis space for word meanings: words label categories within the existing structure, but do not themselves affect peoples' mental representation of the objects' relationships.

- The Whorfian account: the language learner infers the underlying object category structure from both perceptual features and word meanings. If a word labels a category that is not very perceptually distinctive, the word learner may infer that that category is nonetheless important (i.e., distinctive) in the under-

> lying taxonomy. As word learning occurs, the hypothesis space for future word meanings changes over time.
>
> To be clear, there are two ways in which learning words could change the hypothesis space. New words could pick out categories that learners had already inferred based on perceptual features and make these categories more salient, therefore changing the lengths of branches within the tree. It is also possible that learning a word that refers to a category that is not consistent with the previous hypothesis space (e.g., learning that the word *Dalmapeño* applies to Dalmatians and jalapeños but nothing else) could actually alter the structure of a word learner's hypothesis space topologically (i.e., change the categories that exist as hypotheses). We will be focus on the former type of effect of language on thought for the purposes of this experiment.

This experiment tries to tease apart the two accounts of hypothesis space formation.

Before I endeavor to do so, it is worth asking: should we expect language to be able to affect mental representations? Whorf (1956) famously put forth the idea that language shapes thought, stating in one of his strongest formulations, "Language is not simply a reporting device for experience but a defining framework for it." In the intervening years, whether and how much language influences thought have been hotly contested. The strong version of the hypothesis – that language determines thought in an Orwellian manner, such that lacking a word for *war* prevents people of being able to think of war – seems self-evidently ridiculous, and has caused many to argue against all forms of the Whorfian hypothesis (e.g., Pinker, 1994). However, a growing body of evidence demonstrates a number of weaker ways in which language can shape thought (see Gentner and Goldin-Meadow (2003) for an overview).

Word meanings have been shown to have several effects on thought. The way that languages divide up the color spectrum into labeled categories affects color similarity, memory, and discrimination (Winawer et al., 2007; Kay & Regier, 2007; Kay & Kempton, 1984; Roberson, Davies, & Davidoff, 2000). Different languages also divide up the space of spatial relationships and spatial motions differently, using words like prepositions to group different kinds of relationships or motions together (Bowerman & Choi, 2003). For example, the English preposition *in* highlights spatial relationships involving containment, but does not address whether the container fits tightly or loosely around the contained object, whereas Korean uses separate prepositions depending on the fit, while caring less about whether the item is contained. These cross-linguistic differences in spatial language affect spatial cognition, including how speakers categorize spatial relationships and how quickly they learn categories (Bowerman & Choi, 2003; Choi, McDonough, Bowerman, & Mandler, 1999; Brown, 2001; León, 2001). Recent evidence even shows that lacking exact number words limits speakers' ability to reason about exact number ((Frank, Everett, Federenko, & Gibson, 2008); (Everett, 2005); (Gordon, 2004); (Pica, Lemer, Izard, & Dehaene, 2004)).

Clearly, language can change our representations of similarity within spaces that are relatively abstract and/or continuous. Language can also highlight certain features of objects and make groups of objects more salient to speakers by placing objects within a shared grammatical category. For speakers of languages with a grammatical gender system, the grammatical gender of an object affects both the adjectives that the speaker uses to describe the object and how similar the speaker perceives the object as being to items with biological gender (e.g., a king or a ballerina) (Boroditsky, Schmidt, & Phillips, 2003). However, grammatical gender systems divide objects into only a few mutually exclusive categories, and these categories are constantly made salient by the fact that the grammatical gender of an object must be marked through many forms of grammatical agreement (e.g., articles and adjective endings in German). Can noun learning change our representation of objects and their relationships to one another? Bloom and Keil (2001) and others are skeptical, and even some of the proponents of the Whorfian hypothesis are careful to point out that language has limited ability to affect thought in a concrete, discrete domain ((Gentner & Boroditsky, 2001)).

Even if we might expect such an effect, no computational models have previously been proposed showing how language could affect the construction of a hypothesis space for word learning. In this experiment, I investigate whether word learning has an effect on the learner's mental representation of inter-object relationships, and I show how the Bayesian word learning model can quantify such an effect. This experiment swaps the order of the word learning and similarity judgment phases in Experiment 1. If a word learner's hypothesis space is affected by learning new word meanings, then we predict that reversing the order of the two experiment phases will yield a different tree structure than that of Experiment 1 – people will give different similarity ratings before they learn words for the objects (Experiment 2) than after the learn words (Experiment 1). In particular, we predict that categories named in the word-learning task will be more distinctive to learners after the word-learning task.

Method

Participants

Participants were 31 students from Stanford University, participating for pay or course credit. All subjects participated in the similarity judgment task followed by the word learning task. All were native speakers of English.

Materials

The stimuli were the same 45 pictures of 3D modeled objects used in Experiment 1.

Procedure

The experimental procedure was the same as that of Experiment 1, except that the order of the similarity judgment and the word learning phases was reversed.

Word learning data

Participants' generalization patterns in Experiment 2 were substantially the same as in Experiment 1. The same analyses were performed as in Experiment 1, with almost

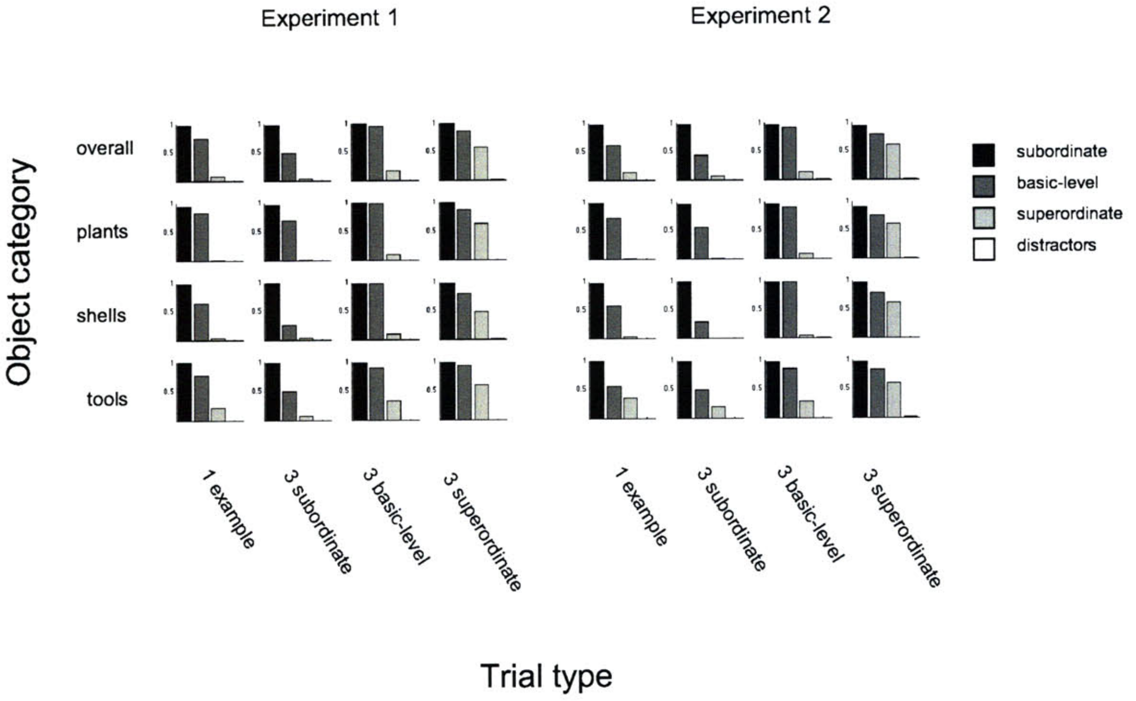


Figure 2-9: The word learning results from Experiment 2, with the results from Experiment 1 reproduced for ease of comparison.

all the results being significant in the same direction. The sole exception was that, whereas in Experiment 1, the generalization from one example dropped significantly more between basic-level and superordinate matches than between subordinate and basic-level matches, in Experiment 2 this difference was not significant ($p = 0.077$). Nevertheless, the same overall pattern of graded generalization from a single example was present in both experiments.

We also performed t-tests comparing the result of each level of generalization specificity in each trial type between the two experiments (i.e., comparing each of the bars in the overall results graphs in Figure 2-9). There were no significant differences between the two sets of results.

Similarity data

The tree recovered based on the similarity ratings in Experiment 2 differed from the tree recovered in Experiment 1 (see Figure 2-6 vs. Figure 2-10). The trees did not

Table 2.1: Branch lengths of the supernodes before and after word learning. The larger number in each row is in bold.

Supernode	Before word learning	After word learning
Plants: subordinate	.088	**.130**
Plants: basic-level	.368	**.374**
Plants: superordinate	.075	**.200**
Tools: subordinate	.090	**.119**
Tools: basic-level	.280	**.334**
Tools: superordinate	.083	**.152**
Shells: subordinate	**.159**	.150
Shells: basic-level	.048	**.089**
Shells: superordinate	.272	**.396**

differ substantially topologically, but did exhibit an effect of word learning on category distinctiveness.

For the purposes of these analyses, I will call the nodes in the tree most closely corresponding to the subordinate, basic-level, and superordinate categories as supernodes. These nodes are labeled in both figures. I compared the topological structure of the two trees by comparing whether any objects changed position relative to the supernodes between the two trees. With one possible exception, none of the objects changed positions with respect to these labeled nodes from Experiment 1 to Experiment 2. The possible exception is object 45, which was grouped with the basic-level shells in Experiment 1, but was ambiguously grouped with either the basic-level or superordinate shells in Experiment 2. This object was grouped with the superordinate matches rather than the basic-level matches for the purposes of these analyses, because this yielded a slightly more distinctive basic-level shell category.

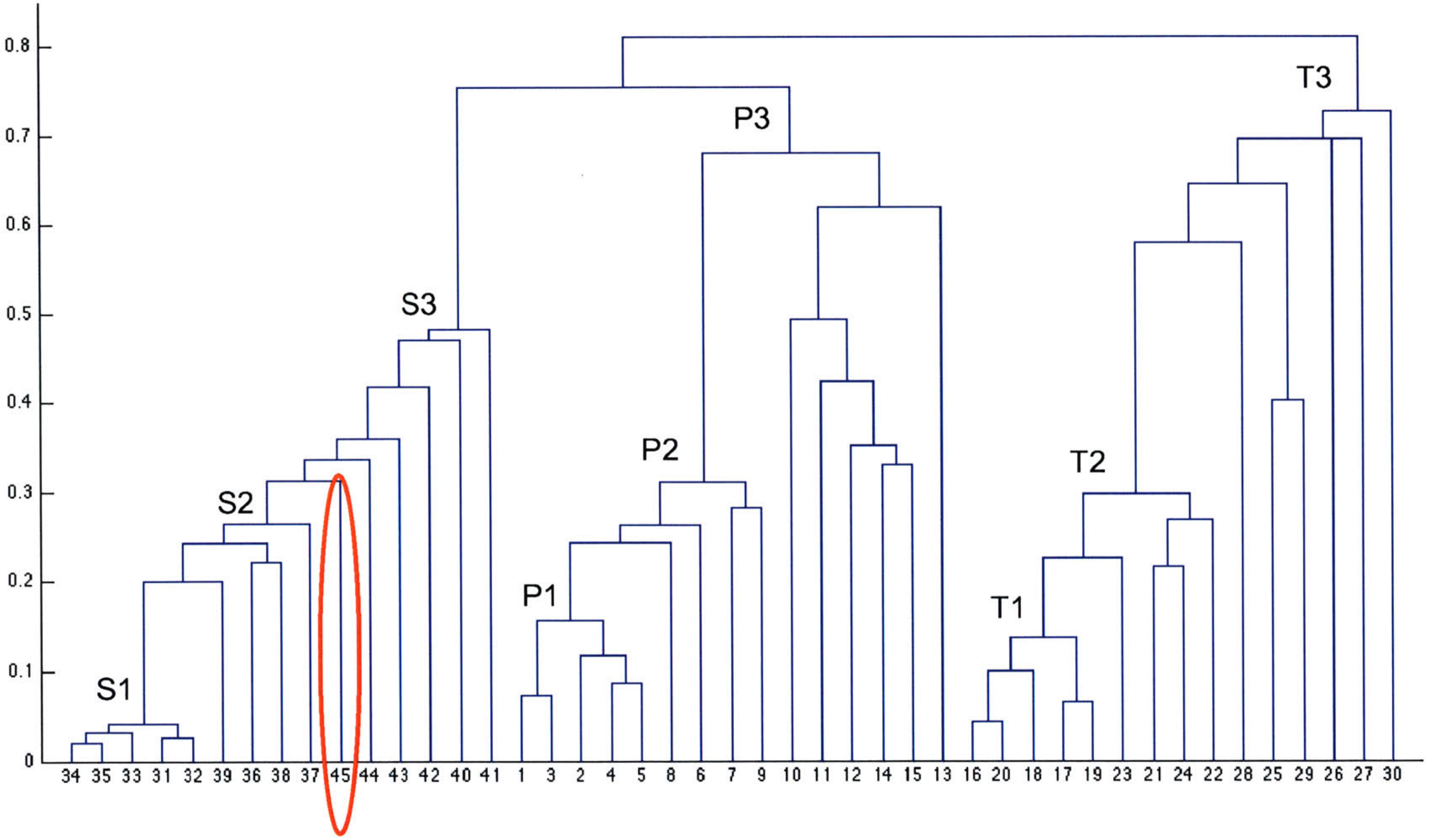


Figure 2-10: The taxonomic structure created from similarity ratings in Experiment 2. The circled object is Object 45, which has shifted position in the tree relative to the Experiment 1 results.

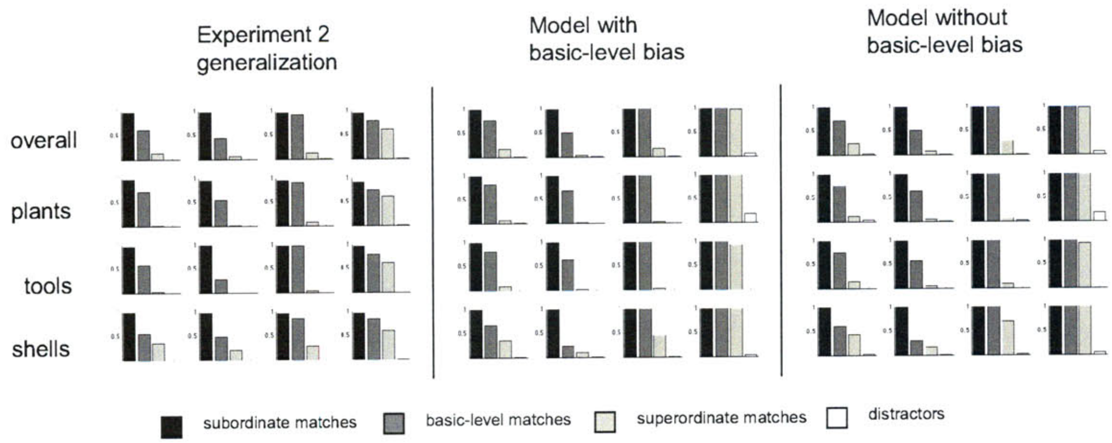


Figure 2-11: The Experiment 1 word learning data reproduced next to the model results, with and without basic-level bias.

The taxonomies from the two experiments had very different branch lengths for each of the supernodes (see Table 2.1). Nodes corresponding to word labels were less salient to participants before they participated in the word-learning phase. After word learning, the branch lengths of these supernodes showed an average increase of 0.053 — a 32% increase over the mean branch length of the supernodes prior to word learning. In the shell category, it appears that the basic-level category only became salient to participants after word learning occurred (compare node S2 in Figure 2-6 and Figure 2-10).

Since people show a strong distinctiveness bias in guessing what new words mean, the different branch lengths pre- and post-word learning correspond to different prior probabilities about the meanings of new words. Thus, word learning can influence thought by influencing how word learners construct a hypothesis space. This structured representation of objects could also serve as a hypothesis space for other category learning (e.g., property induction tasks: "If blickets have an omentum, what other Gazoobian objects are likely to have an omentum?"); noun-learning can potentially influence cognition even beyond the bounds of future word learning.

Modeling results

The model fit was not substantially altered by the different order of the experimental phases, though the best fitting parameters were different for Experiment 2 than for Experiment 1. In evaluating the model, we once again found that a model with a very mild basic-level bias provided the best fit for the word-learning data. We obtained a fit of r = 0.9788 with a basic-level bias parameter of $\beta = 1.1$, and a likelihood parameter of $\epsilon = 0.15$. The model predictions vs. the human generalization data are shown in Figure 2-11. Without a basic-level bias, the model fit was $r = 0.9773$, with an optimal ϵ value of 0.30.

The model fit for both Experiment 1 and Experiment 2 was very good. Here, again, however, we saw the main failing of the model was in predicting human generalizations to superordinate level matches. Presumably there is a great deal of uncertainty over tree structure in both of the experiments.

It may seem initially surprising that the model fit was essentially the same in both experiments, given that word learning did affect the prior probabilities for various hypotheses. However, this is less surprising given that the model combined prior probability with likelihood – especially in the three example trials, the evidence outweighed the prior probabilities. Also, according to our account, the word learners in both experiments started with a hypothesis space close to that gathered in Experiment 2 (the pre-linguistic taxonomy) and over the course of the word learning experiment shifted to a taxonomy close to that of Experiment 1. Therefore, both trees provide accurate representations of the learners' prior probabilities only for a portion of the experiment.

Overall, the Bayesian model predicts word learning data well in both Experiment 1 and Experiment 2.

General discussion

Even in an unfamiliar domain, word learners extend new words based both on prior beliefs and the evidence they have seen. XT's Bayesian account of word learning shows how word learners generalize based on limited evidence, choosing between overlapping hypotheses, whether the objects are familiar or novel. The Bayesian model also reveals some differences between adult word learners learning about novel vs. familiar objects: people appear to be more uncertain of the underlying category structure in the novel Gazoobian domain, and do not show very much of a basic-level bias.

Beyond extending XT's model, this chapter also investigated the origins of the hypothesis space. Both perceptual features and words contribute to the underlying category structure inferred by word learners. The Bayesian modeling approach clarifies how language affects learners' prior beliefs about word meanings and potentially other categories.

Handling uncertainty over tree structures

I discussed the fact that participants appeared uncertain about taxonomic structure, in both Experiment 1 and 2. Modeling word learning with a Bayesian model that took this structural uncertainty into account could potentially confirm this hypothesis and provide a better predictor of learners' generalization of words in an unfamiliar domain. Teh, Daumé, and Roy present a Bayesian agglomerative clustering model using a Kingman's coalescent prior; this model allows inference of a posterior distribution across all possible trees (2008). Using participants' similarity judgments to construct a probabilistic distribution over tree structures instead of a single tree could yield significant improvements over the current model, particularly in predicting generalization to superordinate items where the uncertainty is greatest.

As an alternative approach, making the relationship structure between objects more explicit should yield better fits between the single-tree Bayesian model predictions and adult generalizations in a novel domain. Explicit teaching of the designed stimulus categories prior to word learning, or implicit teaching – by, for instance, highlighting categories by teaching that the category members share a property (e.g., "these objects all have an omentum") – should decrease participant uncertainty about the tree structure.

Additionally, the question remains of how special language is in shaping the hypothesis space. How much of the Whorfian effect demonstrated in Experiment 2 was due simply to increasing familiarity with the objects? How much was due to the fact that the words were highlighting specific categories, but not specifically to the fact that they were words? These questions could be addressed by replacing the word learning phase of the experiments a familiarization task or a property induction task and observing the effects on the hypothesis space.

The development of word learning biases

The Bayesian approach offers a way to answer specific questions about word learning. Rather than just postulating and arguing about constraints, biases, and beliefs may affect word learning, the framework allows us to implement different alternatives and see which ones best predict word learning behavior. In addition, using Bayesian models we can quantify the strength of different biases.

The developmental trajectory of the basic-level bias, as discussed in Experiment 1, is highly speculative. However, by using this modeling paradigm and experimental paradigm on a wide age range of participants, in both novel and familiar domains, there is the potential to find out more about when and how the basic-level bias comes into play.

Language shaping mental representations

The Bayesian framework also offers a method for formalizing and quantifying the effects of language on thought. Recent work has modeled the effect of language (e.g., Dilkina et al., 2007), but this is the first computational model offering insight into the way that language can help form and shape the hypothesis space. Future work could explore whether and to what extent word learning changes learners' property induction hypotheses, as I speculated above. Additionally, whether or not word learning can cause the hypothesis space itself to change structure is still a hotly contested question (Bloom & Keil, 2001); using the same experimental and modeling paradigm with different sets of evidence (e.g., trying to teach participants words for disjoint concepts like *Dalmapeño*) could shed some light on whether this is possible and how much evidence is required.

Conclusion

In this chapter, I have explored how probabilistic methods and categorization can be joined together to explain how people learn words in an unknown domain. I have also presented an account of and how language can affect the probability of learners generalizing to a particular category in the future.

This chapter has addressed category induction when the words involved are labels for object categories. Chapter 3 addresses how people induce combine word meanings and induce the meaning of the resulting phrase. In particular, I look at adjectives like *tall*, which lack a stable extension alone, but instead gain their meaning by modifying object categories. I will be exploring how statistics and categorization can be jointly applied to model this class of words and the phrases they create.

Chapter 3

Composing word meanings

The problem of understanding word combinations

How do people judge whether a phrase, composed of multiple words, refers to an object? If I start using a novel word like *blicket* to refer to objects, you may have to hear it used a few times before you are fairly certain of its meaning (as discussed in the previous chapter). However, if I use a noun phrase you've probably never heard before, like *extremely big green mice*, you can judge whether a particular object falls into this category without having to first see lots of examples of how the phrase is used. This is the power of the compositionality of language: by knowing the meaning of the individual words, and some general rules of how words combine, people can understand phrases.

Cognitive science offers some insights into how people know what some individual words mean, such as the nouns discussed in the previous chapter. But we lack a comprehensive computational framework for describing how semantic composition occurs so that people can understand bigger chunks of language. While some attempts have been made to describe how nouns phrases are modified by adjectives like *green*, which refer to specific categories themselves (things that are green) (e.g., Zadeh, 1975, Huttenlocher & Hedges, 1994), other modifiers, like *big*, remain espe-

The work in this chapter was carried out in collaboration with Noah Goodman, David Barner, and Joshua Tenenbaum.

cially mysterious. *Big* does not refer to a category of similar items in the world – what counts as big depends entirely on the context. A *big mouse* is big for a mouse, and a *big building* is big for a building. Because of this puzzling context dependence of the adjective's extension, words like *big* get at the heart of the problem of compositionality. *Big* is an example of a *gradable adjective* – an adjective that orders objects along a particular dimension of measurement (e.g., size). Gradable adjectives can be used in comparative phrases like *bigger than* and *as big as*, and they all share the property of depending on context for interpretation.

In this chapter, I will be looking at gradable adjectives and modeling how they interact with noun phrases to form new meanings. I will show that gradable adjectives can be understood as statistical functions that operate over one category (described by a noun or noun phrase) to yield a new modified category. As Chapter 2 explored and explained graded judgments of category membership for the categories referred to by nouns (explaining how people answer a question like "How confident are you that this object is a blicket?"), this chapter models graded judgments of category membership for phrases ("How confident are you that this object is a *big* blicket?")

Using gradable adjectives as a starting point, I will discuss the broader implications for understanding semantic compositionality. This work provides insights about a number of major questions in different fields of cognitive science:

- *Vagueness and formal semantics.* The following puzzle of induction is called the Sorites Paradox: if you have a heap of sand and you remove a single grain, it is still a heap. Yet at some point, you are left with a single grain of sand, which is clearly not a heap. Similarly, if you have a clear example of a big chair, and you have an infinitessimally smaller chair, it is still a big chair. Yet at some point it will not be. How can this apparent paradox be computationally defined and understood? Can we quantify the types of judgments that people give in these situations? Can we explain how gradable adjectives can have a consistent meaning when their extensions vary in different situations?

- *Mental representation and processing.* What do people represent about categories? What do they represent about word meaning? How do they combine word meanings?

- *Innate knowledge and word learning.* What do people know about potential words meanings when they start learning language? How does this change over time?

Following the empirical sections of this chapter, I will return to all of these topics in the General discussion.

A clarification, before we begin, of the scope of the models in this paper: In order to use a gradable adjective, two steps must take place. First, the *comparison class* (Hare, 1952), or the objects being compared in a given context, must be identified. This category is defined in part by the noun or noun phrase that the gradable is modifying. However, it may also be affected by factors such as linguistic context (e.g., if particular trees have been under discussion previous to using the phrase *tall tree*), the proximity of items to one another, or items' perceptual similarity. Once the comparison class is determined, the second step of gradable adjective use is running the function that takes the comparison class (e.g., trees) and yields a subcategory (e.g., tall trees). In this chapter, I will not be addressing the first step, but instead proposing a statistical model explaining how the second step occurs. Past approaches to gradable adjectives and to compositionality more broadly have also focused on how categories are modified once the relevant category has already been identified.

Past approaches to compositionality and gradable adjectives

Models of acquisition of compositional semantics have been proposed previously, making use of cross-situational learning (Piantadosi, Goodman, Ellis, & Tenenbaum, 2008) or training data (Zettlemoyer & Collins, 2009). However, the goal of these approaches is to map natural language sentences onto logical forms. While understanding the general role that words play in a sentence is an essential part of understanding compositionality, it is also vital to understand how those combined phrases and sentences pick out referents in the world. I turn now to work that attempts to address that issue.

Intersective adjectives are those like *green* or *square* that define the intersection of two categories when combined with nouns: *green mouse* refers to things that are both mice and green; *square table* refers to things that are both square and tables. Much of the work that has been done on compositionality has addressed how intersective adjectives are combined with nouns to pick out a new category. Some have suggested that this process is, in fact, just the process of taking the intersection of two sets, or two fuzzy sets, to arrive at a new category (Zadeh, 1965, 1975; Rosch & Mervis, 1975; Kamp & Partee, 1995). However, the fuzzy set intersection proposal is insufficient to explain people's generalizations about the typicality of category members, and also yields logical contradictions for some inputs (Osherson & Smith, 1981). Huttenlocher and Hedges (1994) outline an alternative statistical model for combining two such intersective categories: if the categories are defined as probabilistic distributions (with each object belonging to a category with a given probability), then the combination of the categories is the joint distribution. This addresses issues of graded category membership, and also results in typicality judgments for the combined category that more closely match human judgments than other proposed methods such as set intersection. E. Smith, Osherson, Rips, and Keane (1988) propose a model for modifying prototypical categories represented with noun phrases by adjectives or intensifiers to yield a new graded categories. However, the account only addresses attributes that are inherent to an object regardless of the context it is in.

While the work on intersective adjectives provides some insight into semantic compositionality, it does not translate to *subsective adjectives* like the gradable adjectives discussed above: adjectives that pick out a subset of a category defined by a noun phrase, without ever defining a context-independent category themselves (Partee, 1995). Linguists have studied subsective adjectives and explained to some extent how they behave, without providing a computational account of how they function in a particular context. Gradable adjectives like *tall* have been described as functions mapping objects in the world onto a set of degrees by which they can be ordered and compared (Cresswell, 1976; Kennedy, 1999). The ability to compare objects based on this mapping is what makes adjectives *gradable*, as opposed to other adjectives such as *dead* – a tree can be *taller than* a flower, but it cannot be *deader than* a flower. As noted by Klein (1991) and Bierwisch (1989), using gradable adjectives must require comparing the objects in the comparison class in terms of either the objects' values on the appropriate dimension, or possibly the ordering of the objects (without taking the actual values into account). However, the details of such a function have been left unspecified. Bierwisch suggests that perhaps all members of a category that are above average value along the dimension in question (e.g. taller than the mean height for the category) would be tall (e.g.).

Empirical evidence indicates that people are, indeed, sensitive to context in judging which items are picked out by gradable adjectives – and that this is true of children as well as adults (Barner & Snedeker, 2008; Syrett, Bradley, Lidz, & Kennedy, 2006; Ebeling & Gelman, 1994; Gelman & Ebeling, 1989;Sera, Troyer, & Smith, 1988)[1]. Children aged 2-4 can judge whether a typical object like a *mitten* is *big for a mitten* (Ebeling & Gelman, 1988) or whether a button is the right size for an adult or a baby (Sera et al., 1988); they can also understand that an object may be big in some contexts (a shot glass as a glass for a doll) but not others (a shot glass as a glass for a human) (Ebeling & Gelman, 1994). Barner and Snedeker (2008) showed that children aged 4 are sensitive to the distribution of the heights of objects in the context in judging which objects are *short* and *tall.* Adding extra very tall items or very short items to a distribution of objects causes the children to change their judgments about whether some of the items in the middle of the height range are tall.

Despite this empirical evidence, few proposals have come out of psychology either as to how people making judgments about when a gradable adjective applies to a give object. Barner and Snedeker suggest that the metric of comparison used to select the *tall/big/loud*/etc. items in a set could involve the mean of the values (as proposed

[1] Syrett et al. (2006)'s work shows that people's judgments do not appear to be context sensitive in the special case of *Absolute gradable adjectives* such as *full*, where there is an endpoint to the scale (a glass cannot be fuller than *completely full*). Most of this chapter addresses *relative gradable adjectives* which do not have such an endpoint, but I will revisit this issue in the General discussion.

by Bierwisch), but might instead rely on the number of item in the set – e.g., all objects taller than most of the objects in the set might be judged tall. In the field of artificial intelligence, Gold (under review) suggests that *tall* can be learned as picking out those items which are some number of standard deviations above the mean, and similarly for other gradable adjectives.

While a few authors have thus sketched possible functions that gradable adjectives could perform when combined with a category, no models have been empirically tested. In the next section, I formally present a number of possible statistical models of gradable adjectives. I then empirically assess their performance at predicting judgments of *tall* and *long*.

Before I begin this process, it is worth noting that some linguists have, in the past, resisted statistical accounts of linguistic representation. However, these debates have focused on very different questions involving language: (1) whether strict syntactic rules can be replaced with probabilistic rules, and (2) whether statistical learning processes can be used to learn syntactic systems. This work does not address either debate. Instead, I fill in one component of lexical meaning that has previously been left unspecified by linguists and psychologists alike. I show how a system that takes in a category as input and yields a new category must consist of a statistical process, and I empirically evaluate candidate models.

Models of gradable adjectives

This section outlines two types of approaches to modeling gradable adjectives. The first approach suggests that, given a particular context, people search for the threshold above which objects are *tall*[2]. The threshold could be based on the mean height (e.g., one standard deviation above the mean) or other parameters relating to the object heights (e.g., the top quartile of object heights), or it could be determined solely by the ranking of the objects and not their actual heights – e.g., the tallest third of

[2] I will use *tall* for many examples of gradable adjective use in this section, however, all of the models could potentially describe gradable adjective.

the objects are considered tall. The approach of looking for a threshold corresponds directly to the idea that many uses of gradable adjectives establish a *standard of comparison* against which all the objects are judged (see Klein (1991) and Kennedy (2007) for further overview).

An alternate approach proposes that, given a particular category of objects, people seek *the tall ones* as a specific subcategory. Under this approach, people first find the most likely grouping of the objects and then label the tallest group as *the tall ones.* One reason to suggest such an approach is that it corresponds to the idea that noun phrases represent categories, and identifying referents of an adjective-noun combination might reasonably be seen as a categorization problem. Additionally, people are good at categorizing objects and features in a wide number of domains; this approach could make use of domain-general cognitive architecture. A final reason is that some contexts involve irregular distributions of items, many of which naturally make sense to think of categorically. For instance, people trying to judge which items at the Seattle Woodland Park Zoo are *big animals* will notice a gap in size between the elephants and the next smallest species of animals. The lions are unlikely to be grouped together as either being big or not big, even if they are split across one standard deviation above the mean (an example threshold value) – because the lions are generally more like other lions in size than they are like other species of animals in size. Thus there are times when a categorization approach is particularly intuitive.

Another important distinction I will be discussing in this chapter is the separation between *parametric* and *non-parametric* models. Parametric models assume that the observed data comes from some underlying probability distribution (e.g., a normal distribution); they define values like thresholds in terms of the parameters of these distributions, like means and standard deviations. Non-parametric models make no such assumptions about data distribution. A model that says any object taller than 17 feet is a tall object (no matter what the context) is non-parametric. So is a model that sets the tallness threshold to half the height of the tallest object. All the category-based models we will address here make no underlying assumptions about the probability distribution of the data, and are therefore non-parametric.

I will be formalizing and comparing models from both general families: threshold-based models and category-based models. Some of the threshold-based models I propose are parametric, while others are not. Each proposed model is a statistical function that takes in a particular comparison class and yields a graded category of the *tall* members of the category given that context. That is, all models answer the same basic question: "Which Xes are tall Xes?"

In describing both types of models formally, we will be referring to the collection of items in the context, or the comparison class, as C.

Threshold-based models

A threshold function T takes C as input; its output is a real value along the dimension of interest. For *tall*, the value $T(C)$ is interpreted as a height, and anything greater than this height is judged tall.

Human judgments are not so precise that people always agree exactly on the answer to questions like "Within this orchard, which are the tall trees?" Judgments may vary somewhat between subjects and even within subjects, if a person is asked about the same context at different times. To account for this, the threshold is noisy: it is a normally distributed variable with a mean at $T(C)$ and a variance governed by noise-width ϵ. Then the probability of an item x being tall is the cumulative probability that its height $h(x)$ is greater than a random variable drawn from a Gaussian distribution centered at $T(C)$:

$$P(x \text{ is tall}|C) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{h(x) - T(C)}{\epsilon\sqrt{2}}\right)\right], \tag{3.1}$$

where erf is the Gauss error function. See Figure 3-1 for an illustration of how the two free parameters affect the model predictions.

I implemented and compared the following threshold functions (see Appendix A for the mathematical descriptions of each model):

- *Absolute height (AH)*: No matter what the context is, all items taller than a fixed reference height k are tall.

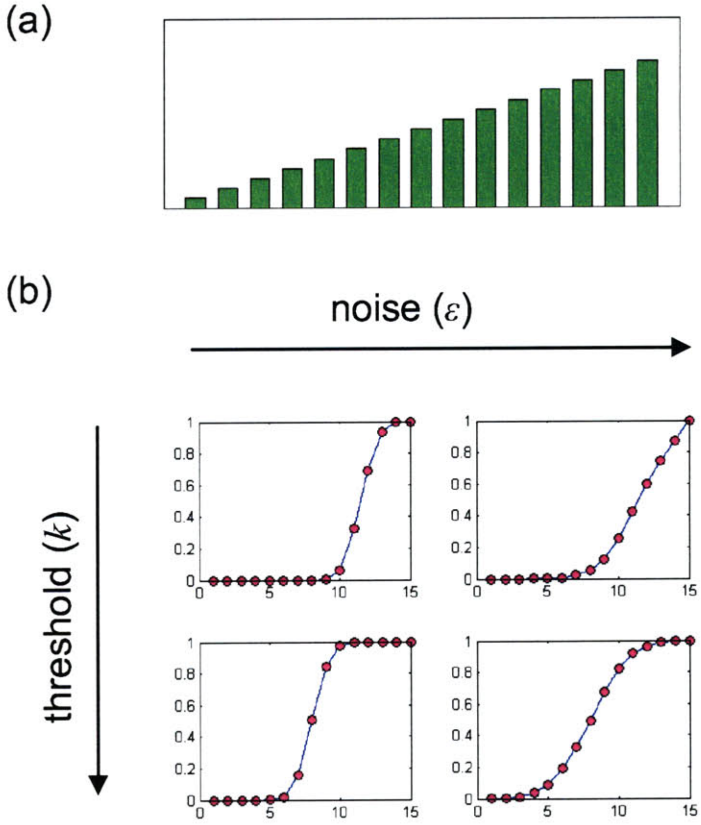


Figure 3-1: The effects of the two free parameters on the predictions of a threshold model about which items in a distribution are tall. (a) shows a distribution and (b) shows the corresponding predictions of a threshold model (RH-R) with different settings of the free parameters. k controls the location of the threshold, while ϵ controls the noisiness of the threshold. The x-axis of each of the graphs represents the items of the distribution shown in (a). The y-axis of each of the graphs shows the probability that the item is tall.

- *Absolute number (AN)*: No matter what the context is, we line up all the items in order by height, and then take the tallest k items as tall.
- *Unique percent number (UAN)*: As in AN, but computed based on only one object of each height (in other words, only unique heights are compared).
- *Percent number (PN)*: Items are ranked by height, and then take the tallest $k\%$ of the items as tall.
- *Unique percent number (UPN)*: As in PN, but computed based on only one object of each height (in other words, as with UAN, only unique heights are compared).

- *Relative height by range (RH-R)*: Any item within the top $k\%$ of the range of heights is tall.
- *Relative height by standard deviation (RH-SD)*: Any item with a height greater than k standard deviations above the mean is tall.

The first two models do not depend on the context, but serve as a baseline comparison for the models that are sensitive to the items in the input distribution.

Category-based models

A category-based model takes the comparison class C as input and searches for likely categories among the items based on the clustering of the item heights. Its output is a value between 0 and 1 for each item x in C indicating how probable it is that x is in the tallest category of items.

I created a category-based model called the Cluster Model (CLUS)[3]. CLUS is a modified version of the infinite Gaussian mixture model (Rasmussen, 2000), a modern version of Anderson's rational model of categorization (1991) . The infinite Gaussian mixture model assumes that the observed data (in the case of generative adjectives, the heights of the objects in the context) come from an unknown number of normal distributions, each generating one cluster of data. The model allows inference backward from the data to determine the probability that each data point came from a particular underlying distribution.

[3] I will use *cluster* to mean groups of objects considered by the model as possibly belonging together, and *category* to mean a linguistic referent like *the tall Xes.* The tallest cluster of items found by the CLUS model is the category of tall items

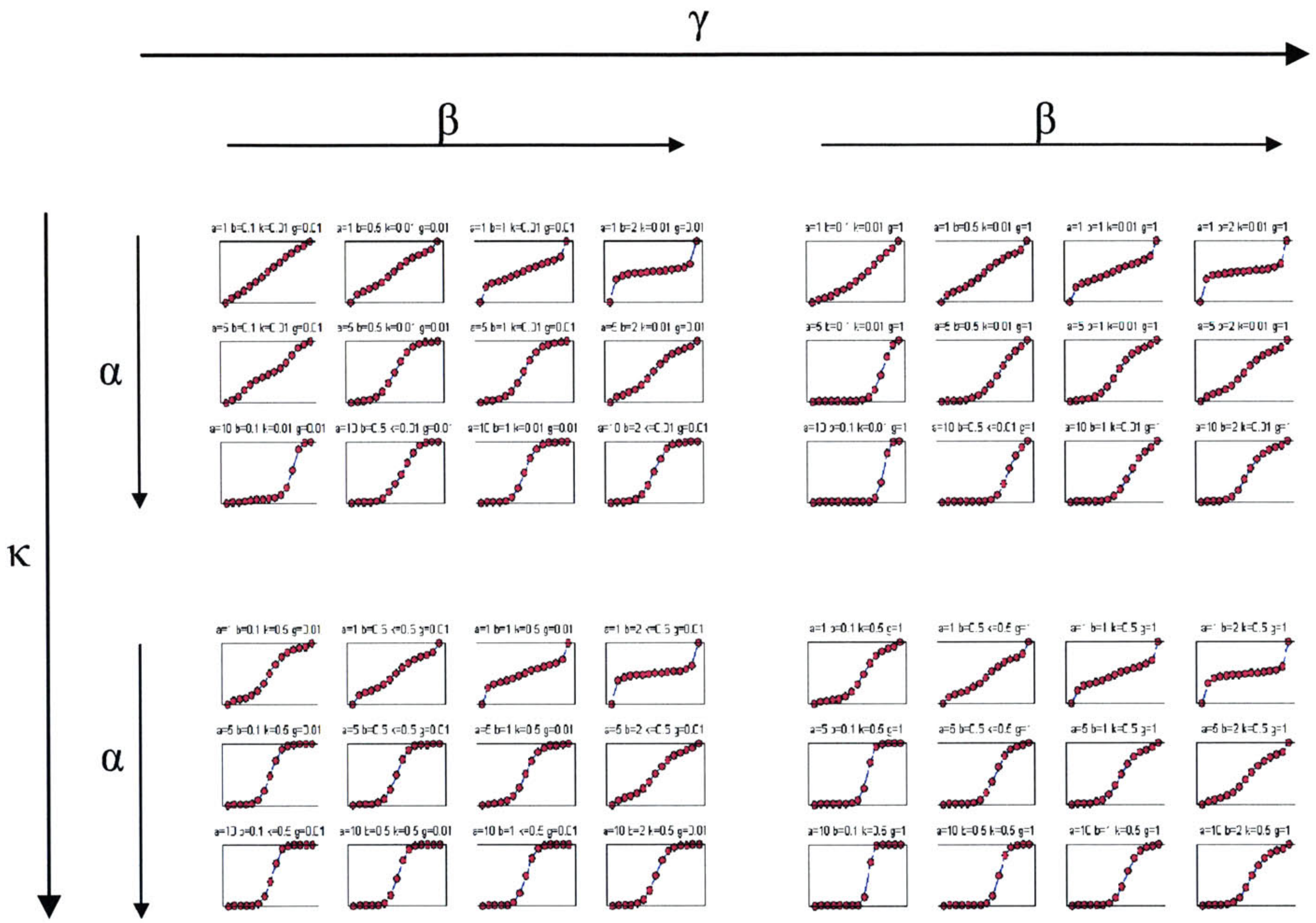


Figure 3-2: The effects of the four free parameters on the CLUS model predictions about whether the items shown in Figure 3-1(a) are tall. γ controls expectations about the numbers of items in a cluster. Jointly, the other parameters control the shape of the prior probability distributions for the cluster means and variances. See Appendix A for details.

The original Rasmussen (2000) model would allow for the inference of overlapping clusters of items. However, gradable adjectives do not involve such overlap; if one item is judged tall, then everything with a greater height is also tall, and short items are never taller than tall items. Therefore, I restricted the model to only consider assignments of items to clusters that yielded contiguous sets along the dimension of interest (i.e., no interspersing of tall and non-tall items). I additionally modified the model such that it only considered assignments of objects to clusters such that there were at least two groups of items (so that for any group of two or more items, there is always at least one tall item and at least one item that is not tall)[4].

How is the probability of a given assignment of objects to clusters, Q, determined? An intuitive explanation is given here; for further detail and a formal mathematical specification, see Appendix A. As is standard within a Bayesian framework, the probability of Q given the context C is composed of two factors: the prior probability of Q and the likelihood that Q generated the data in C.

The prior probability, $P(Q)$, is the probability that the number of clusters in Q would occur, and that those clusters would contain the number of objects that they do. For instance, it is relatively improbable, though not impossible, that each item would be in its own cluster. The likelihood of the data in the context given the assignment to clusters, $P(C|Q)$, depends on how probable it is that that distribution of data would result from the given assignment. For instance, if a context contains five objects which are 20 inches tall and one object is 2 inches tall those data probable were not generated from an assignment that clustered the objects as follows: 20, 20, 20 inches and 20, 20, 2 inches. The mean of the second cluster is 14 inches; it is more likely that an underlying category with a mean of 14 inches would have generated objects with heights clustered around the mean, rather than with a huge gap in them. Therefore, the data are relatively surprising given this assignment, and the likelihood is low.

[4] When asked to find *the tall Xes*, people are strongly pragmatically encouraged to divide the set of Xes. Except in rare cases, which were not represented by any of the empirical contexts addressed in this chapter, people are unlikely to select all the items in a set or none of the items in a set as *the tall Xes*

By combining the prior and the likelihood, we can calculate posterior probability $P(Q|C)$:

$$P(Q|C) \propto P(Q)P(C|Q) \tag{3.2}$$

This results in the following probability that an object is tall – or whether it is in the tallest cluster, q_{tall}:

$$P(x \text{ is tall}|C) = \sum_{Q} P(x \in q_{\text{tall}}|Q)P(Q|C) \tag{3.3}$$

In other words, the probability that a particular item x is tall given a particular assignment Q is 1 if the item is in the tallest cluster and 0 otherwise. To obtain the overall probability that x is tall in a given context C, we average every possible assignment Q weighted by its posterior probability given C.

The CLUS model contains four free hyperparameters governing the prior expectations about cluster size (in terms of number of objects), mean, and variance (in terms of height object height). They are explained in detail in Appendix A, but their effects are illustrated in Figure 3-2.

In addition to the CLUS model, I implemented a simple baseline comparison model called the Maximum Margin (MM) model. This model found the biggest gap in item heights and categorized every item taller than the gap as a *tall X* and every item shorter than the gap as not being a *tall X*. In cases of ties, the tallest of the gaps was used as the category boundary.

Model predictions

Before moving on to empirical tests of the models, it is worth gaining an intuitive understanding of where they are similar and where they differ.[5] Figure 3-3 shows three different distributions of items (represented as green rectangles of varying heights) and

[5] Varying the parameters of any given model can cause its predictions to vary greatly for the same context; this section assumes some typical parameter settings for each model, as empirically determined in the following experiments.

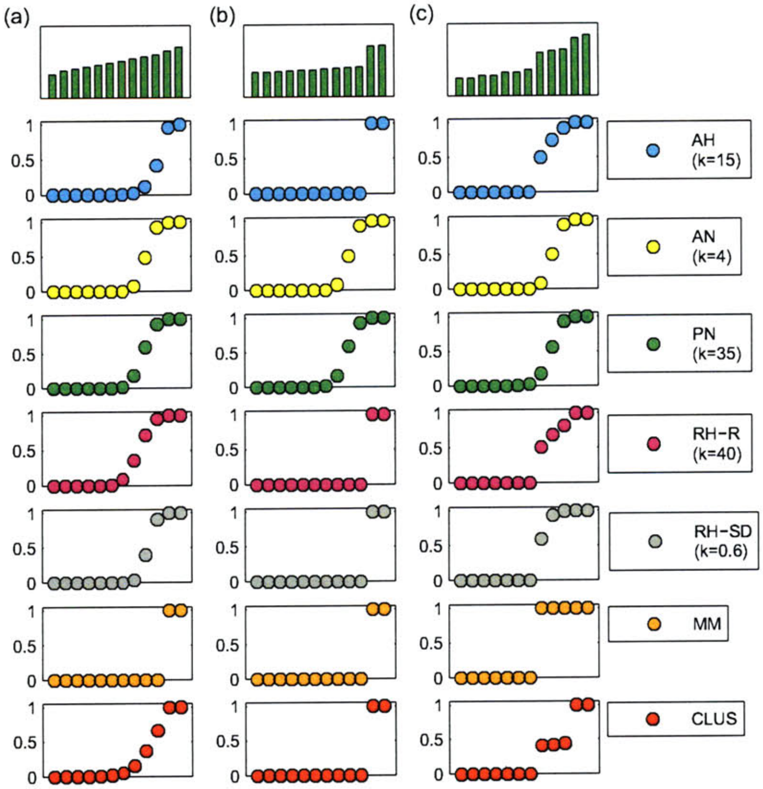


Figure 3-3: Each column shows a distribution of rectangular items of different heights at the top. The graphs below show the probability that each item is tall, according to each model. Model name and threshold parameter setting specified to right of each graph. (Cluster model shown uses typical hyperparameter settings.)

the corresponding predictions of the different models about which the probability that each is a tall item.

In the first distribution (left), the items are approximately uniformly distributed in height. All the threshold-based models make similar predictions; the threshold that they calculate in the given context is similar, as is the sigmoidal drop-in probability due to the noise parameter.

Interestingly, while the MM model shows an abrupt drop from tall to not tall, the CLUS model shows a sigmoidal drop off similar to the threshold-based predictions. Why is this, given that the items all appear to belong to a single undifferentiated category? Because the CLUS model requires at least two clusters be found, the model cannot classify all the items as tall. Given that the data do not strongly point to any particular splitting of the items into clusters, the prior expectations about cluster size play a strong role in the outcome, and we get the graded drop-off around the most probable cluster size.

When, as in the second distribution (center), there are two extremely distinct clusters of items present, once again the models have very similar predictions. Here, the CLUS model is very certain which items belong in the tall cluster, and all of those tall-cluster items also fall well above the threshold in the threshold-based models that depend on object height (e.g., RH-SD and RH-R). The threshold-based models that instead determine the threshold according to the number of items (e.g., PN) still show the same pattern of prediction as before, because the number of objects has not changed.

The third distribution (right), is the first case where we see very different predictions between the threshold-based approaches and the CLUS model. Here, there are some items that are clearly short, some that are clearly tall, and some that are in between, separated by a gap from both other clusters. The threshold-based models show the same kind of sigmoidal drop-off in predicted tallness as seen in the previous distributions. The CLUS model, however, tends to group the mid-height items together, though it is uncertain about whether they are tall or not. Therefore, we see three near-plateaus in the probability predictions of the CLUS Model, in contrast to the gradual drop-off in most of the threshold-based models. The MM model shows plateaus as well, but there is no middle plateau (just tall and short).

This chapter evaluates these models in several steps. First, I look at whether whether any of these models are flexible enough to predict human judgments of *tall* in a wide variety of contexts. I compare the models' performance in predicting the judgments of adults and children. I also evaluate the models' predictions of judgments for another gradable adjective, *long.*

Experiment 1a: Judgments of *tall* across a wide variety of contexts

In this experiment, I present adults with a wide array of distributions of items of different heights. I ask them to find the tall ones, and compare their judgments to the predictions of the models.

Method

Participants

Participants were 181 adults, participating for payment via the online Amazon Mechanical Turk interface. Each participant saw and was asked for judgments about a single distribution of items, yielding 7-10 participants viewing each distribution in the experiment. [6]

Stimuli

The stimuli consisted of 21 distributions of items. Each distribution consisted of a set of green rectangles of varying heights, ranging in number from 9-54 items per distribution. The majority of the sets of items were sampled from one or two underlying Gaussian probability distributions, which varied in height range, mean, variance, distance between clusters, and number of items per cluster. Other sets of items were drawn from non-Gaussian probability distributions, such as uniform or exponential distributions. The distributions were designed to test the proposed models in a wide variety of situations. See Figure 3-4(a) for a representative set of distributions, and Appendix B for a full description of all distributions used.

[6] 210 requests for participants were put online, but after all trials were eliminated in which participants had disregarded instructions and answered questions about more than one distribution, 181 unique participants remained.

In all the Mechanical Turk-based experiments reported in this chapter, the majority of the participants were native English speakers, but some non-native English speakers participated as well. Because of the small number of subjects per condition, all subjects are included in the analyses in this chapter. Preliminary analyses indicate that the results for only the native English speakers are substantially the same.

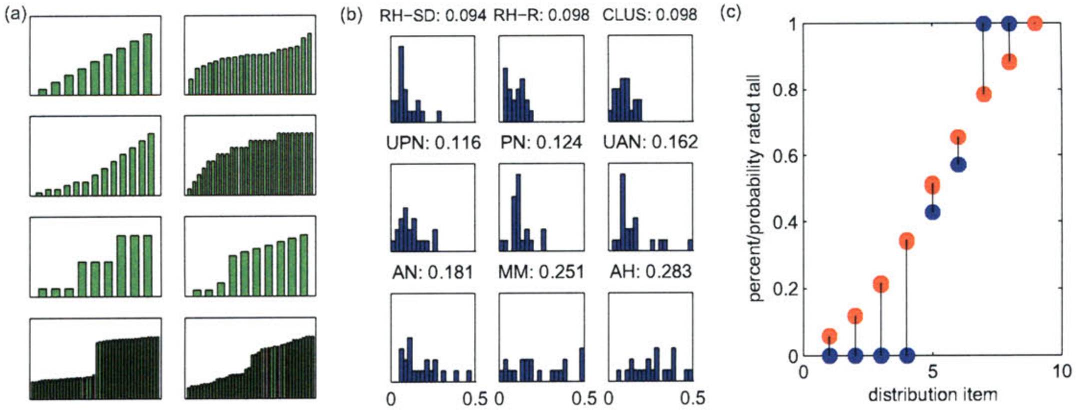


Figure 3-4: (a) A number of representative distributions used as stimuli in Experiment 1 (subjects saw these and other sets of objects in randomized order instead of sorted by height, as shown here). (b) Histograms demonstrating the performance of each model in Experiment 1, where success is measured along the x-axis in terms of mean absolute error (MAE) (see (c) for explanation of measure), with a performance of 0 being best and 1 being worst (for these histograms, all results worse than 0.5 are binned at 0.5). Each histogram represents the MAE of one model across all distributions, with the number of distributions receiving each score measured on the y-axis. The model name and the average MAE for that model is shown at the top of each histogram. The models are sorted from best to worst performance overall. (c) An illustration of how MAE is calculated for one distribution. For each item in the distribution, the distance between people's judgments (blue) and the model's predictions (red) is calculated (black lines) and the MAE is the average of all these distances.

The items were presented onscreen together, arranged in one of two pseudorandom orders. Each item was labeled both above and below with a number, starting with 1 at the left-hand side of the figure and increasing.

All distributions were shown within a 800 x 450 pixel frame. The top of the frame was set to be 1.3 times the height of the maximum item height within the distribution, so as to not have any of the items crowded too close to the top of the frame. The items were generated as bars in a Matlab bar graph, with the width of each item set to 60% of the possible bar width. Because the number of items varied within a static frame size, the absolute width of the bars varied.

Procedure

Each participant saw the following text:

> The green objects shown below are pimwits. Each pimwit is labeled with a number. Can you look at all of the pimwits and find the *tall* pimwits?

Beneath this, they saw the distribution of items. Underneath that figure, participants received the following instructions: "List the numbers of all the pimwits that are tall here." A text entry area allowed participants to fill in their list.

The nine model classes were run on all the distributions. For each of the threshold-based models, 2500 random threshold samples were drawn from the Gaussian distribution governed by k and ϵ. For each sample, all the items in the distribution were judged tall or not tall relative to the threshold. The overall probability that item x was tall was the proportion of samples in which x was taller than the threshold. For the CLUS model, the possible partitions of the distribution items into clusters were evaluated and the probability of each distribution item being in the tallest cluster was found as specified in the model description[7].

The free parameters were varied systematically across different instances of each model. I compared the model predictions to the answers given by the adults and found the parameter settings for each model that best fit the adult data.

Results

The average model performance varied widely across the distributions used in this experiment. The Relative Height by Range, Relative Height by Standard Deviation, and Cluster models performed best overall. The other models not only performed worse on average, but were also proven by specific points of failure to be insufficiently flexible to model adult judgments of *tall*.

For each distribution, I evaluated a given model's performance by comparing that models predictions about the probability that each item was tall to the percent of adults who labeled that item tall. I performed this comparison using a measure called mean absolute error (MAE) which measured the average error per item; an

[7] to reduce computational difficulty, a maximum of 4 clusters per distribution was allowed in this experiment; the change in the CLUS model's predictions when considering 5 or more clusters was found to be negligible for this set of distributions.

MAE of 0 indicates no error (model perfectly predicts adult judgments), while an MAE of 1 indicates maximal error. The method of calculating the MAE is illustrated in Figure 3-4(c).

I selected the best parameter setting for each model by averaging the MAEs between the adult judgments and the model probabilities across all distributions. Figure 3-4(b) shows a histogram of the MAE across all distributions for each of these best models. The models are ranked in order of average performance. All MAEs greater than 0.5 (very poor performances) are binned at 0.5 for the purposes of these histograms.

The RH-R, RH-SD, and CLUS models performed best; all had approximately the same average MAE. All of these models performed well on each individual distribution, with the exception of the RH-SD model's performance on a single distribution, which is discussed further below.

On average, the Unique Percent Number and Percent Number models perform only slightly worse overall; however, there are some distributions where both of these models perform very poorly, as shown in Figure 3-5(a). In each of these distributions, there are two clear clusters of items, and most adults judge all the items in the top cluster to be tall. However, the percent of items in the tall cluster varies between these two contexts. As a result, we can see a strong *bidirectional failure* in performance; on the left, the PN model predicts far too many items are tall, and on the right, far too few (UPN has the same problem in many cases, since it reduces to the same model as PN whenever the distribution values are unique). Because of this bidirectional failure, the PN threshold parameter cannot be adjusted in either direction to better predict all the adult results, and both it and the UPN model systematically fails to capture human judgments about what is tall.

Figure 3-5(b)-(d) also shows similar bidirectional failures for the Absolute Number, Maximum Margin, and Absolute Height models. Figure 3-4(b) demonstrates these models' poor performance on the overall distribution set. Because all of these models fail in both directions – sometimes calling too few items tall and sometimes too many – these models also cannot account for human judgments of tallness. This failure

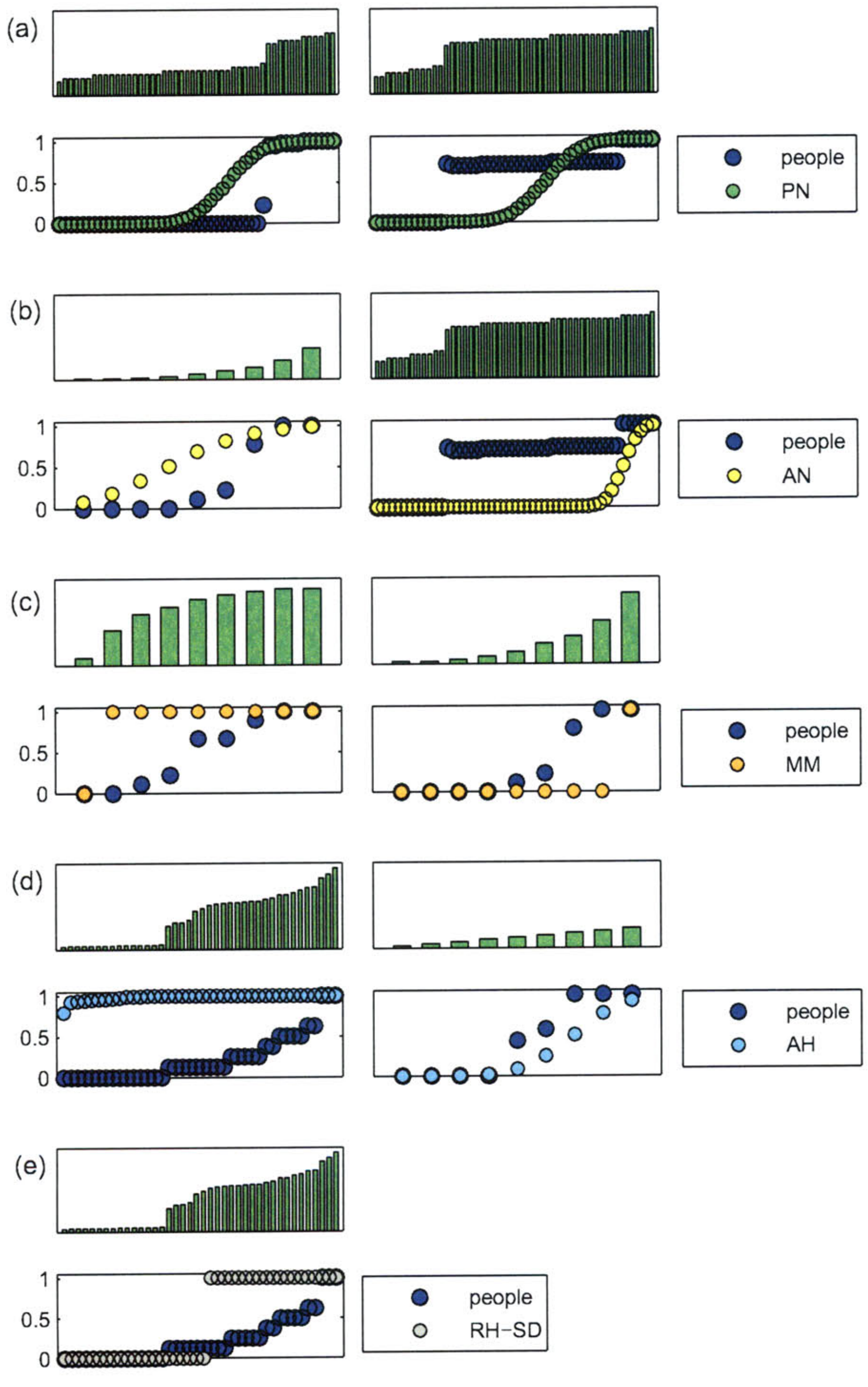


Figure 3-5: Several models failed to predict the tallness judgments of people in Experiment 1. (a)-(d) show four models exhibiting a *bidirectional failure* – they predict too many items being tall given the distribution on the left, and too few items being tall given the distribution on the right. In (e) we see that RH-SD also performed poorly in one of these directions for one distribution.

of the two absolute standard of comparison models (AN and AH) is consistent with previous empirical evidence that people are sensitive to context in making judgments of tallness. The failure of the MM model indicates that people are not performing this simple clustering heuristic in judging the tall items; if people are looking for categories of objects within a context, this is not their strategy.

Figure 3-5(e)) shows the one distribution where the RH-SD model did significantly more poorly than its general performance. In this distribution, there are many items that are taller than 0.4 standard deviations above the mean height (the optimal RH-SD threshold overall), and the RH-SD model labels them all tall. Adult judgments fall off far more quickly, and there are many items that very few people think are tall but which the RH-SD model predicts are tall with 100% probability. Despite this large divergence between model and human performance on one distribution, however, this model does not show a bidirectional failure on the distributions in Experiment 1a. Aside from this one distribution, the RH-SD model performs very well; is the failure shown in Figure 3-5(e) a fluke, or indicative of systematic problems for the model? Experiment 2 further investigates this model's performance in similar contexts, as well as further investigating the RH-R and CLUS models.

Overall, many of the models failed to match human judgments of tallness. The models that failed most dramatically were those threshold-based models which did not depend on the context in choosing a threshold, and also the Maximum Margin model. Models that depended only on the number of items in the context to set the threshold also systematically failed. However, all of the models that depend on the height of objects in a given distribution (RH-R, RH-SD, and CLUS) performed well overall. This suggests that, in assessing which items in a given context are tall, people are judging tallness in a way that can be modeled by a statistical function based on the heights of the objects in the category, and not simply based on the objects' rankings by height or a simple heuristic.

Before continuing to explore which models best describe human judgments, I address a potential worry from Experiment 1a. The widths of the items differed across some of the distributions. It is possible – likely, even – that inherent properties such as the height-width ratio of individual items affect judgments of tallness. Therefore, it may be questionable to evaluate the same parameter settings for a single model across distributions with items of different widths. Experiment 1b addresses this concern by investigating how much differences in item width affect adult judgments of tallness.

Experiment 1b: The effects of item width on judgments of *tall*

Method

Participants

Participants were 29 adults, participating for payment via the online Amazon Mechanical Turk interface. Each participant saw and was asked for judgments about a single distribution of items, yielding 9-10 participants viewing each distribution in the experiment. (One participant's second set of results was eliminated after they participated twice.)

Stimuli

The stimuli consisted of three versions of the same distribution: nine items evenly distributed in height. The rectangular items shown were red instead of green, but in all other ways the stimuli were the same as in Experiment 1a. As in Experiment 1a, each subject saw each distribution in one of two pseudorandom orders. One distribution was identical to one of the distributions used in Experiment 1a except for the coloring; the other distributions were identical to the first, except that in one case the bars were skinnier (half the width of the original bars), and in one case they were wider (3/2 the width of the original bars).

Procedure

The procedure for each participant was the same as Experiment 1a, except that the items in the distributions were called zavs and were red instead of green.

No modeling was performed for this experiment.

Results

I compared adults' judgments across all three conditions. There was no significant effect of object width on judgments of which items were tall.

I performed χ^2 tests on each pair of conditions, comparing the number of people who chose each possible number of items as tall. I performed a 2x5 χ^2 test in each case, as all participants selected between 1 and 5 items out of each distribution[8]. None of the differences were significant – medium x skinny ($\chi^2 = 1.95$, $p = 0.75$), medium x wide ($\chi^2 = 1.38$, $p = 0.85$), or wide x skinny ($\chi^2 = 2.41$, $p = 0.66$).

Varying an item's width by as much as a factor of three did not significantly influence people's judgments of *tall*. It is certainly possible that inherent properties of items such as width, shape, or proportion influence judgments of their tallness, however, and we should be cautious in generalizing from this small population of adults making judgments about a single distribution of items. Additionally, the widths of items in Experiment 1a varied by up to a factor of 6, so further work is necessary to be completely convinced that generalization over that range of widths is reasonable. However, the fact that people's judgments of *tall* are robust to the extent shown here suggests that the potential worries about Experiment 1a are not a concern.

Experiment 1a showed how adults' usage of *tall* varied based on context as a function of the heights of the objects and identified several statistical models that predicted adult judgments well. Do these models also predict the usage of children who have only recently learned the word? I investigate this question in Experiment 1c.

[8] no participants selected non-contiguous sets of items as tall.

Experiment 1c: Children's judgments of *tall*

Method

Participants

Participants were 124 4-year-old children. I used data gathered by Barner and Snedeker (2008) from 16 children about each of 3 distributions (48 children total), and additional data gathered from 10-18 children about each of 5 other distributions (76 children total).

Stimuli

The stimuli consisted of 8 of the distributions used in Experiment 1a. Because of limitations in the attention span of children and the stability of the stimuli, only the simplest distributions, with relatively few items and relatively small ranges of heights, were used (see Appendix B for details). In contrast to the adult experiments, children saw sets of physical items. These items were pink cylinders $\frac{1}{4}$ inch in diameter, varying in height from 0.33 to 9 inches. Each cylinder was decorated with a face and hair.

Procedure

Children were asked to judge categories by placing items within a red circle. Children first participated in a pretest to insure that they understood how to categorize items in this manner. Each child was given a set of toys containing a number of items from multiple categories of fruit. They were asked to perform categorization tasks such as putting all the bananas in the red circle until they had done so with three consecutive successes.

For the experimental trials, children saw the distributions of novel cylindrical items lined up in one of two pseudorandom orders. Children were told that the items were *pimwits* and prompted to touch each of the pimwits. They were then instructed, "Can you look at all of the pimwits and find the tall pimwits and put the tall pimwits in the red circle?" For further details of the procedure, see Barner and Snedeker (2008).

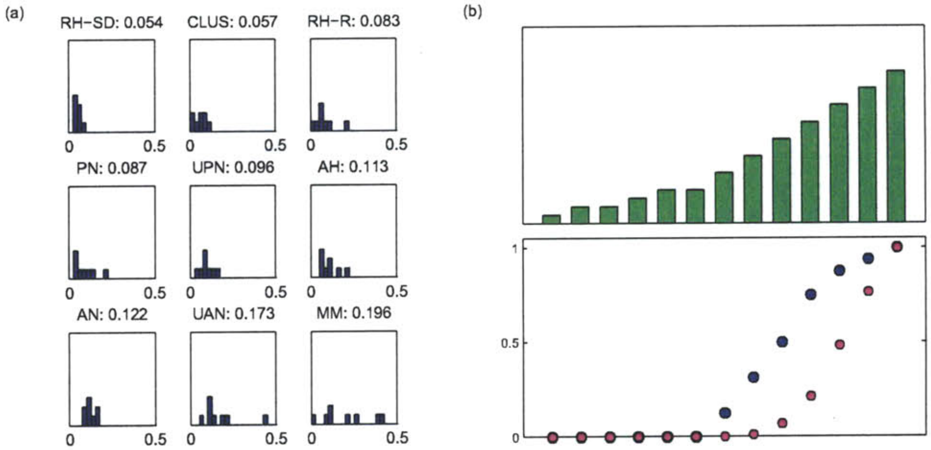


Figure 3-6: (a) Histograms demonstrating the performance of each model in Experiment 1c, comparing model predictions with children's judgments. The model name and the average mean difference score for that model is shown at the top of each histogram. The models are sorted from best to worst performance overall. (b) The distribution where the RH-R model performed worst, with the children's judgments (blue) and the model predictions (magenta) shown below.

The same modeling procedure was used as in Experiment 1a.

Results

The same models that performed well at predicting adults' judgments of tall also successfully predicted children's judgments of tall. Due to the limitations of the distributions, the results of the various models were less differentiated overall.

As in Experiment 1a, I evaluated how well each model predicted the children's judgments using the mean absolute error measure. Figure 3-6(a) shows the histogram of each model's performance across all distributions.

The models' ranking by performance is similar for children and adults. The main difference in ranking is the improved performance of the Absolute Height model overall. This is probably an artifact of the distributions used; the heights of the objects across the distributions in Experiment 1c are far less varied than across the distributions in Experiment 1a due to the limitations of the physical objects used. Most of the distributions range in height from about 1-9 inches, and the AH model is therefore

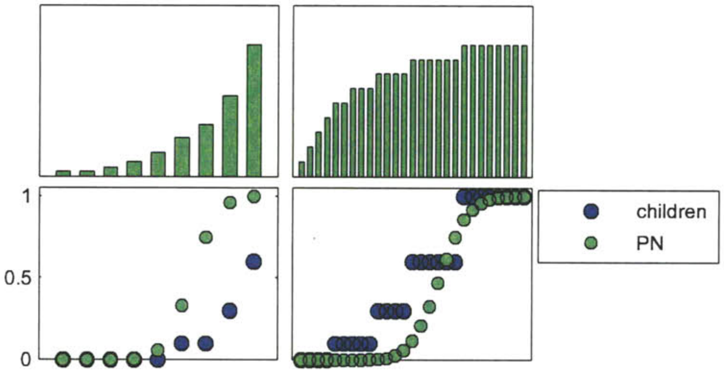


Figure 3-7: (a) The PN model shows a bidirectional failure in predicting children's judgments of *tall*.

able to perform well by setting its threshold to the height that ends up coinciding with the optimal RH-R threshold in most distributions.

Along with this limitation in variation in height among the distributions used for this experiment, the distributions in Experiment 1c vary less in number of items per distribution than the distributions in Experiment 1a. Therefore, all of the models, including those based on the number of items in the context, perform more similarly in predicting the children's judgments than the adults'. Nonetheless, Figure 3-6(a) shows that the top three models remain the same between children and adults. The RH-SD model and the CLUS model perform the best, with RH-R trailing behind, due mainly to the results of a single distribution (shown in Figure 3-6(b)). This worst MAE for the RH-R model is about the same as the worst MAE for both the RH-R and CLUS models in Experiment 1a, and is better than the worst MAE for the RH-SD model in Experiment 1a. Thus, the RH-R model did not perform worse at predicting children's judgments than adults', either on average or in the worst case.

Experiment 1c did not allow for the same degree of testing the RH-SD model that Experiment 1a did. In this experiment, there were no distributions which could test the RH-SD model for the kind of failure exhibited in Figure 3-5(e).

The PN model again shows a bidirectional failure here (see Figure 3-7), despite its good performance on average. Therefore, the models that best predict children's judgments of tallness overall are the same three that best predict adults' judgments

– the RH-R, RH-SD, and CLUS models. This result shows that from early on in children's use of the word *tall*, they are understanding and using it as a flexible word that picks out items in a given context based on a statistical computation over their heights. This matches adults' general use of the word. The best model parameters are similar for adults and children, though adults seem to be choosing a greater proportion of the items as tall than children are, on average. This could be an artifact of the experimental paradigm differences, or a difference in participants' attention span. However, it is also possible that adults' and children's specific parameterization of *tall* changes over time, even if their basic understanding of how the word works remains similar. Possible developmental trajectories of gradable adjectives are explored further in the General discussion.

Experiments 1a-1c show that the best predictors of human usage of *tall*, in the case of both adults and children, are statistical models that operate over the heights of the items in the context. However, it is still not clear whether the RH-SD, RH-R, or the CLUS model best explains judgments of *tall*. In Experiment 2, I further explore this question by seeking contexts where the three models make different predictions.

Experiment 2: Judgments of *tall* in ambiguous circumstances

Method

Participants

Participants were 107 adults, participating for payment via the online Amazon Mechanical Turk interface. Each participant saw and was asked for judgments about a single distribution of items, yielding 17-18 participants viewing each distribution in the experiment. [9]

[9] 108 requests for participants were put online, but one person participated twice and their second contribution was discarded.

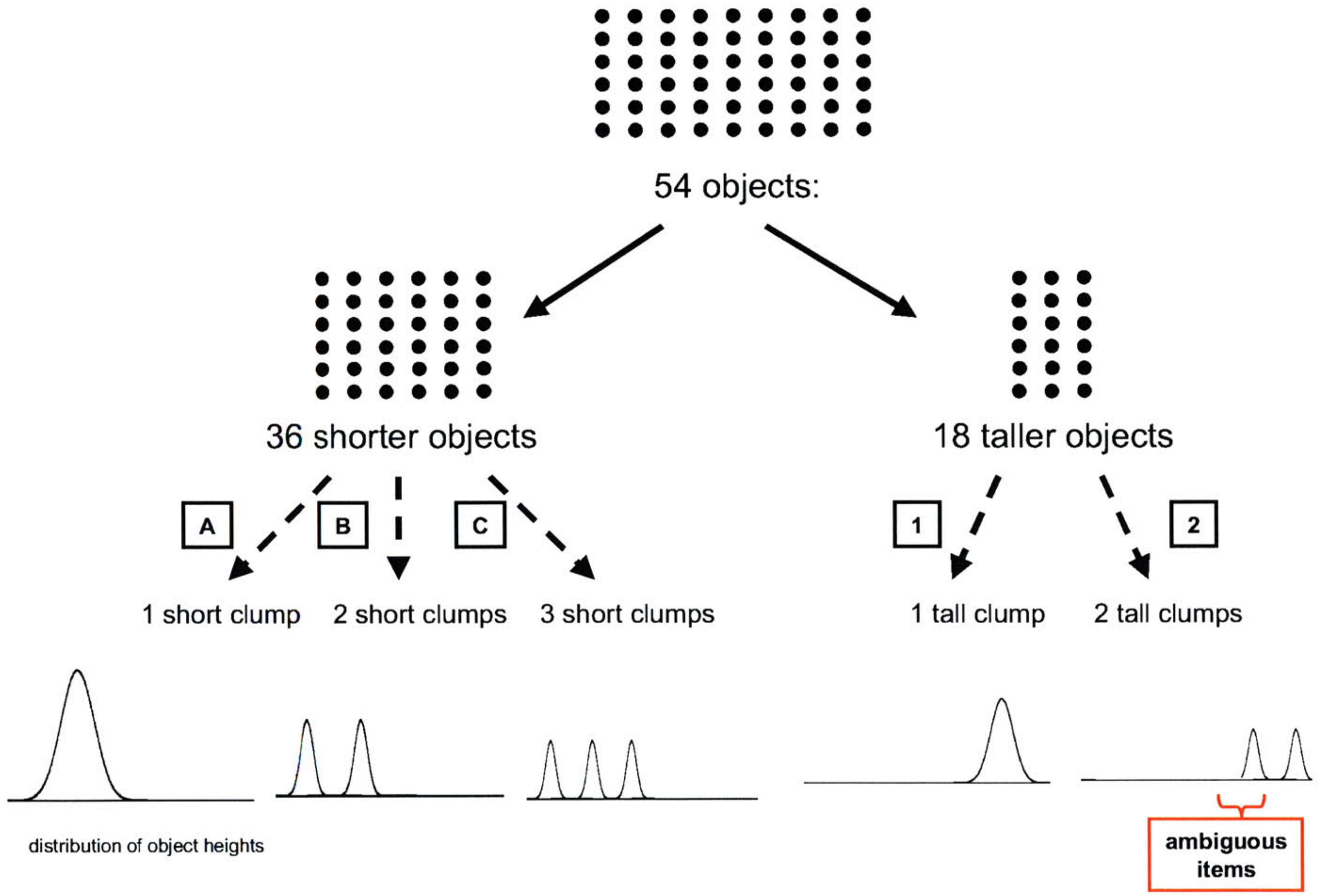


Figure 3-8: A schematic illustration of the process of stimulus generation for Experiment 2. The 54 objects in a distribution are divided into a short clump and a tall clump. From there, one of the three options (indicated by dotted lines labeled by letters) is chosen to subdivide the short clump into clusters , and one of the two options is chosen to subdivide the tall cluster (indicated by dotted lines labeled by numbers). Thus there are six final distributions, indicated by combinations A1, A2, B1, B2, C1, and C2. The actual object heights are sampled from a Gaussian distribution for each cluster of items. The overall mean and standard deviation of the short clump of items is the same across all distributions, and the same is true for the tall clump of items.

Stimuli

The stimuli consisted of six distributions of 54 items. The distributions were the same format as in Experiment 1a. The goal in designing the experiments was to create contexts in which the three best models from Experiment 1a made significantly different predictions. The stimuli were designed such that three of the contexts included an ambiguously tall group of items that the models made different predictions about. The other three contexts were parametrically matched in most ways to the first three, but contained no ambiguous group of items.

See Figure 3-8 for an illustration of how the distributions were generated. Each distribution was constructed by dividing the items into two clumps: a shorter clump of 36 items, and a taller clump of 18 items. The shorter clump was subdivided into 1, 2, or 3 clusters (each consisting of items with heights sampled from a Gaussian distribution), and the taller clump into either 1 relatively high variance or 2 relatively low variance clusters, yielding a 3x2 design. Across all distributions, the short clump of items was matched for mean and variance (no matter how many subclusters it was divided into), and likewise for the tall clump. Therefore, the overall means and variances of the entire distributions were also matched.

The particular parameters of the ambiguous distributions were designed so that the optimal thresholds for the RH-SD and RH-R models, as determined in Experiment 1a, did not coincide for these distributions. In the ambiguous distributions, there was a large number of items which were taller than the best previous RH-SD threshold (0.4 standard deviations above the mean) but not taller than the best previous RH-R threshold. Additionally, the CLUS model made predictions in the ambiguous contexts that were more sharply categorical than the more gradual sigmoidal fall-off of the threshold models.

The items were presented to subjects in one of three pseudorandom orders. See Figure 3-9 for two example distributions, one with one tall cluster and one with two tall clusters.

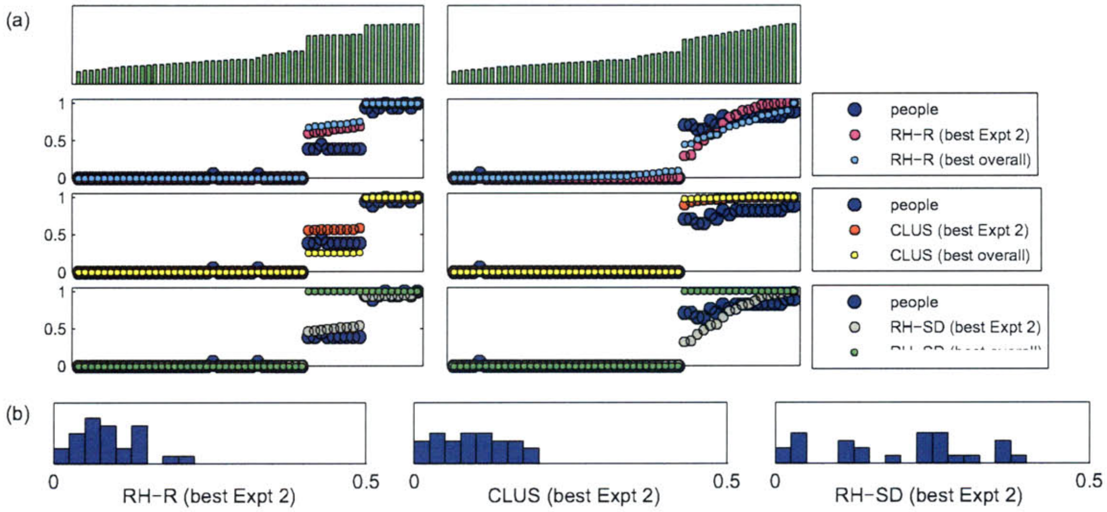


Figure 3-9: (a) Two sample distributions from Experiment 2 are shown at the top, with the performance of the three models shown below. For each model, the best parameter settings for just Experiment 2 and the best parameter settings for all the distributions of both Experiment 1a and Experiment 2 are compared. The best RH-R and CLUS models are very similar for overall results vs. for just the Experiment 2 distributions. The best RH-SD model overall, however, is very different from the best RH-SD model for just the Experiment 2 distributions – so much so that the best overall model does not capture adult judgments for the Experiment 2 distributions well at all. (b) The best Experiment 2 RH-R and CLUS models also perform very well overall, but that is not the case for the best Experiment 2 RH-SD model. The histograms show the mean differences for the best Experiment 2 models across all distributions from Experiments 1 and 2.

Procedure

The experimental procedure was the same as in Experiment 1a. The modeling procedure was the same as in Experiment 1a, except that a maximum of 5 clusters was used for calculating the possible category-based model partitions.

Results

The RH-SD model did a poor job predicting human judgments in the ambiguous cases. The RH-R and CLUS models both were able to fit human data well, with some subtle differences.

I once again assessed model performance on each distribution using MAE. I measured model performance across just the six distributions used in Experiment 2, and also across all the distributions used in both Experiment 1a and Experiment 2.

Figure 3-9(a) compares the performance of the three models to human judgments for two of the distributions, one containing ambiguously tall items and the other not. The performance of both the best parameter fit for just the six Experiment 2 distributions is shown, as well as the performance of the best models overall (for the combination of all Experiment 1a and Experiment 2 distributions).

The best parameters for the RH-R and CLUS models are similar for the Experiment 2 distributions and for all the distributions together, and the performances of these models are very similar in both cases; the best RH-R threshold is $k = 27\%$ of the height range for Experiment 2 (with a noise parameter of $\epsilon = 0.01$) and $k = 29\%$ overall (with a noise parameter of $\epsilon = 0.05$). The CLUS hyperparameters were also similar for the best Experiment 2 and best overall performance[10].

In strong contrast, the optimal parameter values and for the RH-SD model are far different for just the Experiment 2 distributions versus for the combined distribution set. The best overall RH-SD model ($k = 0.4$ standard deviations above the mean, $\epsilon = 2.5$) predicts with certainty that all the items in the tall clump are tall, whereas people are uncertain about the middle cluster of items (which we term the *ambiguous items* for the rest of this discussion). The RH-SD parameters can be adjusted to better match the human judgments for Experiment 2 ($k = 1.0$ standard deviations above the mean, $\epsilon = 75$), but these parameter results do poorly on the overall set of distributions, as shown in the histograms in Figure 3-9(b). These histograms also show, by contrast, the good performance overall of the best RH-R and CLUS models from Experiment 2.

Both the RH-R and the CLUS models predict the human judgments well with their optimal Experiment 2 parameter settings, as well as their optimal overall parameter settings, and it is not clear which model provides the best fit to the data. There are, however, subtle differences in their predictions. The CLUS model predicts almost equal probabilities of being tall for the ambiguous items; the RH-R model shows

[10] Overall: $\alpha = 7.5, \beta = 0.2, \kappa = 0.13, \gamma = 0.1$; Experiment 2: $\alpha = 7.5, \beta = 0.3, \kappa = 0.17, \gamma = 1.0$

a more sigmoidal fall-off for these items. Additionally, when there is only a single cluster of tall items, the RH-R model predicts much more of a fall-off in tallness judgments for the tall items than the CLUS model does, with people showing an ambiguous drop-off that is possibly closer to the RH-R predictions. On the other hand, with the best overall parameters, the RH-R model predicts a greater-than-zero probability of some of the short clump of items being labeled *tall*, though no people rated any of these items tall. The CLUS model, by contrast, does not predict any of the short clump items as tall without drastic parameter variation. Though these results illustrate differences in the models, we do not have a definitive answer as to which of the two models best fit human judgments – given the current amount of data, people's behaviors are well approximated by both models.

Overall, the results of Experiment 2 show clearly that the RH-SD model, an intuitive parametric model based on the mean and standard deviation of a distribution, cannot account for how people make judgments about gradable adjectives. The two remaining models, RH-R and CLUS, are both non-parametric models; the implications of human judgments being best predicted by a non-parametric model are explored in the General discussion. Their qualitatively different predictions suggest that future work will help us understand which of these two models is the strategy used by people, though on average both models predict the Experiment 2 data well.

How widely can we generalize these results in explaining gradable adjectives? Can a non-parametric model explain how we use a word like *long* equally well? In Experiment 3, I explore this question.

Experiment 3: Judgments of *long*

Method

Participants

Participants were 49 adults, participating for payment via the online Amazon Mechanical Turk interface. Each participant saw and was asked for judgments about a single distribution of items, yielding 9-10 participants viewing each distribution in the experiment.

Stimuli

The stimuli consisted of five distributions of items, which were a subset of those used in Experiment 1a and Experiment 2 (see Appendix B for details), except that the object values were interpreted as lengths instead of heights. Thus, the distributions were similar to those used in Experiment 1a, but turned on their side.

The items were blue rectangular items varying in length, lined up at along the left-side of the graph and varying in length (generated by a horizontal bar graph in Matlab, width=60%). Each item was labeled to the left and the right with a number, starting with 1 at the bottom of the figure and increasing. The surrounding frame was 800x600 pixels, and the bottom of the frame was 1.3 times the length of the longest item in the distribution. As in Experiment 1a, each distribution was shown to subjects in two pseudorandom orders.

Procedure

Each participant saw the following text:

> The blue objects shown below are veeks. Each veek is labeled with a number. Can you look at all of the veeks and find the *long* veeks?

The distribution was shown below the question. Beneath the distribution of items was a text box with the following instructions: "List the numbers of all the veeks that are long here."

The same models and same parameter settings were run as in Experiment 1a and 1c, following the same procedure. For the distributions used, all the item heights were unique, so the UPN and PN models made identical predictions, as did the UAN and AN models. Therefore, the unique models are not included in the analysis below.

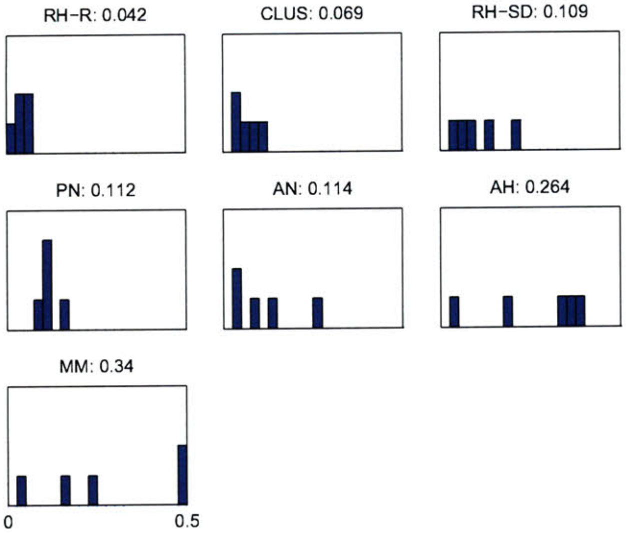


Figure 3-10: Histograms demonstrating the performance of each model in Experiment 3, comparing model predictions with adult judgments of *long*. The model name and the average MAE for that model is shown at the top of each histogram. The models are sorted from best to worst performance overall.

Results

Figure 3-10 shows the performance all the models across all the distributions. The RH-R and CLUS models outperform the other models on this distribution set; RH-SD comes in third, performing most poorly on the same distribution where it performed worst in Experiment 1a. The optimal parameters for the RH-R and CLUS models here are nearly identical to those for modeling adult judgments about the combined set of distributions in Experiments 1a and 2.

The distribution set for Experiment 3 is relatively small, and further work should be done to explore *long* and other gradable adjectives. But these results indicate that *long* can also best be modeled using a non-parametric statistical function operating over the values along the relevant dimension (e.g., heights or lengths) of the objects in the context.

General discussion

In this chapter, I have proposed and empirically evaluated models of gradable adjectives as a method of better understanding semantic compositionality. The empirical evidence supports the idea that gradable adjectives can best be modeled as statistical functions that take in a category of objects in the context and yield graded judgments about membership in a subcategory (e.g., *the tall ones*). What's more, the models that predict human judgments of gradable adjective use are non-parametric functions.

In this discussion, I address the immediate lessons of the empirical evidence with regards to understanding gradable adjectives. I then discuss the broader implications of the work, returning to the broad questions of cognitive science raised in the chapter's introduction.

Gradable adjectives as non-parametric statistical categorizers

Why should it be that the best models of human usage of gradable adjectives are non-parametric? Parametric model such as RH-SD may seem intuitively appealing given the number of data sets that people encounter that are approximately normally distributed. However, people also regularly encounter sets of items that do not follow a parametric distribution. Such contexts include cases where the objects under consideration do not belong to a single natural kind, e.g., "the things on the table". In these situations, the distribution of object heights (or other properties) may be highly irregular or multimodal. Additionally, even in more natural domains, common categories like *animals* or *trees* consist of many subcategories of items which may individually be normally distributed in terms of properties like height, but may form a much more irregular distribution overall. Therefore, a non-parametric approach may help people use and understand words like *tall* more flexibly than a parametric model would allow, applying equally well to *the tall people, the tall trees*, and *the tall things on the table*, whereas the Gaussian parameterization of RH-SD applies most meaningfully to the first situation only.

After Experiments 1-3, the question remains as to which of the two best models is the correct non-parametric approach to describe usage of *tall*. The RH-R model is simpler than the CLUS model in terms of number of free parameters. However, there remain two reasons why the CLUS model is a compelling model from a cognitive perspective. First, it seems to capture a crucial idea of linguistic compositionality: we use noun phrases to pick out complex subcategories that are not lexicalized as single nouns but that nevertheless correspond to interesting chunks of the world. Category-based models correspond to the idea that people are looking for a category of objects every time they use a noun phrase – including when use a noun phrase involving a gradable adjective – rather than reasoning about a threshold along a dimension. The second reason the CLUS model continues to be a strong candidate is that it relies on well-established, domain-general principles of categorization. If these general strategies also apply to understanding gradable adjectives, this would not entail positing any new cognitive complexity. The RH-R model, by contrast, is less widely applicable.

One additional drawback of a range-based model is that, given a normally distributed category such as *people*, as a learner continues to see new category members, she will continue to encounter outliers beyond the minimum or maximum of the previous set of items encountered. Encountering a new data point that changes the range of heights the learner has seen for category items will instantaneously change a range-based model's judgment of how likely each category item is to be tall. In a category-based model, on the other hand, new data points are likely to be grouped with existing clusters, unless they are distant outliers. Therefore, the meaning of *tall* and other gradable adjectives, as composed with a particular category, would be more stable over time using a category-based model.

While this kind of instantaneous change in the range-based model's predictions could potentially be avoided with a more sophisticated range-based model (e.g., if the range-based model discounts outliers), the predicted difference in model response to new outlier data points offers one future avenue for investigation in trying to determine how people use gradable adjectives. Experiment 2 also shows other areas where the two models make different predictions; empirically exploring similar distributions where the differences in model predictions are more pronounced should provide more evidence about which non-parametric function best describes human behavior.

I have shown how gradable adjectives act as a statistical function when they modify nouns or noun phrases, probabilistically identifying members of a subcategory of the given comparison class. Additionally, all of the models presented in this chapter can easily be extended to operate over graded category input. This is important in accommodating semantic composition with uncertain categories or graded categories; people can make judgments about which items are probably "the tall blickets" when they aren't entirely sure which items in the context are blickets, and people can judge which crayons are "the long green crayons" in a cases where the crayons range incrementally from green to blue. The broader implications of this work on gradable adjectives and compositionality are explored in the next sections.

Statistical categorization and formal semantics

Linguists have debated the semantic content of gradable adjectives. As noted previously, some have proposed that gradable adjectives serve as functions that map objects onto an abstract representation of degrees or an ordering along a scale, without specifying the details of how this function operates in a given context (e.g., Kennedy, 1999; Cresswell, 1976). Others have argued that gradable adjectives are *vague predicates* whose extensions vary from context to context – again, without specifying how the extension relates to the objects in a particular context (e.g., Kamp, 1975; Klein, 1980).

Our model of gradable adjectives as a non-parametric function unifies these approaches. According to this model, the extension of a phrase involving a gradable adjective does indeed vary from context to context. However, it does so in a predictable way, as demonstrated in the experiments of this chapter. As well as being functions which map objects to degrees along a dimensional scale like height (e.g., *tall* mapping a set of trees onto their heights), gradable adjectives can be seen as functions that map category members (e.g., *trees*) onto degrees of membership in a new category (e.g., *tall trees*). This work helps clarify the semantic content of adjectives and combines the advantages of the previous proposals.

Vagueness and the Sorites Paradox

The fact that gradable adjectives potentially involve vagueness in their extension has particularly puzzled semanticists. Vague terms are those which do not strictly classify objects into those that have a property and those that do not; while there may be some clear examples in a given context of what is *big* or *not big*, there is also a *range of indeterminacy* in which it is unclear whether the objects have this property (Partee, 1995). The Sorites Paradox of the heap of sand, as outlined in the introduction, is a classic example used to illustrated the problem of vagueness. Restating the inductive problem in terms of *tall*: if you have a tree that is *tall* and another one that is only infinitessimally shorter, the second one must also be *tall* – but continue this process and eventually you will be looking at saplings or bonsais, which are clearly not *tall trees*.

The approach to compositionality outlined in this chapter addresses this supposed paradox. Statistical functions yield probabilistic category membership, and this corresponds to graded rather than binary application of the property specified by the gradable adjective (i.e., if a tree has 80% probability of membership of in the category *tall trees*, then it has an 80% probability of having the property of being tall). Thus, because the model reasons probabilistically, there can be uncertainty as to whether a given tree is tall. The probability of two adjacent items (ordered by height) being tall may be very similar, if they are similar in height (and depending on the rest of the context). In fact, two adjacent items may be perceptually indistinguishable (see Raffman, 2000) and therefore have an infinitessimal difference in their probability of being in the tall category. But none of this leads to a necessary paradox, nor does it lead to a situation that we cannot define beyond labeling it vague. Instead, a statistical compositional model allows us to predict very specifically the probability with which an item will be in the extension of a gradable adjective in a given context. It explains how we can start with an object that is a clear example of a tall item and proceed through a set of just slightly shorter items to an item that is clearly a short item, without necessarily ever crossing a point of clear threshold (if people are indeed performing a threshold-based function, as in the RH-R model, then they will be performing a computation that involves a threshold for tallness, but they may have uncertainty about where exactly that threshold lies; similarly, if people are performing a category-based computation, as with the CLUS model, they will be reasoning implicitly about the threshold between the tallest category and the other categories, but again, they can be uncertain about that category boundary). While semanticists may not be able to assign a definite truth value to "this tree is a tall tree" for all the intermediate cases on the spectrum (at least, not one truth value that everyone will agree with), a the statistical categorization approach to gradable adjectives allows them to understand the probability with which people will agree with the statement.

Standards of comparison

Kennedy (2002) discusses vague terms as depending on a shifting *standards of comparison* in their evaluation — "this tree is a tall tree" may be judged true or false depending on the context, because the standard of comparison against which tallness is judged changes. I show here how the analysis of gradable adjectives in this chapter can give rise to such shifting judgments. The graded category boundary that is the result of combining a gradable adjective with a set of objects can be seen as the (fuzzy) standard of comparison by which objects may be judged tall or not tall, and this shifts depending on the comparison class.

Kennedy notes a number of specific examples of how context can affect standard of comparison, including the examples below, which I reformulate in statistical function-based terms:

- *Standard of differentiation*: "The fat pig" can be used to pick out the larger of two pigs, neither of which is fat compared to the broader population. In the comparison class of two pigs, the fatter one is will be understood as the fat big by either the RH-R or the CLUS model, because it is the fattest in the set. This looks like a binary evaluation of *fat* because it is the tallest item in the comparison class, which, according to our non-parametric statistical models, always has 100% chance of being rated *fat, tall, big*, etc.

- Judgments based on exceeding a "norm of expectation": "Benny is tall!" can be spoken regarding a 2-year-old who is short for his age, but has grown since last time he was observed by the speaker. If the comparison class is past examples of Benny, or the last Benny observed, then Benny will be judged tall by the statistical models, as again being the tallest item in the set. This will always be true of current Benny vs. past Bennies (so long as Benny is growing), but in particular, if Benny has grown substantially as compared to the previously observed examples of Benny, then the past examples will not necessarily seem tall in comparison to current Benny. So if Benny has grown a lot since the last time the speaker saw him, it may seem worth remarking upon the apparent categorical difference.

- Judgments based on a "norm of average" vs. judgments based on a "norm of life expectancy": People may remark, "Fido is an old dog," regarding a 14-year-old dog. They can also say, "Rover is an old dog," for a 20-year-old dog, and mean something different; Rover is unusually old for a dog, while Fido is at the end of a normal dog lifespan. Here it seems that Rover is being compared to the set of dogs at the time of death, or the oldest dogs in existence, while Fido is being compared to the population of dogs of all ages.

Kennedy's assessment of the different norms at play in using gradable adjectives revealing important factors that can influence the context in which an item is being evaluated in a given expression. According to the statistical categorization approach, the gradable adjectives maintain a consistent meaning throughout all these cases, performing the same probabilistic function on different comparison classes. There is a large role for both semantics and pragmatics to play in identifying the comparison class, or context, used as input by the gradable adjective in a given expression.

The suggestion that adjectives are functions is not a new one in the field of semantics (see, e.g., Parsons, 1970; Montague, 1970; Partee, 1995); adjectives have previously been proposed as mapping nouns onto ADJ + N intensions, or properties rather than sets. The intension combined with the given context then lead to a given extension, or set of objects in the world. The contribution of the work in this chapter to the semantics literature is three-fold: (1) showing exactly how, given the contextual set of circumstances, a gradable adjective and noun combination gets interpreted, (2) explaining and quantifying the apparent vagueness of terms, and when people will be certain of their use vs. when the example will fall into the "range of indeterminacy" (and how uncertain people will be), and (3) providing a simple and unified account of the shifting standards of comparison. I next examine how this can be generalized beyond the kinds of gradable adjectives we have so far examined.

Typology of gradable adjectives

I predict that the model we have presented should be applicable to all relative gradable adjectives, including more abstract ones such as *smart* or *talented.* These dimensions are harder to measure exactly than dimensions such as height. But at the least, items can often be ordered at least approximately along these dimensions (as in grading students or judging a talent competition), and people can make judgments about small vs. large gaps in intelligence, talent, attractiveness, etc. The input values (the degrees of intelligence, talent, beauty, etc.) may be more uncertain, but a statistical function could still apply in these more abstract domains.

Some gradable adjectives, unlike the ones I have studied here, are *absolute*, e.g., *full* and *wet.* One glass of water can be *fuller than* another, but there is a specific meaning for a *full glass*: it is at the maximal endpoint of the scale. This is an example of a *maximal absolute adjective. wet* is a *minimal absolute adjective*; one table can be *wetter than* another table, but each table individually is either wet or it is not (it must simple surpass the minimal wetness endpoint to qualify as *wet*).

Syrett et al. (2006) empirically show that people's judgments of relative gradable adjectives like *big* and *long* vary based on context (in keeping with Barner and Snedeker (2008)'s findings for *tall*), but that the same does not appear to hold for absolute gradable adjectives. For instance, given two partially-filled containers, children will not find "Give me the full container" a felicitous request. Additionally, if given a distribution of containers varying from empty to full, adults will only pick the completely full item when asked to select the full containers. Syrett et al. take this evidence to mean that the applicability of an absolute adjective to a particular item does not change based on context, unlike judgments for relative gradable adjectives.

The studies in this chapter are agnostic on this point. However, it seems possible that both relative and absolute gradable adjectives could be described by versions of the context-sensitive non-parametric models developed here. It is possible that absolute gradable adjectives are sensitive to different input contexts that Syrett et al. did not investigate. Intuitively, context could make a difference to whether something is judged full; a beer mug that is filled to an inch from the top is generally considered a full glass of beer when compared to a half-filled mug, but compared to a completely filled-to-the-brim mug, the same container might not be judged full. However, if both the completely-filled and the filled-to-an-inch-from-the-top mugs are compared to a bunch of mugs with only a trace of liquid in them, both of the mostly filled containers might be selected when someone is asked to fetch the "full mugs". Therefore, it seems likely that people are both taking context into account and also performing a statistical computation based on context, though undoubtedly with very different parameters for threshold or category size than in the case of relative adjectives. In contrast to the model for relative gradable adjectives, the fullest item in any given comparison class would not be automatically labeled *full* — this label would be anchored to the end of the spectrum only.

Beyond gradable adjectives

Gradable adjectives are not the only vague terms, nor the only terms whose meaning must always be interpreted based on compositionality with the given context. Intensifiers such as *very* or *extremely* could also be functions that operate over a comparison class and return, a probabilistic subcategory of a category such as *tall trees* (e.g., the *very tall trees*).

Some quantifiers are also vague lexical items. Quantifiers operate on a scale of ordered cardinalities of sets (ranging from the empty set to the set of all items in the context); *none* <*some* <*most* <*all* (partial ordering only). Some have puzzled over the fact that people's usage of quantifiers like *most* seem to be best described using "mushy Gaussians" despite the fact that quantifiers pick out sets of individuals. Quantifiers allegedly have an exact boundary, such as "50%+1" for *most*, but children and possibly adults seem to employ a vague FUZZY-MOST concept (Halberda, Taing, & Lidz, 2008). What's more, context may affect quantifier usage (3 out of 5 students seems to be *most of the class*, but if 1 million people vote and only 500,001 vote for the winner, then intuitions are mixed as to whether *most voters* voted for the winner). A statistical model of compositionality, operating over set cardinalities instead of over a set of measurements, could also potentially explain and more precisely describe how people use quantifiers.

These cases need to be further investigated empirically to better understand the patterns of human usage of these compositional phrases. But each word or phrase that both involves graded categorization of items or groups on an ordered scale and also depends on context for its interpretation could potentially be described using a similar model to the ones outlined in this chapter.

Statistical categorization and mental representation

How does this theory of gradable adjectives address the concerns of conceptual representation and combination? Under this theory, in order to understand a lexical concept corresponding to a gradable adjective, a speaker must represent the following information:

- the relevant dimension upon which items are compared (e.g., the dimension of height for *tall*)
- the polarity of the lexical item (e.g., positive for *tall* vs. negative for *short*)
- the statistical computation that the adjective performs
- the parameters governing the computation (i.e., governing either the threshold location and noise level, or the prior expectations about cluster sizes)

In order to combine a gradable adjective with a category to yield a more complex concept, sufficient information about the relevant category must also be represented. For a finite comparison class like those used in Experiments 1-3, this is trivial. For a category like *tree*, however, what must people represent in order to judge the *tall trees*? For a range-based model, making such a judgment requires representing the minimum and maximum values seen along the relevant dimension (e.g., the tallest and shortest trees ever encountered). For a category-based model like the CLUS model, people could make judgments based on a stored set of category exemplars (which would also be sufficient for the range-based model, but not necessary).

This work is agnostic about the actual processing that goes into the function of gradable adjectives. It is quite possible that people do not recompute the statistical function over a given category every time they think about what is *tall* or *big* within that category. Perhaps, for example, people store a summary representation of an item on the boundary of *tall*, or a prototypical tall item within the category, and only recalculate this if their information about the category changes. However, people are clearly able to easily perform these computations on the fly any time they run into a novel category like "the items on this table".

People clearly do not have to store examples of every phrase that they hear used in order to understand the meaning of phrases. This chapter provides insight into how the categories and functions referred to by words can be combined to form more complex categories referred to by phrases. These studies and models address the particular example of compositionality involving gradable adjectives modifying noun phrases, but I have also discussed how other classes of lexical items like intensifiers could potentially be incorporated into this computational account. The work in this chapter serves as a starting point for a broader model of compositionality – one that explains how people can make judgments about category membership given phrases that they have never heard used before, knowing only the meanings of the individual words within the phrase.

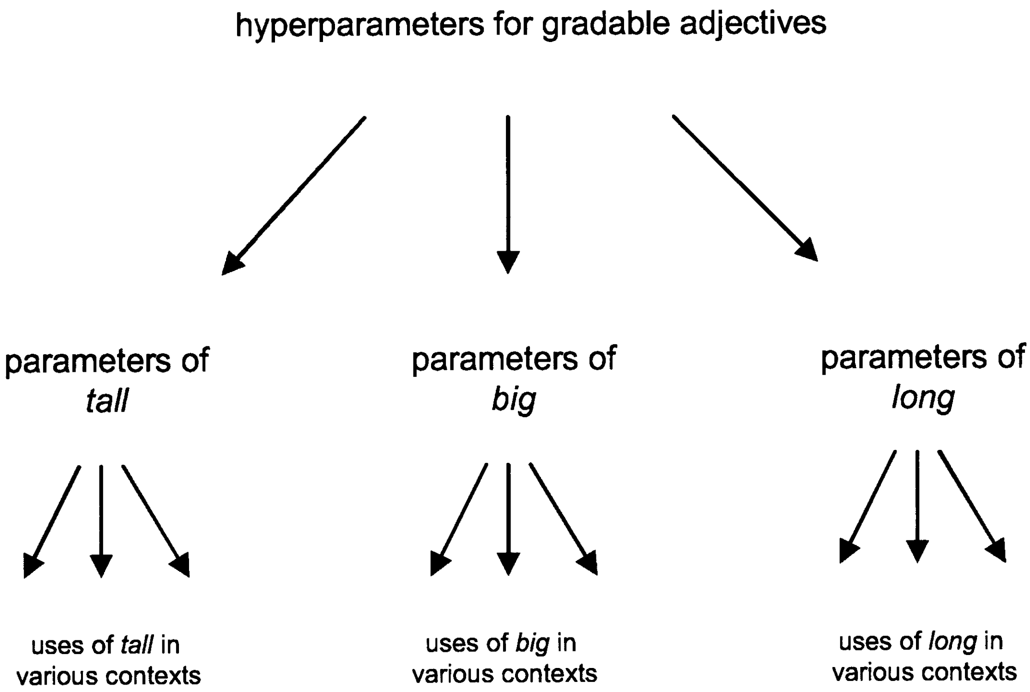

hyperparameters for gradable adjectives
parameters of *tall*
parameters of *big*
parameters of *long*
uses of *tall* in various contexts
uses of *big* in various contexts
uses of *long* in various contexts

Figure 3-11: The learner may form an overhypothesis about the meanings of gradable adjectives as a class of words. From individual uses of *tall*, a learner infers the parameters governing that word; from the combined parameters of multiple gradable adjectives, general expectations about gradable adjective parameters may be inferred.

Learning words that are compositional in meaning

Children use gradable adjectives like *big* and *tall* similarly to adults by age four (though there are many gradable adjectives that they do not yet know at this age). How do children learn gradable adjectives, and what pre-linguistic expectations aid them in this endeavor? These questions remain open; however, the work in this chapter provides a framework for understanding and empirically investigating the possibilities.

Perhaps children start out expecting that there will be a class of words which operate as statistical functions to yield graded categories as adjectives do, and perhaps they know what form the functions take (e.g., they expect the class of words to be based on finding categories, or based on range). Maybe the process of learning a gradable adjective entails identifying the right dimension and polarity, and then fine-tuning the parameters of the known function. This view of gradable adjectives shows how a child's judgments of a gradable adjective *tall* could potentially change over time without any drastic conceptual change; all of the components of the child's mental representation can potentially remain stable over time, with only the parameters of the function being adjusted based on new evidence.

At the other extreme, it's possible that children start out not expecting such a class corresponding to statistical functions and must learn their first few gradable adjectives independently and laboriously, figuring out what function each word corresponds to independently. Work by Keil and Carroll (1980) suggests that children may even start out learning individual adjectives like *tall* as a separate modifier for each category that it modifies (for instance, meaning *narrow* when applied to some categories and *large* for others), and then eventually generalize about what *tall* means across all situations. Something similar might occur eventually for the broader class of gradable adjectives: at some point, by recognizing common linguistic patterns, children might generalize that all gradable adjectives act as set of similar statistical functions with slightly varying parameter settings. In between the two extremes, there is the possibility that children expect all gradable adjectives to act as the same kind of statistical function, but don't know what the function is to begin with.

In any of these scenarios, there is the possibility that the child could learn about the meaning of gradable adjectives as a whole from the meaning of individual lexical items. Forming an *overhypothesis* (see Kemp et al., 2007) about the meaning of the class of words would mean that children would form prior expectations about the parameters of the statistical functions across all gradable adjectives, such that thresholds or category sizes would generally be similar even when making judgments about different dimensions (see Figure 3-11). On the other hand, it is possible that children learn the parameter values of each gradable adjective individually without any expectation that parameters will be similar across different dimensions of measurement.

Experiment 1c does not provide conclusive evidence about how gradable adjectives are learned. The experiment is in part limited by the fact that only a single gradable adjective is studied, but also by the fact that the 4-year-olds in the study had a well-developed understanding of *tall.* Preliminary modeling of children's and adult's judgments of *short* (gathered by Barner and Snedeker (2008) and in additional unpublished data) suggest that adult-like understanding of *short* takes longer to develop than understanding of *tall.* Even so, children at age 4 are already making judgments that are well predicted by a statistical non-parametric function, though the parameters of that function differ from that which describes adult judgments. Further evidence is needed to answer questions about the development of understanding gradable adjectives, and what conceptual understanding children begin with in this domain. More judgments from children of a wider range of ages about a wider variety of gradable adjectives will help answer these questions. It would also be informative to teach adults or children gradable adjectives that describe novel dimensions and test their initial judgments of which items the adjectives apply to.

In this chapter, I have provided some initial data of how children use gradable adjectives relative to adults. I have developed the basis for modeling such data and suggested how the empirical and modeling techniques from this chapter can be turned to answering questions about the developmental trajectory of understanding gradable adjectives.

Conclusion

In this chapter, I proposed that gradable adjectives can best be modeled as a non-parametric statistical function over a category (or context) which yields probabilistic judgments about membership in a subcategory. I provided empirical evidence that adult and child judgments match such a function for *tall*, and provided preliminary evidence that such a function also describes *long.* This approach explains how gradable adjectives (and potentially other modifiers which are compositional in meaning) can compose with noun phrases to form more complex concepts. This theory of gradable adjectives also matches with knowledge from formal semantics of people's vague judgments about when gradable adjectives apply, providing an explanation and quantified description of this vagueness. The non-parametric statistical model of gradable adjectives not only matches empirical evidence of usage of these lexical items, but provides far-reaching cognitive and linguistic insights into lexical meaning and compositionality.

In the next chapter, I look at how semantic composition relates to the notion of making sense. In particular, I examine how we instantly can judge whether a combination of predicate and noun phrase makes sense or is nonsensical (even if the phrase refers to something that we've never seen before), and how we learn this distinction from limited evidence.

Chapter 4

Distinguishing sense and nonsense

The problem of learning what makes sense

How do we learn what combinations of words make sense together? In the last chapter we looked at the meanings of phrases, and how language users judge category membership given a compositional meaning. However, some words can be syntactically combined, but do not yield a phrase that is semantically sensible (e.g., *hour-long banana*). We seem to know the difference intuitively and instantaneously between combinations that do and don't make sense. What governs this distinction? In this chapter, I present a model for learning sense and nonsense based on limited evidence. I also discuss whether the constraints that make this learning possible can themselves be learned.

Sensibleness and the puzzling lack of evidence

Imagine that your friend is coming over to visit after going shopping at the market. She tells you that she bought a kind of banana that she had never run across before – a very unusual banana. She tells you you will be surprised, and as she reaches into her grocery bag, you try to predict what properties this fruit has that will surprise

The work in this chapter was carried out in collaboration with Charles Kemp and Joshua Tenenbaum.

you. Maybe these bananas taste like watermelon. Maybe they are bright blue. Maybe these bananas are an hour long.

Wait, scratch that. Chances are, that last thought never crosses your mind. It might occur to you to wonder if the bananas look or taste odd. But you are unlikely to hypothesize about their duration. You are also unlikely to wonder whether these bananas believe the Earth is flat, or whether the bananas are the President's fault. Why not? It's not a matter of evidence; you've probably never seen a blue banana or an hour-long banana – or any of the other hypothetical bananas. Still, there is a distinction. Tasting like a watermelon or being bright blue would definitely be surprising qualities for a banana to possess, but are imaginable. The other properties wouldn't be surprising so much as flat out baffling – they just don't make any sense.

What is the difference between these properties? And how do you instantly and intuitively know that, of all these banana types you have never encountered, and probably never contemplated, watermelon-flavored bananas and blue bananas make sense, but hour-long bananas and the others do not?

The fact that people never encounter any of these counterfactual bananas is a large part of what makes this inference question so interesting. Many judgments that we make involve figuring out what is likely to happen based on what we have previously seen. This problem is different, as we have no past evidence about either blue bananas or hour-long bananas, nor are we likely to encounter any in the future.

Perhaps we are looking at the problem of evidence incorrectly, however — we have no direct evidence about our counterfactual bananas, but perhaps indirect evidence can help us to make a guess as to what is sensible. We may never have seen a blue banana, but we have seen many other blue things – blueberries, bicycles, and bedspreads, for example -- which share many properties with bananas. All of these items are visible objects, and we can touch them, move them, or give them to our friends. If these objects can sensibly be blue, then perhaps it would be sensible for bananas to be blue as well. None of the other blue objects is an hour long, however, which might make us more doubtful that bananas could sensibly have this property.

This insight — that indirect evidence can help us recognize categories of objects which share properties in common — is key to the work I present in this chapter. I will be demonstrating how it is possible to use such evidence to learn categories of objects, categories of properties, and the relationships between these categories — and how to, from there, infer what makes sense and what is nonsense. Before I proceed, however, let's clearly define sense and nonsense.

What are sense and nonsense?

It seems to be common sense that bananas cannot be an hour long. But what is common sense? Common sense is generally taken to be the set of facts that most members of the general population are expected to know. This can include facts about what is generally true ("Cows are not green") as well as facts about how things are usually done ("If you go to the store, you have to pay for items before you leave"). Nonsense and sensibleness, as I will be discussing them, are a strict subset of the commonsense domain. People may sometimes grow up lacking some "commonsense" knowledge (e.g., not knowing that you have to pay for items before leaving a store), but few (if any) adults are likely to believe that bananas could sensibly be an hour long.

As suggested above, there are many ways in which a statement can fail to make sense. The sentence *Colorless green ideas sleep furiously* is packed full of different kinds of nonsense. In part the sentence does not make sense because ideas cannot be green, nor can they sleep – this has to do with which types of properties that can modify which objects. Additionally, things that are green cannot also be colorless – these two properties are contradictory. And sleeping cannot be done furiously; that modifier makes no sense with that property.

In this chapter, I address the specific kind of commonsense knowledge that involve which types of objects can have which types of properties. I will refer to properties also by their corresponding linguistic term, *predicates*; predicates include feature-based properties such as *is green*, and also behavioral properties such as *dances* or *has been kissed.* Objects are things in the world or the mind – anything describable using a noun phrase is an object.

What types of object-predicate combinations are nonsense as opposed to simply being untrue? Nonsensical pairs are those that are semantically anomalous due to *category mistakes* ((Horn, 1989)). Statements containing category mistakes can be distinguished from statements that are untrue by negating the sentence; an untrue sentence will become true when negated, whereas a nonsensical statement will not become true when negated. "Bananas are usually an hour long" is not true; "Bananas are not usually an hour long" is also not true, as the duration of bananas cannot be measured — it is a category mistake to talk about bananas as if they are the kind of thing that can be an hour long. By contrast, "Cows are green" is untrue, but "Cows are not green" is true. Therefore, *an hour long banana* lacks sense, while *a green cow* is merely unlikely.

Is sensibleness a binary distinction? Drange (1966) presents a set of sentences which he claims show graded sensibleness, including the following (in decreasing order of sensibleness):

- Englishmen like coffee better than tea.
- Squirrels like coffee better than tea.
- Bacteria like coffee better than tea.
- Stones like coffee better than tea.
- Quadratic equations like coffee better than tea.

Drange ultimately concludes that some combinations are completely unthinkable (i.e., nonsensical), despite the apparent gradation of plausibility of Drange's sentences. Sensibleness could be binary, entirely graded, or usually binary but also allowing for some degree of gradedness (possibly only in certain well-defined cases). For the purposes of the model presented in this paper, I will treat sensibility as a binary distinction, but this assumption could be relaxed. This point is taken up again in the General discussion.

The learning problem and the M constraint

How can children acquire the knowledge of (or the ability to rapidly infer) which object-predicate pairings are sensible and which are nonsense, based only on the indirect evidence that they gain about which object-predicate pairings are true in the world? This is a hard problem both because (i) there is, for the most part, only positive evidence to work from – children do not get very many pieces of evidence of the form "soccer games can't sensibly be heavy" – and (ii) the set of object-predicate pairs which are true are a strict subset of those which are possible. The second fact means that the learner must do more than guess from positive evidence what other evidence they are likely to see. She must also perform a second level of inference, deducing which pairs are unlikely to be seen because the object and predicate just don't happen to occur due to, e.g., environmental circumstances (flamingos are not blue because their diet of krill turns them pink) and which pairs are not seen because they are nonsensical (flamingos are not an hour long). In explaining how a child (or learner more generally) can acquire this ability, I will address the following questions:

1. What biases or constraints on a learner's mental representation of the world would allow her to learn what is sensible, and how could these be applied?

2. To what degree could these biases or constraints could themselves be learned from the data in the world (as opposed to being innate)?

In answer to Question 1, I present a model that learns sense and nonsense given information about what actually occurs in the world. This model employs a hierarchical constraint that may also make human learning of sense and nonsense possible. In addressing Question 2, I investigate what kinds of evidence would be sufficient for a learner to infer that such a hierarchical constraint is appropriate.

My model for learning sense and nonsense is based on a theory of predicates and objects proposed formally by Sommers (1959, 1963, 1965, 1971). Sommers' work formalized and expanded on ideas about the structure of object categories and predicates that were recognized as far back as Aristotle in his *Categories*: that objects

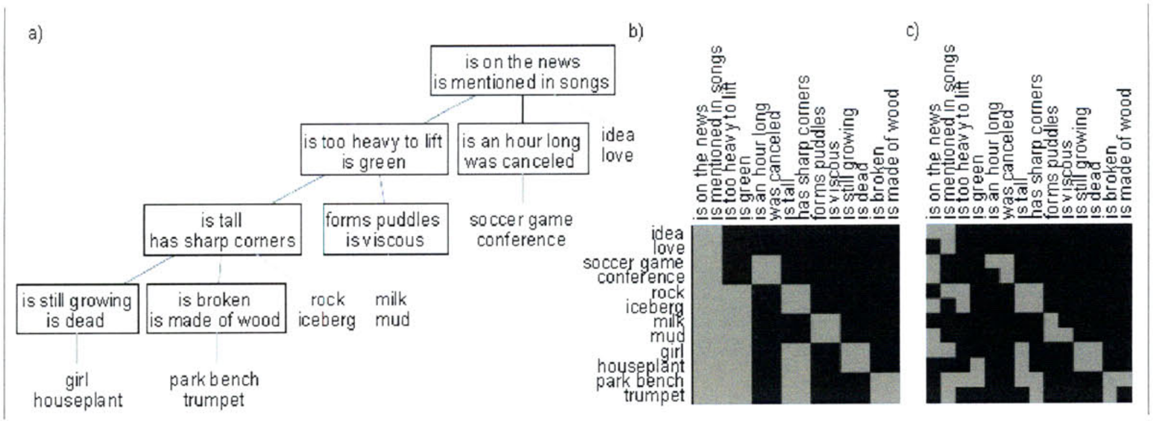


Figure 4-1: (a) A predicability tree. The predicates located at a node of a tree span all the objects at that node and all those in the subtree below it. (b) The corresponding predicability matrix, R. Squares in gray indicate predicable pairs. (c) A truth matrix, T, indicating the predicable pairs that are actually true.

belong to nested categories, and that predicates apply sensibly to entire categories.[1] Sommers proposed that objects are organized hierarchically in an ontology (e.g., a puppy is a member of many ontological categories: it is a DOG, an ANIMAL, a BIOLOGICAL ORGANISM, a PHYSICAL OBJECT, and so on). Predicates are placed at nodes in the tree, and are *span* the objects in the subtree beneath them. When a predicate spans an object, it can sensibly be paired with that object, but is not necessarily true of that object; the predicate *is green* spans *cow* even though cows are not generally green. The resulting structure of predicates and objects is called a *predicability tree.* See Figure 4-1 for an illustration of the structure of a (simplified) predicability tree, and the corresponding sensibility and truth matrices.

The strict hierarchical constraint that governs the predicability tree was termed by Sommers the *M constraint.* The M constraint is so named because the strict hierarchy prevents "M" shapes from appearing in the tree. If, for all pairs of predicates, both predicates either span the same set of objects, or else one predicate spans a strict

[1] The idea that mental representations of natural kinds follow a hierarchical structure that does not include disjoint sets has been formulated by many others, as well. See, e.g., Atran (1998) for a discussion of conceptual representation of biological taxonomies; G. L. Murphy and Smith (1982) and Markman (1989) for a discussion of tree structures in cognitive representations of categories; and Berwick (1985) for a discussion of how the Subset Principle and the M constraint are congruent in eliminating disjoint categories.

subset of the objects spanned by the other predicate, then the M constraint is honored. If, however, the predicates span overlapping sets of objects such that each spans at least one object not spanned by the other, then an "M" appears in the tree, and the constraint is broken.

Is this M constraint a reasonable limit to place on the information that people learn and store about the world? It is unclear whether the M constraint reflects the actual structure of the world; there is some evidence that the real world predicability structure is more complex (see the General discussion). However, even if the M constraint does not entirely reflect the organization of objects and predicates in reality, it can be a powerful tool for learning and representation. The structure governing which object-predicate pairs are true is a very complex lattice, as opposed to a tree. This lattice structure is not compact and does not allow for generalizations about kinds of objects or properties, nor does it lead to quick or easy inferences about new objects or predicates that a learner encounters. Simon (1969) showed that a strict hierarchical structure was the optimal organization for making such inferences.

For an intuitive example, imagine that we learn one fact about a novel object called a *blicket.* "I had to fix my blicket with a screwdriver today," a friend says. What else can we deduce from this besides that a blicket can sensibly (and truthfully) be fixed by a screwdriver? We can infer that a blicket must be somewhere in the subtree of the predicability tree containing things that can be fixed by a screwdriver — the subtree containing PHYSICAL OBJECTS, and more specifically, ARTIFACTS. From this, we can deduce other predicates that can be sensibly applied to blickets: e.g., *can be seen*, *can be picked up*, *is used in the kitchen*, *was made in a factory.* Note that not all of these are necessarily true of blickets, but can sensibly be combined with them. But if any of these predicates did not apply to blickets given that they can be fixed by screwdrivers, then the strict hierarchical structure of the tree would be broken. We can also judge predicates that would be nonsense when applied to blickets: e.g., *is an hour long*, *believes in Santa Claus*, *is a lie.* If any of these predicates *did* apply to blickets, then the M constraint would also be broken. Clearly, the M constraint is a powerful tool for inference about what is sensible.

The M constraint may be efficient, but do people follow such a constraint? Keil (1979) found strong evidence that both adults and children seem to represent things in a strict hierarchy like that Sommers proposed. While human judgments converged to be consistent with a single predicability tree by adulthood, Keil found that predicability judgments varied greatly between children, and changed as children grew older. However, as early as kindergarten, each individual child's judgments were almost always tree-consistent. Additionally, Keil's cross-cultural studies additionally found that it was not just English speakers who followed the M constraint in making predicability judgments; the constraint appears to be a more universal cognitive property. Keil's findings provide evidence for the psychological plausibility of the computational model I propose.

In addition to needing a representation governing which predicate and object pairs can sensibly go together, someone learning about what is sensible would also need to be able to represent which of those sensible pairs are also true in the world. The model I will propose combines Sommers' hierarchical M constraint representation of sensibility with additional information about which sensible pairs are also true. This allows the learner to perform both levels of inference ("what is true?" and "what is sensible?") based on the evidence from the world about how often various object-predicate pairs occur.

Given the M constraint model, Question 2 asks: Where does the M constraint come from? Keil's answer was that the constraint is innate. He believed that children are able to learn over time the specifics of which predicates apply to which objects with the aid of the M constraint. However, he believed that the M constraint itself cannot be learned. Keil did not believe that the M constraint reflected some property of the world, but was instead built into human minds to allow for rapid inference. He thought that the M constraint was a powerful tool for learning, but as he did not believe it was a reflection of real world structure, and as it appeared in even the youngest children he tested, he did not think it was learned.

In this paper, I will not take Keil's assumption about the innateness of the M constraint for granted. I will address both the question of how the specific hierarchy of predicates and objects can be learned, and also whether the M constraint itself could be learned.

A note on language and concepts

Keil addressed predicability largely as a matter of conceptual representation and reasoning, but both he and Sommers discussed its applicability to language data as well. For instance, Sommers's example of the word *bat* as potentially violating the M constraint is an example of lexical but not conceptual ambiguity. All of Keil's empirical evidence about the M constraint was gathered via asking people for judgments about English or Spanish statements or in response to questions in those languages. My statements about what the M constraint constrains should be taken to apply primarily to conceptual behavior in humans. However, language provides a reflection of thought, and while language can be both more flexible and more ambiguous than thought in many cases, we may expect to see the M constraint reflected in language as well.

People presumably use information from both real-world encounters and linguistic input as sources of information about predicability. Thus, both an actual yellow banana and the phrase *a yellow banana* can potentially count as predication events. Obviously, language data is not the same as real-world data; we can utter the phrase *hour-long banana* even if we cannot conceive of such a thing, and we can describe things that we can conceive of using figurative language that is not literally possible. Even *hour-long banana* can relatively easily be assigned meaning; it could be a banana that takes an hour to eat, or that is only ripe for an hour, for instance. However, I expect literally sensible object-predicate pairs to appear more regularly in language data than pairs which can only be made sense of through metaphor. In determining whether a phrase that is used in language is literally sensible or only figuratively sensible, Keil suggests the paraphrase test: does a paraphrase of the original also make sense? One can talk about meaningfully about "killing an idea", but "forever ending the life of an idea" or "causing an idea to cease its vital bodily functions" does not make sense; thus, an idea cannot sensibly be killed.

Language-based data is easier to gather than data about real-world predication events. Thus, I will continue in the tradition of Keil and investigate the M constraint in part using real-world data involving language, while keeping in mind the ways in which language data potentially differ from conceptual representations.

Past work on sense and nonsense

Past work on sensibleness comes from a number of different fields of cognitive science, addressing the question from different angles. Keil's initial work describing the M constraint was followed by a number of related studies on children's learning and inference using the M constraint (Keil, 1983; Keil & Kelly, 1985), and responses to this work arguing that the M constraint was invalid (e.g., Carey, 1983) — an argument which we shall return to in the General discussion. Keil showed how children can use the M constraint to rapidly make inferences about the location of a new object in an existing predicability tree (1983), but none of the papers in this body of work suggested a method by which the M constraint could be used to learn an entire predicability tree based on data about individual true predicate-object pairs. Nor did they suggest a method for learning the M constraint itself; as discussed above, Keil believed the M constraint to be innate.

Other psychological work has addressed *plausibility*, a concept related to sense and nonsense, but also to truth. The plausibility of a sentence is one of several factors that governs how easily the sentence is processed by a reader or listener (see, e.g., Gibson & Pearlmutter, 1998; Pearlmutter & MacDonald, 1992). Both semantic anomalies and statements that are untrue or unlikely have low plausibility and can cause slower comprehension. Similarly, both nonsensical sentences and unlikely sentences generate an N400 response in the brain — "He took a sip from the waterfall" (an unlikely statement) elicits a smaller N400 wave than "He took a sip from the transmitter" (a category mistake), however (Kutas & Hillyard, 1980).

To the best of my knowledge, no work has been done to establish how people learn to judge semantic anomaly in the plausibility literature, though work on adjective-noun plausibility specifically shows that the plausibility of a given adjective-noun pair is highly correlated with the frequency of that pair co-occurring in a corpus of natural language data (Lapata, McDonald, & Keller, 1999). This suggests that pairs that are both true and sensible could potentially be learned as being sensible from language data (though the causal direction of the correlation with the frequency data is unclear — possibly the linguistic data are merely a reflection of facts that people learn in some other manner, rather than providing some of the data by which they learn). But it does not suggest any method by which people could learn to distinguish the class of object-predicate pairs that are untrue, yet still sensible from those that are neither true nor sensible.

Linguistics provides a number of approaches to predicability. Verb selectional constraints govern the semantic types of arguments that verbs take. Resnick (1993) has postulated the importance of a hierarchical organization of concepts in reasoning about predicability, but neither he nor other cognitive scientists studying semantic selectional constraints have proposed a method for learning a predicability tree from input data. VerbNet represents semantic information about verbs in a hierarchical structure, specifying the ontological categories that verbs can apply to (Kipper, Dang, & Palmer, 2000). However, the verbs are not organized in the same kind of object-based ontology that a predicability tree is, instead being organized by types of actions. FrameNet attempts to represent semantic frames that contain information constraining how words may interact; however, the organization of the frames is not strictly hierarchically constrained, and the lattice is organized by frames, which do not map neatly onto any element of a predicability tree (Baker, Fillmore, & Lowe, 1998). WordNet implicitly represents ontological structure through its organization of nouns, and represents some information about the arguments that verbs can take, but does not explicitly map out clusters and relationships of nouns and predicates (Fellbaum, 1998). These structures, while each similar in some ways to a predicability tree, do not contain all the information of a predicability tree. They are also constructed in part or entirely by human users rather than being learned.

A number of artificial intelligence projects exist that have tried to accumulate commonsense knowledge about the world. Projects such as Cyc (Lenat, 1995) and

OpenMind (Stork, 1999) have attempted to gather true statements about the world both from human input and from other sources such as the World Wide Web. Many of these projects also include inference engines which help draw further conclusions from existing facts. Most of these projects have tried to amass a much greater body of knowledge than just what predicability knowledge, learning what is likely (e.g., that cows are not green, and that people usually have to pay a bill before leaving a restaurant). To the best of my knowledge, no work in any of these fields has attempted to explicitly learn predicability, and thus to be able to distinguish sense and nonsense as we present it here.

Ontology learners (e.g., Snow, Jurafsky, & Ng, 2006; see also Buitelaar, Cimiano, & Magnini, 2005 for overview) are more analogous to the current learning endeavor. However, while methods for learning ontologies from input data exist, solving this problem alone is insufficient to help a learner make predicability judgments. Likewise, some work has been done to automatically learn verb relationships (Chklovski & Pantel, 2004), including clusters of similar verbs, but this work has not learned the relationships of these predicate clusters to object categories. A full-fledged predicability tree, containing both clusters of objects and clusters of predicates, and constraining the relationships between these clusters, is necessary in order to distinguish sense and nonsense. In the next section I present a model for learning a predicability tree using the M constraint.

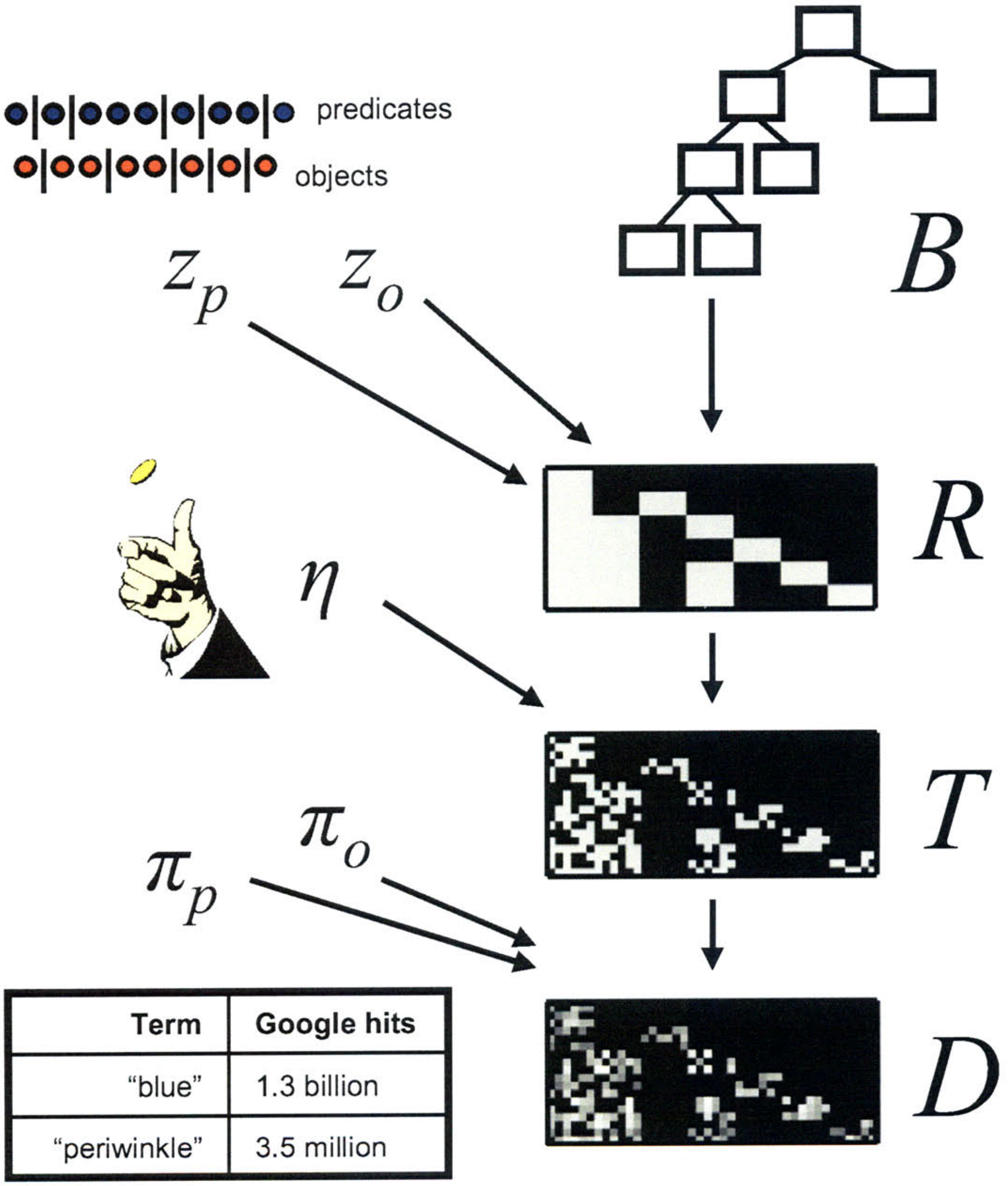
A generative model of sensibleness

In this section I will describe a generative model of sense and nonsense which incorporates the M constraint. The model describes the process by which object-predicate pair frequencies are generated in the data set (e.g., the number of times *yellow banana* is seen). Inverting this model, allows us to take the observed data and figure out the most probable model settings that would have generated that particular data – as an example, if the phrase *yellow banana* shows up frequently in speech (one possible source of data), then we will infer that our model parameters should probably be set up to allow *is yellow* to predicate *banana.* This section provides an overview of the model; for the formal mathematical details, see Appendix A.

The model described in this section is illustrated in Figure 4-2. The generative model begins by specifying what is sensible and what is not via a predicability tree, B. This tree by definition enforces the M constraint, and specifies which object-predicate pairs are sensible, shown in corresponding sensibleness matrix R. From there, a subset of the object-predicate pairs which are sensible are also true in the world (yellow bananas are both sensible and true, whereas blue bananas are sensible but don't generally occur in truth). To determine whether a sensible pair is also true, a coin with weight η is flipped. The results are shown in truth matrix T. For the purposes of this model, sensibility and truth are both considered to be binary distinctions in this model, and there is presumed to be a single relevant hierarchy that objects can be organized into for determining sensibility. These assumptions provide a simplified learning problem to begin with by reducing the computational complexity of the search for the best tree to match the data. However, the general approach of this model is valid without all of these assumptions, and each of them could be relaxed in future work. See the General discussion for further discussion of these points.

The frequencies of object-predicate pairs in the data set (D) are generated in part based upon whether the pair is true in the model. For instance, blue bananas would be generated less frequently than yellow bananas, if the model has parameter settings corresponding to our beliefs that blue bananas are not true and yellow bananas are (though this is a soft constraint; blue bananas are not deterministically prevented from occurring). The frequency of an object-predicate pair also depends upon the popularity of both the predicate and the object (π_p and π_o, respectively). That is, "blue bicycle" should show up more frequently in the data than "periwinkle bicycle" if "blue" is a much more popular term than "periwinkle" in general, and "blue bicycle" should show up more frequently than "blue unicycle" if "bicycle" shows up much more often than "unicycle."

In addition to the above factors, the model has a prior preference for simpler



Term	Google hits
"blue"	1.3 billion
"periwinkle"	3.5 million

Figure 4-2: The generative M constraint model. The blue and red dots represent the predicates and objects, divided into a set of clusters, z_p and z_o. The clusters are arranged at the nodes of the tree, B, yielding a corresponding predicability matrix, R. To find out whether an object-predicate pair which is sensible is also true, a coin with weight η is flipped, generating the truth matrix, T. The truth of an object-predicate pair combines with the popularity of the object and the predicate (represented schematically by the number of Google hits – one possible source of popularity data) to generate the frequency with which the pair is seen in the data, D.

trees. That is, before seeing any data, the model prefers trees with fewer nodes. The advantages of this prior probability will be discussed later.

The model allows us to compute the joint probability:

$$P(D, T, R) = P(D|T, R)P(T|R)P(R) \tag{4.1}$$

This joint probability indicates how probable a particular predicability tree, truth matrix, and set of data observations are to occur together. We can then invert the model to calculate the posterior probability of the truth and predicability matrices given the data:

$$P(T, R|D) \propto P(D|T, R)P(T|R)P(R) \tag{4.2}$$

How can we recover the predicability tree and truth matrix that best explain the data? It is a computationally infeasible process to evaluate how probable every possible set of parameters is given the data — especially for a large data set. We therefore use a stochastic search process to identify the values of the parameters that best fit the data. The search process is described in Appendix A.

Alternative methods

In seeking comparisons for the M constraint model, I looked for other methods that would organize objects and predicates into a tree and allow inference about which predicates apply to which objects based on this tree structure. Hierarchical clustering (Duda & Hart, 1973) is a popular method for learning tree structures involving objects only; it was used in Chapter 2 to represent the tree structure of the Gazoobian objects based on similarity ratings. To review the process of hierarchical clustering: each item starts in its own cluster, and then the two most similar clusters are joined to form a larger cluster. This process continues until all the items have been connected in a single strict hierarchy.

Hierarchical clustering can be used to cluster all the objects by calculating the similarity of two objects i and j based on the similarity of the two data vectors d_i

and d_j. However, there are two differences between such a tree and the structures learned by the M constraint model: first, the hierarchical clustering tree branches maximally, with each object ending in its own node, therefore it does not allow us to recover the ontological categories of the objects which allow us to generalize usefully about what properties objects are likely to have in common. More problematic is the fact that the predicates are not a part of the clustering process (except insofar as they help to cluster the objects); the predicates do not end up with a fixed location in the tree, and there is therefore no way of inferring which predicates apply to which objects. Hierarchical clustering is therefore insufficient for our needs. The same is true of Bayesian tree learning techniques proposed in the past (e.g., Kemp, Perfors, & Tenenbaum, 2004). D. M. Roy, Kemp, Mansinghka, and Tenenbaum (2007) propose a model for learning *annotated hierarchies* that allows probabilistic features to be applied to subtrees within an inferred tree, but predicate clusters are not learned.

I adapted the hierarchical clustering algorithm so that it would construct a full predicability tree and serve as a reasonable comparison model for the M constraint model. The modified algorithm starts with the same object-clustering method described above to form a maximally branching object hierarchy. I then developed a metric for scoring possible predicate locations and then placed each predicate at the best scoring node in the tree. See Appendix A for details.

This modified hierarchical clustering method does not make the same commitments to an underlying generative process and variables that the M constraint model entails. The hierarchical clustering model is not searching for combinations of truth and predicability matrices, but is instead maximizing the predicability matrix based directly on the data. If this method performs as well as or better than the M constraint model, it is computationally easier to compute and involves fewer variables, so it has an overall advantage in terms of simplicity. If the M constraint model performs better, then that will provide support for the theoretical assumptions underlying the M constraint model's generative process of sensibleness.

In addition to the two tree-based methods, I also tried a simple thresholding method that involved no generalization from the data. The thresholding method predicted that any predicate-object pair that had been seen one or more times in the data set was considered sensible, and all other pairs were not sensible.

Can sensibleness be learned using the M constraint?

The first question I ask in testing the performance of the M constraint model is: Can the model properly recover the underlying predicability tree from a data set when the process used to generate the data is exactly that described by the model (with a small amount of noise added)? Can it do so better than the comparison methods described above?

As I have discussed, the M constraint model makes a number of simplifying assumptions, so most real world data sets of predication events are unlikely to actually be generated according to this process. However, using data sets generated according to this process is a useful first test of the M constraint model. This is also by no means a trivial test; the model must infer several layers of hidden parameters, clustering predicates and objects simultaneously, and finding the best combination of truth and predicability matrices. Using a noisy version of the generative process described by the model additionally allows us to vary different parameters and observe each of their effects on the performance of the M constraint model and comparison methods.

Data

I generated two data sets based on a tree with the same structure as that shown in Figure 4-1, thus using a tree that approximated a real world predication tree as the basis for generating our data. Each node contained six predicates and three objects, with the exception of one internal node containing six predicates and no objects, yielding a total of 42 predicates and 16 objects. This yielded the predicability tree shown at the top of Figure 4-3.

I sampled the popularity parameters, π, from a uniform distribution. λ, the parameter governing the penalty for violating the truth matrix, was set to 10. For different data sets, I generated truth matrices from predicability matrices using biases

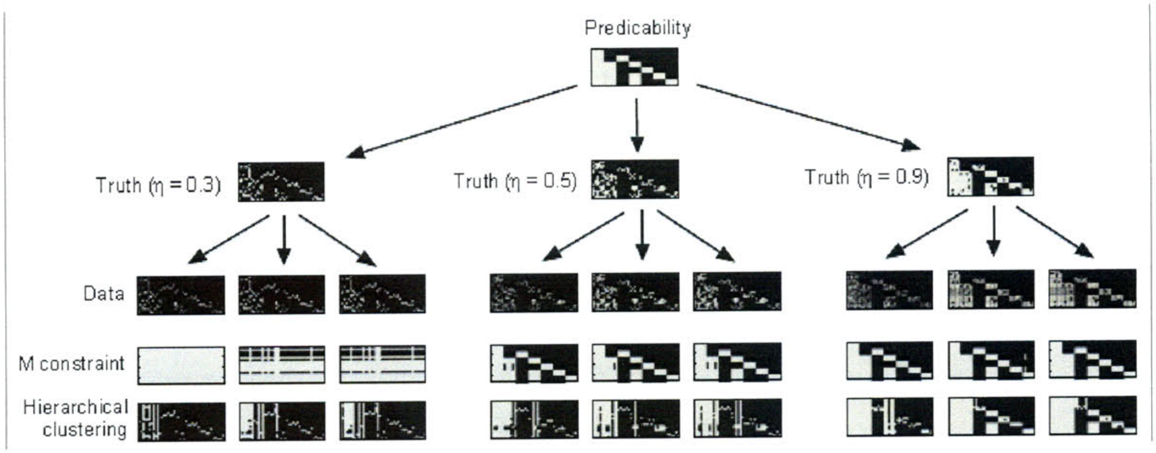


Figure 4-3: Data sets and results for the M constraint model and the hierarchical clustering model. The three truth matrices were generated by selecting random subsets of predicable pairs. The input for each model is a frequency matrix generated by sampling pairs from the truth matrix. Three data sets with increasing numbers of observations are shown for each truth matrix. The final two rows show predicability judgments made by the two models.

of $\eta = \{0.3, 0.5, 0.7, 0.9\}$. For example, a value of $\eta = 0.3$ means that a predicate-object pair that is sensible has a 30% chance of being true.

Having set these parameters, I sampled data according to our model. We generated two truth matrices per η value. For each truth matrix generated, I generated three data sets with different numbers of samples, N, where N took the values {1000, 10000, 100000}. The average number of times each event which was both predicable and true was seen in the data set ranged from 3.63 to 1089. Additionally, 1% of the data points were sampled randomly across all pairs, creating some noise in the data set. The generative process and a selection of the resulting data sets are illustrated in Figure 4-3.

These data sets were presented as the observed data D to the M constraint model and the comparison methods, and the best predicability matrices were inferred by each model.

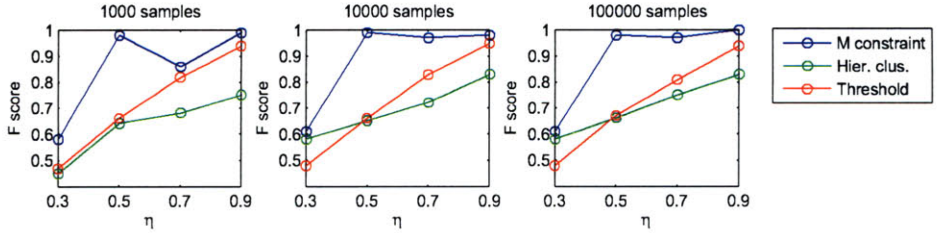


Figure 4-4: The performance of the three models on the simulated data sets. The M constraint model outperforms the modified hierarchical clustering model and simple thresholding for all values of η.

Results

A sample of the matrices recovered by the M constraint model and hierarchical clustering method are shown in Figure 4-3. The matrices recovered by the thresholding method are not shown; they are binary versions of the data matrices shown in the figure.

The recovered predicability matrices, R', were compared to the original underlying predicability matrix, R, using the F-measure — the harmonic mean of precision and recall. A high precision score indicates that most of the pairs that are predicted to be sensible in R' actually were sensible in R. If H is the number of "hits", or predicate-object pairs that are correctly predicted to be sensible in R', and FA is the number of "false alarms", or predicate-object pairs that R' incorrectly predicts as sensible, then

$$\text{precision} = \frac{H}{H + FA} \tag{4.3}$$

A matrix R' is said to have high recall if most of the predicable pairs in R are successfully recovered in R'. If M is the number of "misses", or pairs that are actually predicable according to R but not recovered in R', then

$$\text{recall} = \frac{H}{H + M} \tag{4.4}$$

where H is again the number of hits. Then the F-measure is the following:

$$F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{4.5}$$

F penalizes the overall score more greatly for a low precision or recall than either the arithmetic or geometric mean of the two values, and is therefore a conservative measure of performance.

The average F-measure for the recovered predicability matrices across the different data sets is shown in Figure 4-4. The truth matrices recovered by the M constraint model were all very close to the actual truth matrices, with $F \geq 0.97$ in all cases; the details of those results are omitted.

With relatively dense data, the M constraint model performs nearly perfectly, frequently recovering the original predicability matrix. The comparison methods also perform well with dense input data, though neither of the alternative methods recovered the underlying predicability matrix perfectly for any of the test cases.

The M constraint model outperforms both the thresholding method and the hierarchical clustering method on all the data sets; however, given the data sets corresponding to the sparsest truth matrices ($\eta = 0.3$), none of the algorithms perform well. In those cases, the M constraint model drastically overgeneralizes, while the other two methods both do the opposite. I would suggest that the M constraint model is performing in an intelligent and psychologically plausible way when working from these sparse data sets, and I will explore the reasons behind this behavior in the next section.

There is a dip in the performance of the M constraint model for $\eta = 0.7$ as compared to $\eta = 0.5$ when the number of samples is small ($N = 1000$), although it still outperforms the comparison methods in this case. This appears to be an artifact of the particular samples generated in one of the two data sets; for that data set, one of the nodes of the recovered tree was incorrectly placed, though the tree structure was otherwise correct. This pattern is not seen when there are more data samples drawn. I predict that, averaged across larger numbers of data sets, the M constraint model would perform at least as well for the $\eta = 0.7$ cases as the $\eta = 0.5$ cases.

Interestingly, the hierarchical clustering method is outperformed by the simple thresholding method as well as the M constraint model on many of the data sets, particularly the denser ones. Why does hierarchical clustering perform relatively poorly? The thresholding method does not generalize at all, and does not learn any underlying structure to fit the data. It simply fits the data it has seen exactly, and if, e.g., 90% of the predicable pairs are also truthful, and are sampled at least once in the data set, then the thresholding method will recover a large number of predicable pairs successfully in the densest data sets. Thus, the thresholding method performs well whenever T is close to R and there is a sufficient number of samples and a low level of noise. The hierarchical clustering model, by contrast, often overgeneralizes because of the tree structure it infers.

The M constraint model, given input data, must learn an underlying structure – but it infers two separate levels of structure, postulating both what is sensible and what is true. The M constraint model is therefore able to learn that some number of sensible pairs are not true and thus do not show up in the data even though they are predicable in the recovered R'. Hierarchical clustering, on the other hand, has to build the best structure possible to accommodate both the pairs that have been seen in the data set and those that have not been seen – and it lacks a distinction between possibility and truth. The hierarchical clustering method therefore has no way to explain why a pair does not occur in the data set except to make it not predicable in the tree, and it penalizes predicates that span objects with which they do not occur in the data set. Thus, hierarchical clustering, while a simple and elegant method for clustering objects, lacks the power of the generative M constraint model for inferring sensibility based on the indirect evidence found in such data sets as these.

The M constraint model recovers a tree that is very close to the correct one for values of $\eta \geq 0.5$ (with the sole example for $N = 1000$ discussed above. But should we expect that at least 50% of predicate-object pairs that are sensible in the world are also true? This is probably not the case. However, as the number of objects and predicates per tree node increases, overlapping properties will be observed even for lower values of η. Therefore, with more predicates and objects in the data set, it should be possible to infer the correct predicability tree even from sparser data.

Overall, the M constraint model's structural assumptions give it an advantage over the comparison methods when attempting to infer predicability matrices used to generate data sets according to this generative process. The M constraint model outperforms the comparison methods on simulated data sets, and given sufficient data, it recovers the correct underlying predicability matrix.

The developmental parallel

I now take a closer look at the inferences of the M constraint model when the input data is sparsest. These results were the most surprising in the previous section, since the predictions of the M constraint model and the other methods differed so drastically in these cases. Why did the M constraint model overgeneralize based on such sparse sets of observations?

I would suggest that the M constraint model is doing something both intelligent and psychologically plausible in the cases where it overgeneralizes based on sparse data. Consider a parallel case for a child learner. When someone learning about the world has only made a few observations, then her confidence about the structure of the world should be very low. Such a learner cannot yet distinguish between occurrences that are likely, but have not yet been observed, and ones that are extremely unlikely. She also does not know with any confidence which of the observations she has made so far are due to noise. Because of all this uncertainty, it would be unwise for the learner to speculate that the structure of the world is some highly complicated tree structure that fits the handful of observations she has made. It is better to postulate a simpler structure until the learner has more information.

Our M constraint model captures this intuition, postulating relatively simple trees until the data warrant otherwise. The hierarchical clustering method, on the other hand, is like a learner who creates maximally complex tree structures to fit every data set, regardless of how much evidence she has seen. The thresholding method does not learn a tree structure to fit the data set, but it also undergeneralizes in its predicability predictions in a way that does not match the human developmental

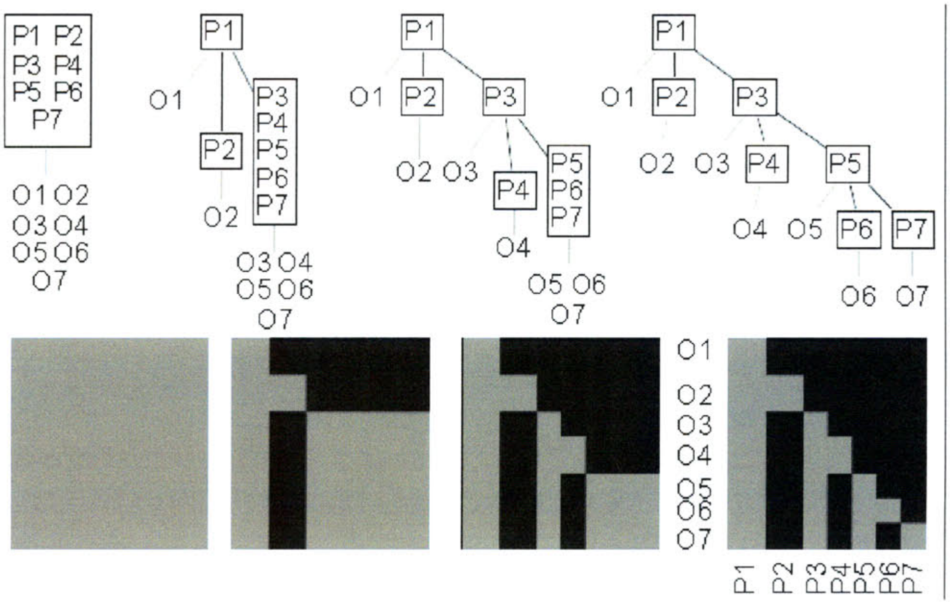


Figure 4-5: A developmental progression showing the predicability trees and corresponding predicability matrices learned by the M constraint model as it receives increasing amounts of data. The labels P1...P7 and O1...O6 represent clusters of predicates and objects, each of which has two or three members.

data. This difference stems from the fact that the M constraint model contains a prior distribution over tree structures that favors simpler trees. This *a priori* preference can be overcome by sufficient evidence about the structure of the world, but data sets that are too sparse do not overcome this preference.

To demonstrate in a more controlled way the behavior of the M constraint model when operating on sparse data, I generated a progression of simple data sets with the same underlying tree structure B as in the previous section, but with only $2-3$ predicates and objects at each node. I fixed the value of $\eta = 0.85$. Instead of sampling randomly to create a set of observations based on T and the popularities of the predicates and objects, every true predicate-object pair was observed a fixed number of times (N). Thus, for $N = 1$, every true pair was seen one time.

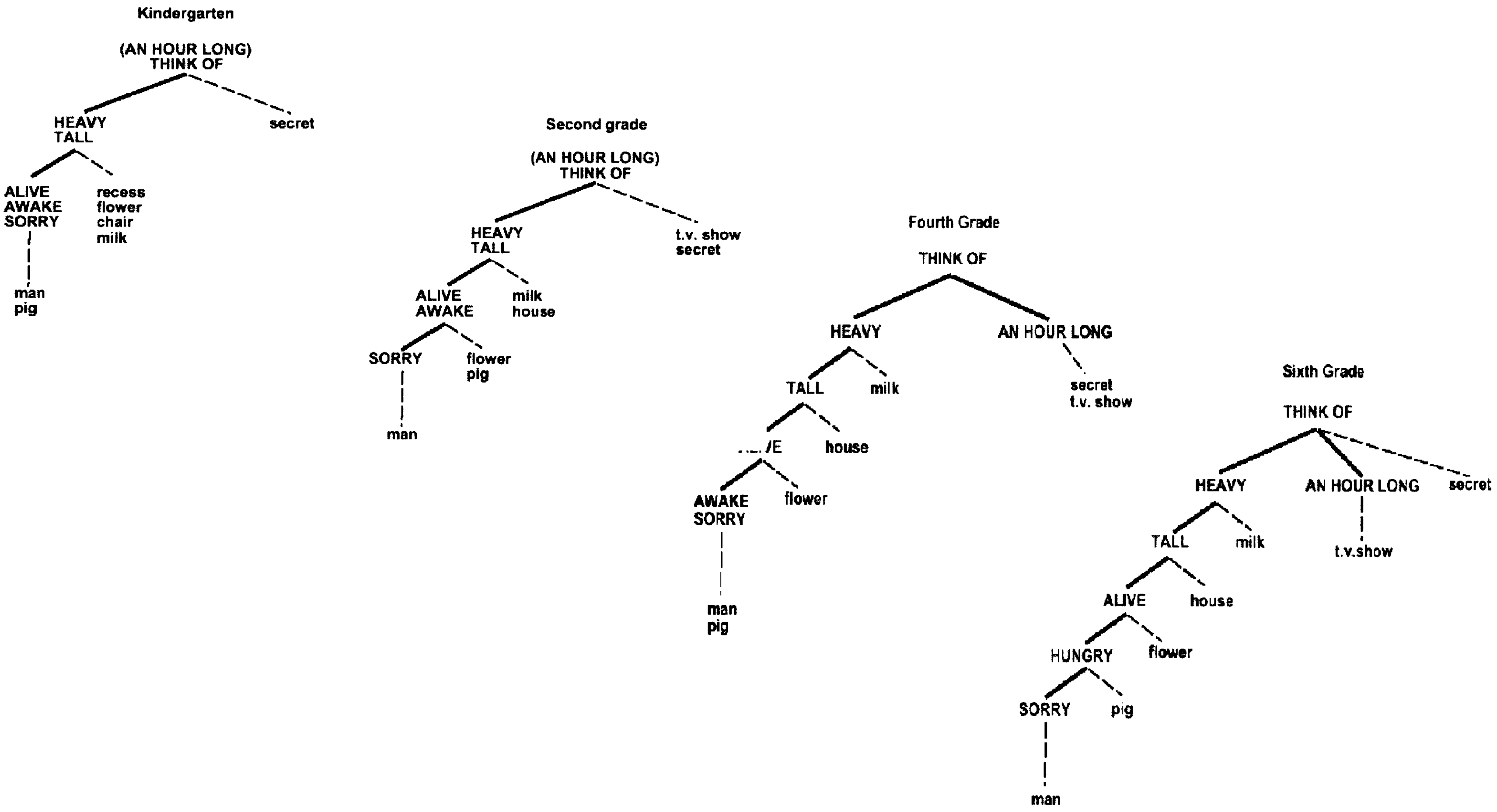


Figure 4-6: A sample of the developmental progression in predicability judgments given by children (figures originally from Keil (1979)). Each tree corresponds to an individual child's judgments.

I ran the M constraint model on data sets generated for $N = 1, 2, 3$, and 6. The best scoring predicability trees and corresponding matrices for each value of N can be seen in Figure 4-5. The developmental progression of the model is similar to the human developmental progression reported by Keil (1979) (see Figure 4-6). When the M constraint model has seen very little evidence, it behaves like the younger learners, choosing simpler trees. As more data are provided, the trees recovered by the model grow to look more like those of Keil's older subjects, who also had more data about the world. The similarity between these developmental curves argues for the psychological plausibility of a model that develops more complex theories only when sufficient evidence is available.

I end this section by noting that the developmental story could be a bit more subtle than Keil's story: children's judgments of what is *possible* show a more complex trajectory than that of the predicability trees, as described by Keil. Shtulman and Carey (2007) showed that children age 4–8 undergeneralize about what is possible at times, claiming that improbable events are impossible (for instance, stating that it is not possible to find an alligator under the bed, despite the fact that *alligator–is under the bed* is a sensible object-predicate pair, and one which adults judge to be possible). This is in contrast to the idea that predicability is overgeneralized throughout development. It is not clear how predicability and possibility differ based on the very different experimental paradigms, but Shtulman and Carey point out that various factors may be going into children's judgments of possibility – for instance, children may not admit the possibility of things that they find surprising, or events where they cannot imagine the circumstances arising. However, it seems likely that a model of sensibleness is a part of what constrains these judgments of possibility (children in these experiments uniformly denied the possibility of eating lightning and other category mistakes), and that other world knowledge then further constraints what a child thinks is possible. Children think that toy alligators can be found under the bed, and that household pets can be found under the bed, so clearly their understanding of the predicability of physical objects is not actually undergeneralized from adults — but some additional information causes them to judge this scenario impossible. The relationship between sense and possibility, and how children understand these ideas, remains to be further investigated. But a prior preference for simpler trees provides a powerful tool for avoiding overfitting of sparse data, and matches the developmental data provided by Keil (1979).

Learning sense and nonsense from real world data

The previous section shows that the M constraint model can perform well at the task of inferring sensibleness based only on noisy, sparse truth data when the data is generated from the model. On the one hand, this is no mean feat, as the model must infer several hidden layers of parameters. On the other hand, this is only informative about how the model performs when the data are generated according to its expectations. Given that there is reason to believe that the model's assumptions of how object-predicate pairings occur in the world are simplified in several ways, I also wish to examine how the model fares when presented with real world data.

Data

What data can be used to represent the object-predicate pairs that children encounter? How can data be gathered about how many times a human learner encounters yellow bananas, bicycles that are discussed in conversation, refrigerators that are visible, soccer games that last an hour, or toys that are fixed with screwdrivers? How can we know which object-predicate pairs are encountered by and salient to children as they learn about the world? Estimating this seems a difficult, if not impossible, task.

As an approximation to the problem, I choose to examine object-predicate pairings found in a natural language corpus. This has the advantage that object-predicate pairs are easily identifiable and countable. However, in many ways it is a poor approximation for the richer data gathered by observing the world directly. Language is filled with idioms, metaphors, and other figurative speech that allows for object-predicate pairings not actually seen in the world. It is possible, though by no means certain, that the model might treat these pairs which show up in language but not literally in the world as noise. Potentially more worrisome is the fact that language corpora are apt to be missing a great deal of important data about what predicate-object pairs commonly co-occur in the world. People rarely choose to comment on everyday occurrences like yellow bananas or visible refrigerators, so very common object-predicate pairs in the world are going to be very underrepresented in language data. It is therefore unclear whether language data can provide sufficient evidence for the M constraint model to learn sensibleness. Language data therefore provides a conservative test of how much the model can learn based on a child's evidence about the world.

In using natural language data, I had to decide how to find object-predicate pairs within the corpus. Keil's objects translated fairly obviously and directly to nouns. Some of the predicates used by Keil in his work were of the form *can be discussed.* Some were also adjectival, such as *is tall* or *is heavy.* In my corpus searches, I decided to focus on the interactions of nouns and verbs. For instance, in the sentence "The girl threw the ball", two object-predicate pairs are present: *girl-can throw* and *ball-can be thrown.* This corresponds to *girl* being the subject of *throw* and *ball* being the direct object. The predicates I gathered from natural language therefore take the form *throw (subject)* or *throw (object).*

I selected 68 objects (nouns) and 45 predicates that fit four criteria: (1) as a group, they covered all the ontological categories used by Keil, (2) individually, the nouns and verbs were as non-polysemous as possible (though avoiding polysemy entirely was impossible), (3) the nouns and verbs were relatively common words in general and high frequency in the chosen corpus in particular, and (4) the nouns co-occurred with more than one of the predicates in the data set (and vice versa). The non-polysemy requirement may or may not be necessary in a learning algorithm; I return to this in the General discussion. For the purposes of this experiment, however, I tried to ensure that there was approximately one sense per word used in the input.

I used the British National Corpus (BNC) as the source of natural language data (2007). The BNC consists of 100 million words of British English, taken from a wide

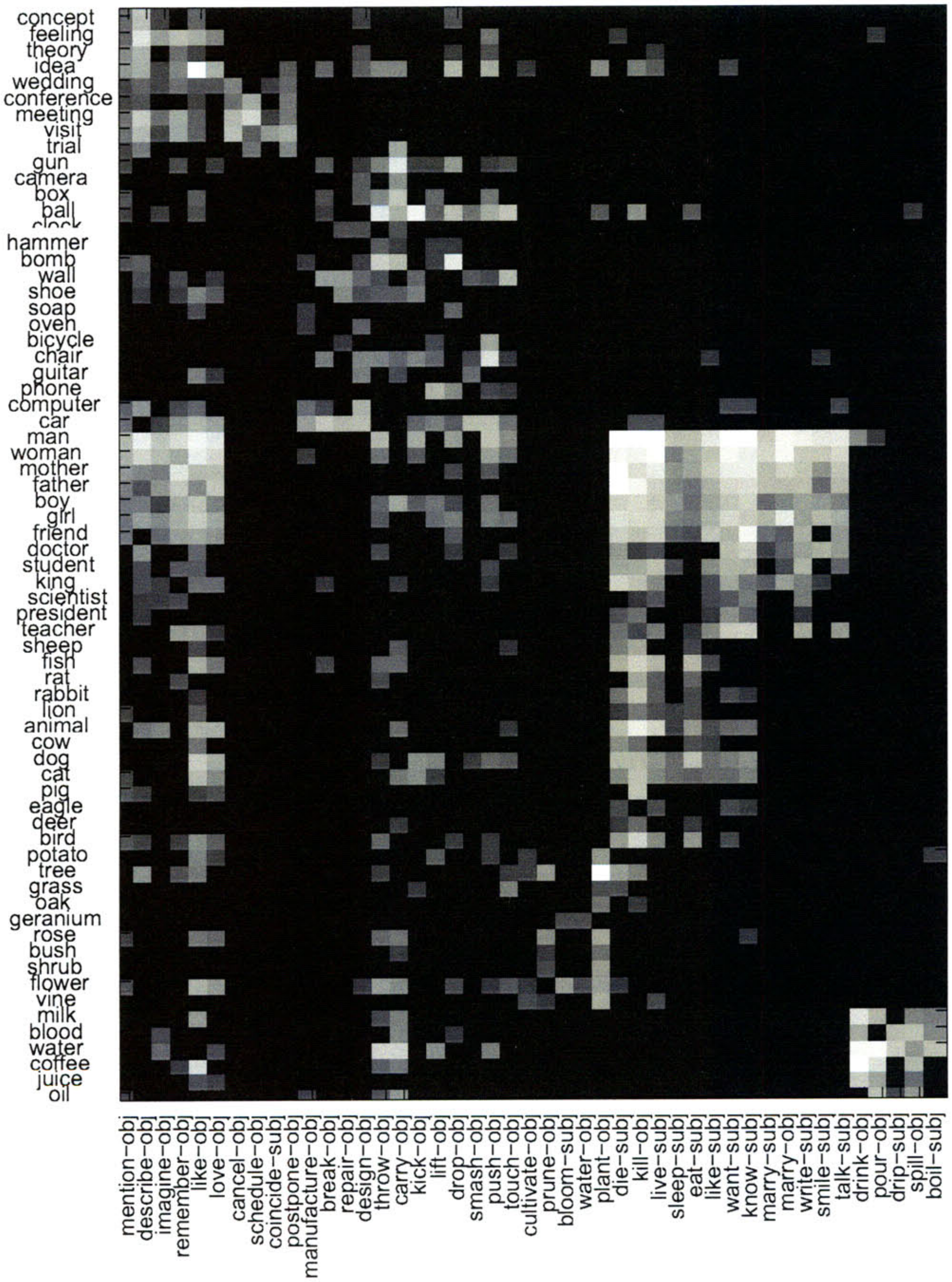


Figure 4-7: The data gathered from the British National Corpus, shown here in log frequencies.

variety of sources, both spoken and written. In searching the corpus, I used the Sketch Engine tool (Kilgarriff, Rychly, Smrz, & Tugwell, 2004) to search the BNC for collocations of each noun-predicate pair in our matrix. I recorded the number of times that each noun had occurred with each verb in the argument position(s) of interest. Some of the labeled co-occurrences were erroneous due to misparsings, but these were left in the data set.

Each predicate was seen an average of 270 times (min = 6; max = 1341), with an average of 14.4 objects (min = 2; max = 43). Each object was seen an average of 179 times (min = 6; max = 2116), with an average of 9.5 predicates (min = 2; max = 26). The average number of times that each object-predicate pair predicted to be sensible was seen in the data was 8.8 (min = 0; max = 336). The data matrix, D, is shown in Figure 4-7.

I used this frequency matrix as input for the M constraint model. The tree that I predict the M constraint model should learn if it is able to recover sensibleness perfectly from the data using the primary sense of each word is shown in Figure 4-8.

Results

The tree learned during the by the M constraint model from the BNC data set is shown in Figure 4-9.

There are many obvious differences between the two trees, but some successes on the part of the M constraint model as well. First, I will discuss some commonalities between the predicted tree and the learned tree. Then I will address the areas in which they differ.

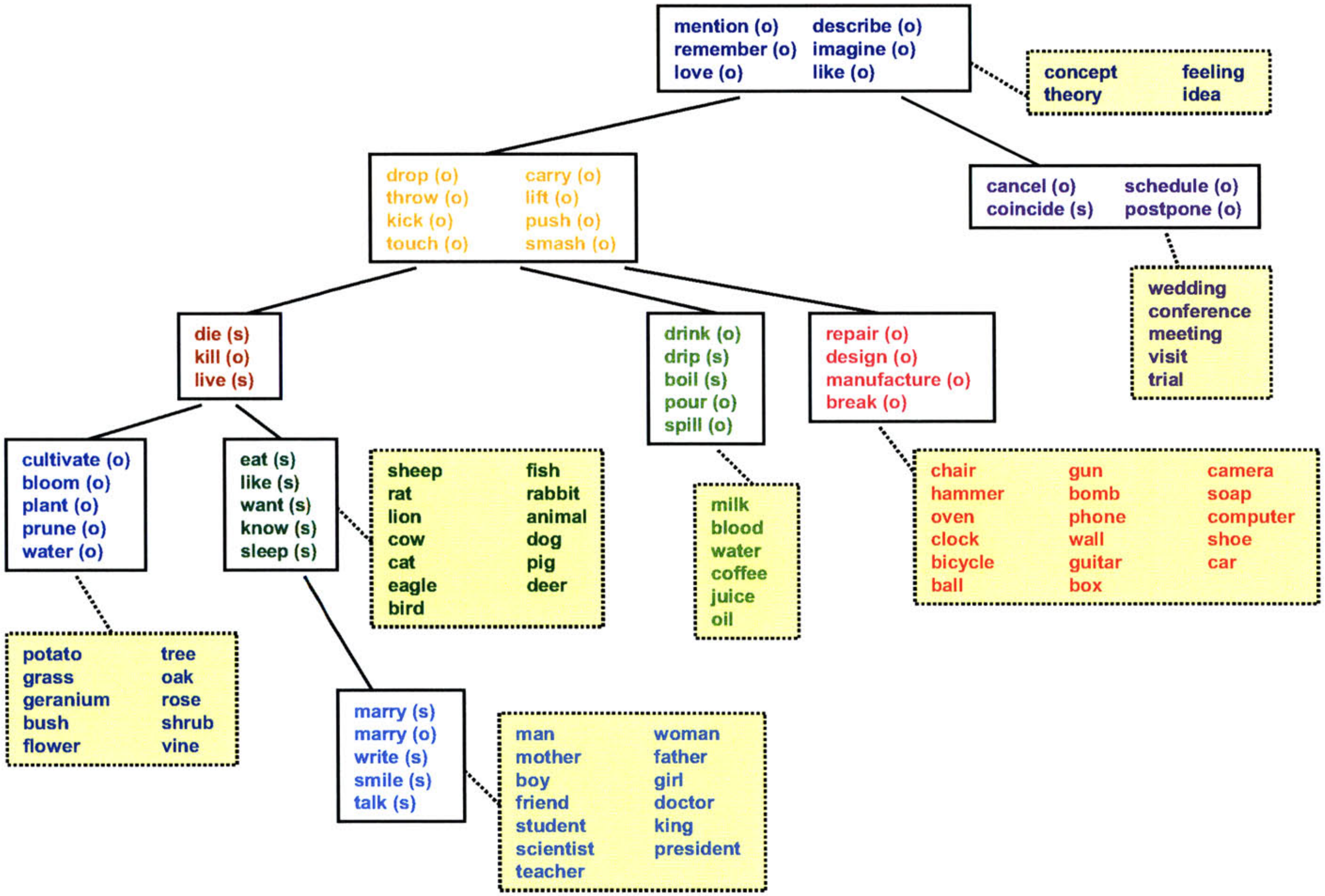


Figure 4-8: The predicted predicability tree incorporating the objects and predicates in the BNC data set.

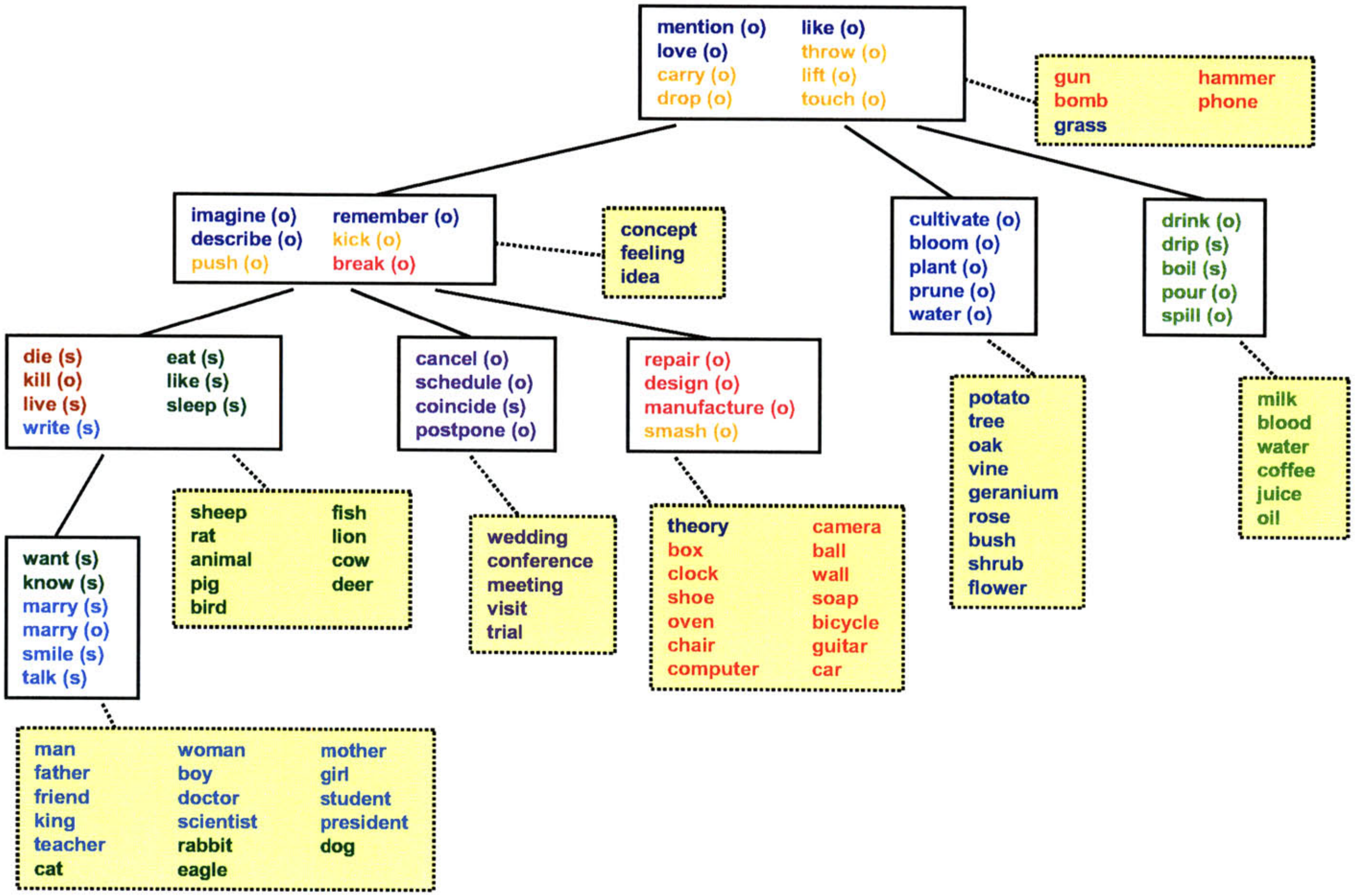


Figure 4-9: The structure recovered by the M constraint model from the BNC data.

The M constraint model successfully recovered many of the ontological categories that match human judgments about kinds. The liquids are grouped together, as are the events, the plants (with the exception of *grass*), and the animals, with the animals being mostly correctly divided in to humans and other animals. The artifacts and abstract items are less successfully divided, though the majority of the artifacts are clustered together, with only *theory* mistakenly placed in the group. The corresponding predicate clusters for many of these groups - plants, liquids, events, and animals particularly - have also been, for the most part, properly learned. Overall, if we compare the matrix corresponding to the learned tree to that corresponding to the predicted tree, the M constraint model's matrix has an F score of 0.81.

Despite the successes of the M constraint model, however, there are many things that are mismatched between the two trees, and we would be dismayed by an adult human who produced some of the sensibleness judgments licensed by the M constraint model's tree. The model commits three types of errors in learning from the BNC data: (1) undergeneralization due to sparse data, (2) overgeneralization due to uneven data sparseness, and (3) misclassification due to noisy data or figurative language.

The model's undergeneralizations about both plants and liquids are examples of error type 1. Due to the sparseness of the data, neither plants nor liquids are classified with the other physical objects. Examining the data set (Figure 4-7) shows that plants are represented very sparsely in the data set overall, and are only occasionally mentioned with any non-plant-specific predicates. Liquids occur more frequently in the data set, but again, mostly with liquid-specific predicates. There is little or no evidence in the data set that some of the abstract predicates like *describe (obj)* or some of the physical predicates like *smash (obj)* could occur with any of the plants or liquids; the little evidence that does occur that a predicate like *kick (obj)* could occur with plants - a single data point in the case of *kick (obj)* - gets discounted as noise. There is some evidence in the data set that some plants can *die* or *be killed*, however, it is insufficient to overcome the fact that the living things group has many other predicates that also apply to it (all the physical and abstract predicates), and that most of those have never been seen with plants. The model therefore prefers to treat the *die (subj)* and *kill (obj)* co-occurrences as noise, having seen no other evidence that the plants can have most of the abstract and physical predicates apply to them. Additionally, the model undergeneralizes about the object *grass*, which co-occurs only 3 times in the data with a plant-specific predicate (*plant-obj*) out of a total of 16 occurrences. The model treats this as noise, due to lack of further evidence that *grass* belongs with the plants.

The other objects at the root node of the tree (gun, hammer, bomb, and phone) are also undergeneralized for reasons of sparse data. They each occur an order of magnitude more with predicates at the root of the tree (e.g., in phrases like, "carry a gun" or "drop the bomb") than with other predicates.

Why does the model undergeneralize here given insufficient evidence, whereas with the developmental data it overgeneralized given sparse data? The difference is that here, there is strong evidence about the objects and predicates in some areas - there is evidence that plants and liquids are coherent clusters, and both have cohesive predicate clusters that apply to them. There is also evidence that most of the abstract and physical predicates apply to many other physical objects. Therefore, the model would predict that if those predicates could apply to plants and liquids, more data would have been encountered. In the developmental data, by contrast, there was initially insufficient data to predict any clusters with any confidence. Here, it is the relative sparseness of the most of the abstract and physical predicates with relation to plants and liquids versus other physical items that causes the undergeneralizations. Because the model predicts that frequencies depend only on truth of an object-predicate pair and the popularity of object and predicate independently, the model has no way to predict that people are less likely to talk about remembering or imagining plants or liquids than they are to talk about remembering a theory or imagining a wedding, despite all of those combinations being sensible and truthful. Therefore, the failure of those predicates to occur with the plants and liquids is accounted for by making them untrue of plants and liquids in the learned tree.

Further undergeneralization occurs in the category of living things, where the data about many animals is also sparse. There is little evidence for most animals other than cats and dogs sleeping or knowing things, even though neither knowledge nor mental representations are limited to humans. Thus, the predicates *sleep (subj)* and *know (subj)* are incorrectly undergeneralized to only apply to the smallest ontological class containing humans. Cats and dogs, as common pets, get discussed far more in literature than most animals, and so it is unsurprising that there are more predication events involving cats and dogs, and a wider variety of predicates applying to them in general, in the BNC corpus than other animals. Therefore, cats, dogs, and the few other objects shown to be capable of sleeping or knowing things are grouped with the humans — thus the model actually *overgeneralizes* about a few animals because it has *undergeneralized* about some of the predicates that should apply to all animals as well as to humans. Therefore, an uneven sparseness in the data about which predicates apply to which animals (with lots of data about only a few animals) leads to some overgeneralization about a few animals — an error of type 2.

This type of error also explains why the model overgeneralizes somewhat about which predicates apply to events. Why can events be *thrown, dropped, touched, carried, lifted, broken, kicked*, and *pushed* according to the learned predicability tree? This is a result of two factors. The first is the undergeneralization of plants and liquids, as described above. There is so little evidence that plants or liquids share most traits with physical objects that they are placed at the root node. However, this is ample evidence that plants and liquids can be *dropped, carried, lifted*, and all the other predicates that are at the top of the tree. Therefore, if these predicates predicates were not to apply to the events, then events would have to reside in a completely separate portion of the tree from everything else.

The events should be assigned to a separate subtree from all the physical objects, however! That's where they belong. So why aren't they? This is partly due to uneven sparseness of data: there is strong evidence that events can be *described, remembered,* and *imagined*, while there is little evidence supporting the idea that these apply to plants or liquids. The events are therefore moved underneath the node containing these predicates. In addition to the strong evidence that the events should be spanned by these three abstract predicates, figurative speech causes events to be grouped with some artifacts: *trial* co-occurs more with *carry (obj)* than any other predicate, all concerning clinical trials (e.g., "Two trials were carried out"). This use of *carry* also provides further reason for the model to overgeneralize about the application of physical predicates to events.

Theory is, apparently incorrectly, grouped with most of the artifacts. The data set shows that people sometimes speak of a particular theory being pushed, and they also talk of designing theories. While the first usage is a polysemous use of *push*, meaning *advocate*, it does not seem entirely anomalous to speak of designing theories. This could be seen as a figure of speech, but it could possibly be a violation of the M constraint. The latter possibility is picked up again in the General discussion.

Despite the limitations of the tree recovered by the M constraint model from the BNC data, the tree still manages to capture many human beliefs about ontological organization and predicability, as indicated by Figure 4-9 and the high F score. Considering that human learners have data about more predicates and objects, a greater amount of predication events to work from, and real-life encounters as well as natural language data, the M constraint model performs remarkably well on this limited data set.

Adding a physicality constraint

The results from the BNC language data are promising, but fall short of what we would hope a real-world learner to be able to produce. However, as discussed, the input data also fall short of what we would expect a real-world learner to observe in learning about the world around them. Object-predicate pairs that are so common as to be unremarkable (like yellow bananas) show up a great deal more in real-world encounters than in language data. And we commonly experience many predication events that involve predicates that are low frequency in language data. In the real world we do imagine a warm cup of coffee, we do hear flowers described, we do remember the old phone we had before we got our new cell phone. Additionally, the objects that we encounter in the real world have many properties that are true of them, but we generally only describe a small subset of those out loud or on paper;

we answer the phone, we drink our coffee, and we smell the flowers.

All of these examples point to ways in which language fails to capture the rich extra information we have from real-world encounters with objects, events, and things in the mental realm — for instance, it is generally fairly obvious to us whether an item is a physical, tangible item, as opposed to an event or abstract idea. If we physically touch, see, and interact with the item, we have ample evidence that it is physical, but we may never remark upon these properties (how often have you felt the need to point out that bananas are physical, tangible objects?). This fundamental distinction of physicality is one of many types of data not included in the data set.

I reran the M constraint model with an additional piece of information about each object. I specified whether or not each item was item was in the realm of physical objects or not. I implemented this as a firm constraint by requiring that all the physical objects be in a distinct subtree from all non-physical objects. Given this additional requirement, the M constraint model's results were greatly improved, despite the limitations that remained in the data set gathered from natural language input. Figure 4-10 shows the tree recovered by the M constraint model given the physicality constraint. The corresponding predicability matrix had an F score of 0.86.

There are many improvements in this recovered tree. Events and abstract ideas are now in their proper locations, and plants and artifacts are all much closer to their correct locations. The abstract predicates are all correctly grouped at the root node of the tree, as well.

The main remaining misclassification is that plants are still not classified as living things in this tree. The same sparse evidence that led these objects to not be classified as living things previously still causes a problem even given the physicality constraint. However, the placement of plants within the ontology now looks much more like children's early beliefs about plants. Children generally do not learn that plants are alive until well after pre-school age, and possibly as late as age 10 (Carey, 1985).

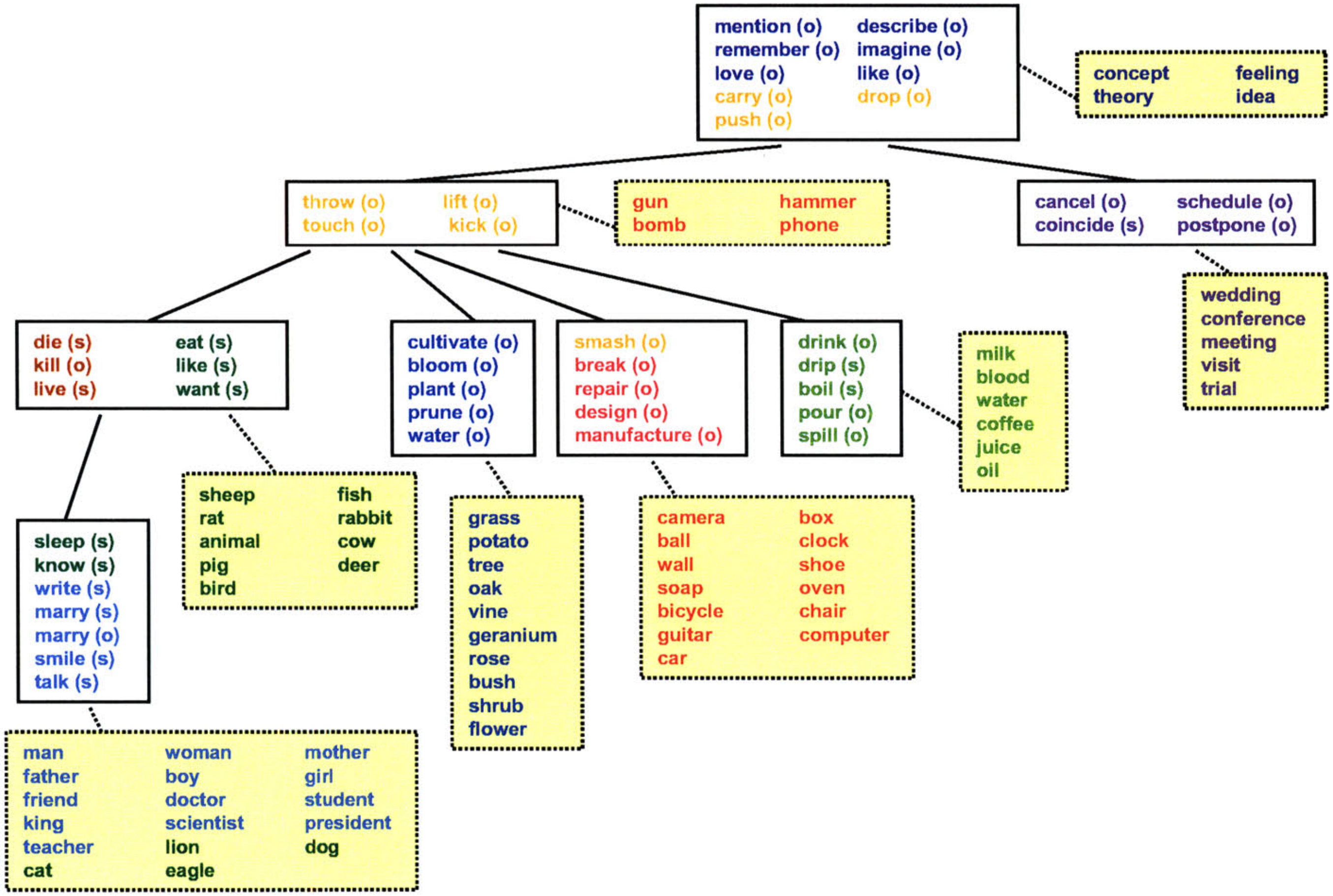


Figure 4-10: The structure recovered by the M constraint model, with added physicality constraint, from the BNC data.

For similar reasons of sparse data, the artifacts that were previously undergeneralized about and placed at the root of the tree are now placed at the root of the physical objects subtree. There is still undergeneralization about what predicates apply to these objects, but the objects have shifted closer to their correct position in the tree.

The physicality constraint — a single additional piece of data about objects, readily available from our everyday interactions with the world — improves the performance of the M constraint model. This makes it more plausible that the M constraint model could learn the correct sensibility judgments from the real world input that human learners are exposed to.

Additionally, the physicality constraint shows how the tree structure can change drastically based on a new learned fact. Children's mental representations can also shift radically based on new data; for instance, children's ideas about what it means to be a part of the biological domain change as they learn about the properties of animals and plants (Inagaki & Hatano, 2003; Opfer & Siegler, 2004; Carey, 1985). Both the M constraint model and children's notions of predicability can change for a large number of objects nearly instantaneously, based on new knowledge about predicates.

Overall, the M constraint model recovers a great deal of structure about the world given a very limited data set. Its performance is improved by adding information about which objects are physical objects. Human learners have a great deal more data about a great many more predicates and objects, much of it from non-linguistic input. However, even without all these advantages, the M constraint model performs remarkably well on real-world data.

Can the M constraint be learned?

I have shown that the M constraint can indeed be used to learn sense and nonsense given appropriate input data. The results of the BNC experiment suggest that a child learning from real-world experiences using the M constraint could indeed learn sense

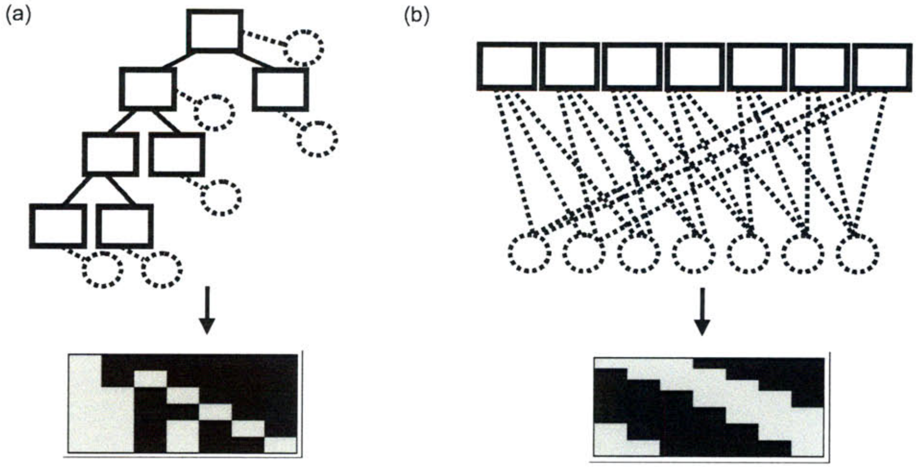


Figure 4-11: The two structures used to generate data sets for model selection, and the resulting data sets – one tree-consistent and one tree-inconsistent.

and nonsense. But can the M constraint itself be learned?

Keil postulated that a child learner must start with the idea that objects and predicates are organized in a strict hierarchy, or else they would not be able to infer the correct structure. What would a child have to do in order to instead learn that the M constraint was useful in organizing information about the world? She would need to have a hypothesis space that included a hierarchical organization as one possibility. She would also have to have a way of determining which possible hypothesis best fit the data in the world. And she would have to receive sufficient evidence to infer that a tree-based structure best described the organization of predicates and objects.

In this section I describe a method of model selection that can choose between multiple possible structures governing predicates and objects and determine which model is most likely to have generated a data set. This model selection method could be used to choose from among any number of possible hypotheses. However, as a proof of concept, I begin with only two models: the original M constraint model, and a comparison model which lacks the M constraint, called the flat model.

The flat model

The flat model is identical to the M constraint model, but lacks the hierarchical constraint. Predicates and objects still form clusters, and predicate clusters apply to objects clusters, but they are not constrained to form a tree. Thus, for instance, one predicate cluster could span object clusters 1, 2, and 3, while another predicate cluster could span object clusters 2, 3, and 4, an arrangement which breaks the M constraint (see Figure 4-11(b) for an example). I emphasize again that this model is only one of many models which a learner might consider in trying to learn about the structure of predication in the world. Model selection could equally well be used to choose among a large group of models as between these two. However, this model is an ideal starting point from a learning standpoint. Since the flat model is exactly the M constraint model without the hierarchical assumption, it can be used to investigate what a learner would pull out of the data if they simply presumed that similar objects and properties clustered, but had no other learning constraints, and I can thus investigate whether the M constraint itself can be learned from a data set. The flat model is also an ideal comparison model from a mathematical standpoint, as the two models produce easily comparable joint posterior probabilities $P(T, R|D)$ given the same data set. Therefore, I proceed in comparing a non-hierarchically constrained, or flat, model to the M constraint model as an interesting comparison in and of itself, and as a proof of concept of Bayesian model selection for learning which model is best supported by the data. The mathematical details of the flat model are given in Appendix A.

Bayesian model selection

In order to determine whether the M constraint or flat model – M_{tree} and M_{flat} respectively – is best supported by a given data set, I use Bayesian model selection. Given the data set D, I search for the combination of predicability matrix, truth matrix, and model with the highest joint posterior probability:

$$P(R,T,M|D) \propto P(D|M,R,T)P(T|R,M),P(R|M)P(M) \qquad (4.6)$$

Informally, this means I look for the best predicability and truth combination according to each model. Then I weight the scores of these winning combinations by the prior probability of the models that generated them.

I use equal prior probabilities for the two models:

$$P(M_{tree}) = P(M_{flat}) = 0.5 \qquad (4.7)$$

Let R_{tree} and T_{tree} indicate the best predicability and truth matrix under the M constraint model — those that maximize the joint posterior probability $P(R,T|D,M_{tree})$. Likewise, let R_{flat} and T_{flat} indicate the predicability and truth matrix that maximize $P(R,T|D,M_{flat})$. If the data are not consistent with a tree structure, then the flat model will contain the best (non-hierarchical) predicability matrix within its hypothesis space, but the M constraint model will not. Since R_{flat} and T_{flat} will be consistent with the data set and R_{tree} and T_{tree} will not, $P(R_{flat},T_{flat}|M_{flat}) > P(R_{tree},T_{tree}|M_{tree})$ — in such a case the flat model will perform better.

By contrast, when the data set is consistent with an underlying tree structure, then the best scoring predicability matrix and truth matrix will be in the hypothesis space of both models. Thus, R_{flat} and R_{tree} are identical in such a case, as are T_{flat} and T_{tree}. However, $P(R|M)$, the probability of the predicability matrix given the model, differs between the two cases. This is because the flat model is considering a much broader hypothesis space of predicability matrices than the M constraint model is. Given that all predicability matrices with the same number of predicate partitions are considered to be equally probable *a priori* within a given model, the probability of any particular R_{flat} will be will be far lower than the probability of the same matrix when it is R_{tree}, because the flat model is distributing probability mass across so many more matrices than the M constraint model.

Bayesian model selection chooses which model is most probable given an observed set of data. In the next section I illustrate this process using synthetic data sets.

Table 4.1: Log-posterior scores for the best possible configuration each model recovered given each set of simulated data.

Model	Tree-consistent data	Tree-inconsistent data
M constraint	**-94726**	-94255
Flat	-94748	**-93933**

Model selection with simulated data

Data

I generated two data sets of predicate-object pair frequencies based on two underlying predicability matrices. One of the predicability matrices was the same predicability matrix used to generate synthetic data in the first set of simulations, and was thus consistent with a hierarchical structure. The other predicability matrix did not correspond to any tree structure. Both predicability matrices contained the same number of predicates and objects, and the same number of predicate partitions. Figure 4-11 shows the predicability matrices and the structures that generated them.

Frequency matrices were generated by sampling each predicable pair 100 times. No noise went into the process.

Results

Table 4.1 shows the log-posterior probabilities for each model given the two data sets. As expected, the flat model outperforms the M constraint model on data that is not tree-consistent; the M constraint model is unable to find a tree that produces predicability judgments at all close to the at of the data set and must therefore overgeneralize greatly to find a tree that could produce those predicability judgments. Figure 4-12(c) shows how the best predicability matrix recovered by the M constraint model overgeneralizes.

In the case where the data set is consistent with a tree structure, both models recover the same predicability and truth matrix, but the M constraint model scores better than the flat model, for the reasons described above. The difference in perfor-

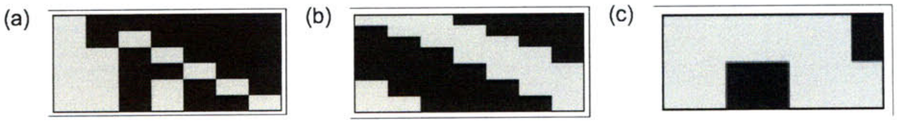


Figure 4-12: (a) The predicability matrix recovered by both models, given the tree-consistent data. (b) The predicability matrix recovered by the flat model, given the tree-inconsistent data. (c) The predicability matrix recovered by the M constraint model, given the tree-inconsistent data.

mance may appear subtle, but the scores are log-posteriors and represent a difference of 22 orders of magnitude.

These results confirm the intuition that a learner with a hypothesis space including several different representations could choose the representation best supported by a given data set. The simulation above shows how Bayesian model selection can be used to determine the model best supported by the data. The hypothesis space of models could be much larger than that used here, but the flat model and M constraint model provide an example of how the model selection process works.

Model selection with real-world data

The simulation used two data sets where there was ample evidence as to whether the data set was tree-consistent, and where every predicable pair was observed in the data set. As we have seen previously, however, real-world data sets can contain noise, and may have sparse or missing data for many predicable pairs. Due to polysemy, language data sets may even include a few exceptions to the M constraint. How clearly do the data have to be tree-consistent in order for the M constraint to be learned from the data set? Does real-world data provide sufficient evidence of being tree-consistent for a child learner to learn the M constraint?

The BNC data set used to test the M constraint model is unlike the data set encountered by real-world learners for all of the reasons discussed in previously. However, I compared the M constraint model and the flat model on the data set, both with and without the physicality constraint. In the flat model, the physicality constraint was implemented by requiring that physical and non-physical objects never

Table 4.2: Log-posterior scores for the best possible configuration each model recovered given the BNC data set, both with and without a physicality constraint.

Model	Unconstrained	With physicality constraint
M constraint	-31439	-31472
Flat	**-31392**	**-31409**

be clustered together.

Results are shown in Table 4.2. The flat model had a higher log-posterior score than the M constraint model on the BNC data set, both with and without the physicality constraint. Therefore, under model selection, the flat model is chosen as best representing the data. However, the predicability matrices recovered by the flat model do not agree as well with the predicted sensibility judgments (from Figure 4-8) as the predicability matrices recovered by the M constraint model: $F = 0.75$ vs. $F = 0.81$ without the physicality constraint; $F = 0.76$ vs. $F = 0.86$ with the physicality constraint. Figures 4-13 and 4-14 show the structures recovered by the flat model in each condition.

As the BNC data set is not representative of the real data encountered by human learners in a number of ways discussed previously, this result is inconclusive. If real-world data is not clearly tree-consistent — if the data are sparse, noisy, or lacking in data about certain ontological categories, as in the BNC data set — then having the M constraint innately built in could be a valuable aid to the child in learning about what is sensible and what is nonsense as they learn about the world around them. However, if the data that the child encounter are far richer than those of the BNC data set *and* are sufficiently tree-consistent, the child learner can learn the M constraint as well as learning the best predicability tree to match the data around them.

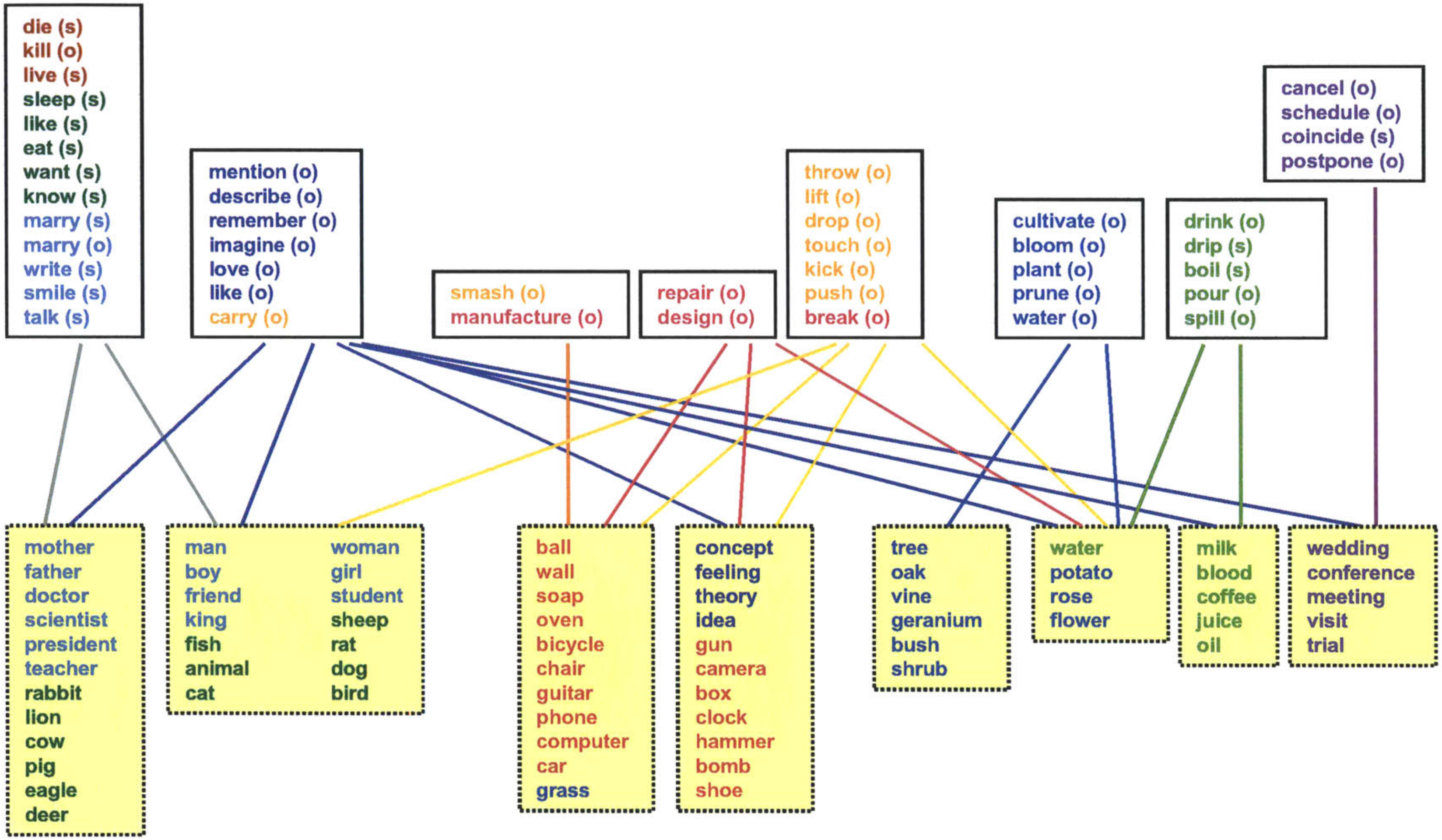


Figure 4-13: The structure recovered by the flat model from the BNC data.

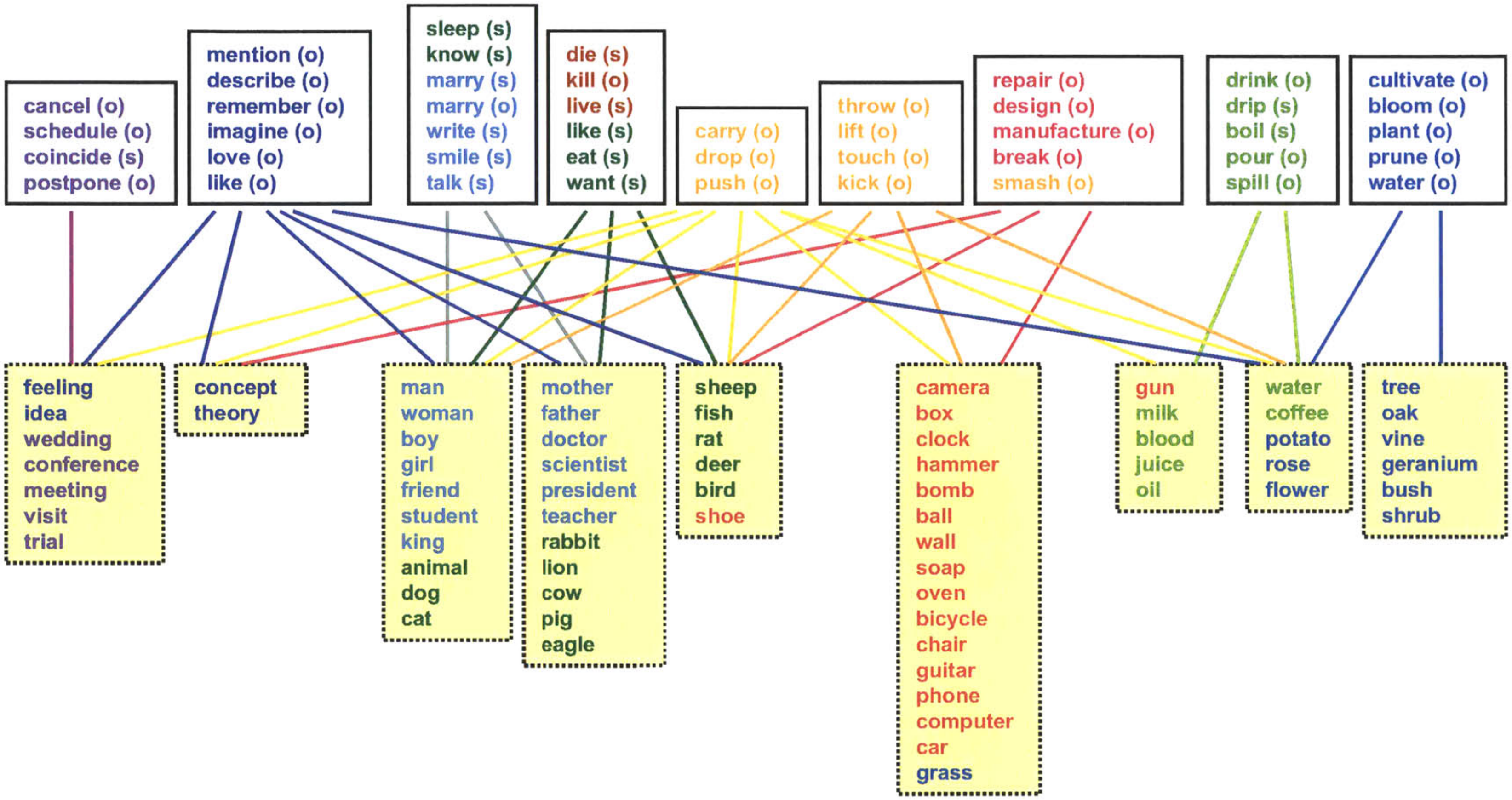


Figure 4-14: The structure recovered by the flat model, with an added physicality constraint, from the BNC data.

The model selection simulated data text showed how the child could learn the M constraint, but future work remains to be done to investigate the kind of data that children have about the world at different ages. The richness and diversity of the data (in terms of numbers of predicates, objects, and ontological categories that the child has knowledge of, and how complete their knowledge is) will determine whether the predication events they have seen are sufficient for children to learn the M constraint.

General discussion

I have shown how the M constraint can be used to learn about the world and make inferences about predicability based only on truth data. I have demonstrated the model's effectiveness on simulated and real-world data sets. I have also shown parallels between the model's behavior with increasing amounts of data and the developmental trajectory of children learning sense and nonsense. Additionally, I have demonstrated how the M constraint itself could potentially be learned. Here I discuss future work, and the broader implications of the model.

Model extensions

Relaxing simplifying assumptions

As discussed in the description of the M constraint model, many simplifying assumptions were made in designing the model. The model is therefore in some ways an inaccurate representation of how sense is related to real-world occurrences. Various extensions could potentially remedy this.

As one possible model extension, non-binary truth and predicability matrices could be explored further to account for the fact that human judgments of truth and sensibility are frequently graded rather than strict binary decisions. Sensibility could be graded in a predictable way. Drange's suggestions (1966) that some things are unthinkable (complete nonsense) while others have graded degrees of sensibleness could still be the consequence of a strict hierarchy; it could be that combining a predicate with an object that it nearly spans (i.e., that it would span if it were moved one node in the tree) makes a more sensible predication event than combining a predicate with an object that would only be spanned by that predicate if the predicate's location was altered more greatly. Thus, a strictly hierarchical predicability tree could correspond to graded sensibility judgments instead of binary ones. Other ways of implementing graded sensibility judgments could be explored as well.

Regardless of whether sensibility is graded, a graded version of the truth matrix makes more sense than a binary version of the matrix, philosophically; bananas might occasionally be painted blue, even if they don't naturally come in that color, for instance. Future work might try implementing a probabilistic matrix in place of the current binary truth matrix and see whether that improves the resulting predicability matrices recovered by the model.

It is also worth exploring what additional factors go into the probability of a predicate-object pair appearing in the data in addition to the popularities of the predicate and object, and the truth of the pair. Particularly in looking at language data, certain phrases occur frequently enough to be stored as lexical items themselves, and certain words hardly ever occur outside of the context of a particular phrase (Becker, 1975). These lexical factors could conceivably be incorporated into the model for superior learning of predicability from language data.

Handling exceptions to the M constraint

There are many possible exceptions to the M constraint that can arise, particularly in dealing with natural language data. I have discussed already the issue of polysemy, which may cause language data in particular to lead to objects or predicates having multiple locations within the tree (e.g., *bat* can be both an animal and an artifact), and also figurative speech, whereby we might talk about a *pushing a deadline back* or a *car dying*. Additionally, there are some non-linguistic challenges to the M constraint as well. (Carey, 1983) argued that real world data may contain many exceptions to the M constraint. Within the domain of animals (including humans), for instance, there may be things that are sensible to say of females but not males. It may be true or untrue of a female that she *is pregnant*, but Carey argues that that predicate doesn't make sense when applied to a male; similarly, it may not make sense to discuss whether a female *is impotent.* Therefore, there may be cross-cutting distinctions within some parts of the tree, such as taxonomic and gender distinctions within the ontological category of animals.

Carey also points out that some items are heterotypical, belonging to multiple ontological categories. For instance, *book* refers both to the content of a book and the physical object; *Italy* refers to a geographical location as well as a political and cultural entity. Different meanings are predicated very differently, but the meanings are all closely bound together. Finally, Carey points out an apparent M constraint violation involving spatial objects with no mass (e.g., *holes*, which can be *tall* or *wide* but not *heavy*) and objects that have mass but no set spatial configuration (e.g., *milk*, which can be *heavy* but not *tall* or *wide*).

What are we to do with all of these potential M constraint violations? Keil and Kelly (1986) argue that many of these potential issues are not truly M constraint violations. They argue, for instance, that a pregnant male is perfectly conceivable (if you'll pardon the pun), and that it is merely a contradiction due to the definition of *male* that makes *pregnant male* sound odd — not a category mistake leading to true nonsense. Keil and Kelly, however, admit that the real world may contain actual violations of the M constraint, including the overlapping predications involving spatial objects without mass, discussed above. They provide empirical evidence that the number of M constraint violations is small, however, even in the parts of the tree Carey is most concerned by, and they state that therefore the M constraint remains a powerful learning tool for making rapid predicability inferences, rather than one which would handicap a child learning about the world.

While Keil and Kelly convincingly argue that some of the potential M constraint violations raised by Carey are probably not true category mistakes, there remain multiple issues that the M constraint model must handle in learning from experiential and linguistic data. These issues include polysemic or heterotypical objects (particularly when using language data as input), and genuine exceptions to the M constraint like *holes*. One advantage of a probabilistic model is that it can tolerate noise. It may be possible that learners have a strong hierarchical bias in place of a strict constraint – one which allows them to make exceptions when there is sufficient evidence. The model presented in this chapter could be adapted to do the same, breaking the hierarchy in order to accommodate items like *holes* that have some properties of physical objects but lack others. Additionally, the M constraint model could consider multiple locations for a single object or predicate, given sufficient data, but the model could have a prior preference for fewer locations per lexical item. This would address the issues of homonymic or polysemic lexical items like *bat* and also heterotypic objects such as *books*.

What if there are cross-cutting distinctions that affect sensibility judgments in some ontological domains (such as reasoning based on taxonomy, gender, or ecology within the domain of living things)? Potentially, additional reasoning constraints outside the model could be responsible for judgments like *men cannot be pregnant*; this is the solution proposed by Keil and Kelly. However, it is also possible that the existing M constraint model could be modified to learn more than one structure to describe the same set of predicates and objects. Previous work suggests that multiple context-sensitive models, each representing a different hierarchical (or other) structure, can be learned to capture reasoning about different aspects of a complex domain (Shafto, Kemp, Mansinghka, Gordon, & Tenenbaum, 2006). Thus, the M constraint could potentially learn multiple cross-cutting structures instead of a single hierarchy, to govern sensibility judgments. The M constraint, as a probabilistic model, has a great deal of flexibility, allowing it to potentially accommodate exceptions in a number of ways.

Scaling the model

Currently, the process of learning a predicability tree using this model is computationally difficult, limiting the size of data sets that are practical to use as input. Relaxing any of the simplifications made by the model would only increase the computational complexity. However, the search problem could potentially be simplified by increasing the number of objects and predicates in the data set over time, and changing the learning approach to an incremental one. Children are obviously incorporating objects and predicates into their existing trees without changing the structure of their overall predicability trees much of the time. Periodically, as children learn that events are not physical objects, or that plants are alive, their trees undergo major revisions. However, when this occurs, many of the ontological categories remain the same, but merely move in the tree. The M constraint search problem could potentially be made more tractable by implementing a search strategy that learns about the world more incrementally and only considers certain types of major revisions to the tree, given sufficient evidence. This would also more closely approximate the problem that child learners face than the simulations in which the M constraint model received all the data at once; child learners start with sparse data about relatively few predicates and objects, and gain more information as they grow.

Applications of the M constraint model

Psychology

Keil has provided strong evidence that people do honor the M constraint in reasoning about the world, even from a very young age. However, no psychological evidence currently exists that indicates that the M constraint is learned — or that it is not. This chapter has provided evidence that it is possible to learn the M constraint given certain kinds of data sets. This work opens the door to investigations of whether the data that children encounter would support the learning of the M constraint, and whether there is any evidence of children considering non-hierarchical predicability structures at any age.

Returning to the topic of sentence processing, plausibility seems to be based on both what makes sense and what tends to happen. According to the M constraint model, however, sensibleness is distinct from truth. Examining whether there is any quantitatively larger jump in difficulty of processing sentences that are untrue and sentences that involve category mistakes may help clarify the relationship between plausibility and sensibleness. It is possible that plausibility judgments are predictable from a combination of the truth and predicability matrices inferred by the M constraint model.

Artificial intelligence

The M constraint is potentially a powerful tool for automatically learning about ontological structures and common sense — in terms of what is sensible as well as what is true — based on frequency data from the real world, including natural language data, which is now abundantly available on the World Wide Web as well as various corpora. Automatic parsers (e.g., Collins, 1999) readily yield verb-subject and verb-direct object pairs, as used in the BNC-based tests of the model.

Harvesting large amounts of natural language data may overcome many of the issues of sparseness that caused the model to fail to capture certain human intuitions given the BNC data. Additionally, more kinds of input can be used; in learning about the world from language data, we should seek to incorporate more than just verb-noun interactions. Adjective-noun co-occurrences are also predication events. Additionally, adverbs can yield predication information: *ideas* not only can't *sleep furiously*, they can't do anything furiously. Only creatures that have mental states can do so. Natural language data provides more information about the world than the limited set of predication types I examined in this chapter.

Beyond natural language data, a number of data sets have been collected from human participants in various studies addressing features that various objects possess (e.g., Ross & Murphy, 1999, Rosch & Mervis, 1975). Such object-feature matrices could additionally be used as input for the M constraint model. These data sources share many of the same limitations of natural language data, however; people are most likely to generate facts about distinct features, and not common ones like *eats food* or *can be thought about.* An incremental M constraint learner could also be used to add new information to already existing ontologies and commonsense reasoners like Cyc.

In general, the M constraint could potentially aid greatly in AI endeavors involving commonsense reasoning. However, if the model is to handle large amounts of data, scaling issues in the search process will become of paramount importance to address. And when dealing with natural language data, an automatic learner would certainly need to be able to handle polysemy; possible extensions in these directions are discussed above.

Conclusion

In this chapter, I have shown how a generative Bayesian model incorporating the M constraint can be used to learn what is sensible about the world given sparse observations of what is true in the world. Additionally, I have demonstrated how Bayesian model selection can be used to learn the M constraint given a hypothesis space including alternate models. If people do organize predicates and objects hierarchically, this result suggests that the hierarchical bias may be learned rather than innate, depending on the structure of the data in the world. In Chapter 5, I discuss the overall implications of this and the previous chapters for our understanding of meaning and meaningfulness.

Chapter 5

Discussion

This thesis has explored a number of different areas related to semantics, word learning, and word use. What are the overarching lessons we can learn from this work? How can these lessons be applied in the future? Here I discuss the conclusions we can draw about the way people learn and use words and phrases, and about the cognitive science dichotomies discussed in Chapter 1. I then discuss the applications of this work.

What have we learned about people?

In Chapter 2, we saw that people have powerful cognitive tools to help them generalize about word meaning based on limited evidence, even in a novel domain. People generalize to distinctive groups of objects. They restrict their generalizations according to the size principle as they gain more examples of how a word applies. They have a great deal of uncertainty about the structure of a domain – especially at the superordinate level of categorization – when they are first introduced to unfamiliar objects and categories. However, they still apply the same principles of word learning that people do in a familiar domain; the results are simply more uncertain when generalizing to superordinate level objects.

The exception to the idea that word learners behave the same with novel objects as with familiar objects is that the basic-level bias does not seem to exist, or exists only very weakly, when learning about a new domain. It appears that the preference for generalization to the basic level develops over time — I hypothesized that this may occur because basic level categories are so frequently picked out by language that they are made more salient due to language use. Chapter 2 showed an effect of language on thought; word learning in a novel domain changes how distinctive people find different categories of objects to be. This, in turn, shapes prior expectations about the categories of objects that are likely to share labels or properties.

In Chapter 3, we saw how people can understand what objects a phrase refers to by assembling the meanings of the words in the phrase. In particular, we looked at the puzzling case of adjectives whose extension varies depending on the context — for instance, something that is *a large Chihuahua* is not *a large dog.* People understand gradable adjectives as statistical functions that take in one category (specified by the noun phrase they modify, and other contextual factors) and probabilistically identify a subcategory of items that are *the large Xes* (e.g.). Gradable adjectives act as non-parametric functions, not assuming any underlying distribution of the category of items that they operate over. This allows them the flexibility to function just as well in phrases like *the tall things in this room* as in cases where the category can be assumed to be normally distributed. Understanding this class of words paves the way for investigations of how people use other modifiers across different contexts.

In Chapter 4, we saw that people can learn about which combinations of words are sensible, using a hierarchical constraint on objects and predicates called the M constraint. The M constraint allows people to infer what is sensible based on the very limited evidence of what actually occurs in the world (a strict subset of what is sensible). By applying a hierarchical constraint to what is sensible, and by reasoning about truth as a subset of sensibleness, people can learn about sense and nonsense from the evidence they see in the world around them. We saw that a prior preference for simpler tree structures predicts the initial overgeneralization and the following developmental trajectory that children show in judging what makes sense and what is nonsense. Additionally, we saw that the M constraint could itself be learned from among a set of possible models. Whether or not it is actually learned depends in part on the data that children receive about the world, and whether a hierarchical structure is strongly reflected in that data. However, even if the data could support the learning of the M constraint, children may or may not start out already equipped with this useful learning constraint.

What have we learned about logic and statistics?

The Bayesian paradigm allows us to combine logical constraints and structures with probability to predict the graded judgments that people will give about word meanings and phrase meanings. It also allows us to recover the most probable underlying structure that gave rise to a particular set of data. This combination is powerful, uniting the graded measurements of psychology with the formal representational structures postulated in linguistics, philosophy, and AI. It also allows us to empirically evaluate the power and predictiveness of logical constraints that people may be using in word learning and reasoning about meaning and sensibleness.

In Chapter 1, we saw that people can infer a set of categories and category structure based on the perceptual relationships between items — though there can be uncertainty in their structural representation. They can reason about the probabilities of various categories being the correct word meaning, incorporating both prior biases (the distinctiveness bias) and constraints (the taxonomic constraint) and the evidence that they see about how the word is used. Combining these factors, people show graded generalizations about word meaning, with growing certainty about the correct category the more examples they see.

In Chapter 2, we saw that gradable adjective usage can be seen as a matter of finding a subcategory of items within a category described by the noun phrase. By probabilistically identifying the subcategory, people show graded judgments about which items are *tall* within a given context. This fits with the general intuition that both nouns and noun phrases should refer to categories in the world, and allows us with a probabilistic framework for identifying those categories – and the degree of uncertainty about the categories – within a given context.

In Chapter 3, we saw that people can use graded, limited evidence about how often object-predicate pairs occur in the world to infer an underlying structure of categories and category relationships. The M constraint allows for this structural inference. But as shown by the probabilistic model, people's theories of the correct set of categories and the relationships between those categories can change over time as they gain new evidence about what occurs in the world.

What have we learned about prior constraints and learning from data?

The Bayesian paradigm also allows us to identify how prior beliefs and constraints combine with current evidence to yield a set of judgments about the world. "Nature vs. nurture" is overly simplistic; people start out with some assumptions about the world, and these change over time given the evidence — leaving them with new prior beliefs for a given context. The probabilistic modeling approach allows us to tell what can be learned from a given set of data, and how different sets of prior beliefs and constraints affect a learner's inferences.

In Chapter 1, we saw that a preference for distinctiveness is present even when learning about novel domains — but that the salience, or distinctiveness, of a given category can be changed based on new evidence about word meanings. We saw that a basic-level bias need not be built in initially to explain how people can learn words, but may develop over time. We saw that people assume that the evidence that they see in the world around them is sampled in a representational manner from an underlying category; this leads to the size principle in generalizing about the correct category for a new word to apply to.

In Chapter 2, we saw that people have prior expectations about the size of a subcategory like *the tall trees* within the larger category of *trees.* A language user using *tall* in a given context already has parameters governing the range of heights and possibly the number of items in the category picked out by *the tall trees.* However, the particular set of items in a given context will combine with these parameters to yield a set of category judgments specific to the context. These prior expectations about category size could be learned and fine-tuned over time; children may have different expectations than adults. Additionally, it is possible that the exact nature of the function itself has to be inferred by the child over time based on evidence of gradable adjective usage across different phrases.

In Chapter 3, we saw that having a prior expectation that the world is hierarchically structured allows people to draw rapid structural inferences from limited data. This allows people to learn about what makes sense and what doesn't from early on. Judgments about sensibleness change over time, however; if a child hears that a blicket can be fixed with a screwdriver, her generalizations about what else a blicket can sensibly do will be governed by her current predicability tree. We also saw that the M constraint itself does not have to be innate. Given data that sufficiently supports a tree structure, the M constraint itself could potentially be learned.

Applications

The work in this thesis provides a framework for investigating many questions about language meaning and use. In the area of word learning, we now have a framework for investigating how prior expectations about word meaning change over time, and more generally how word-learning shapes mental representations of objects and categories. Additionally, this approach can be used to investigate other possible constraints on word learning. For instance, within the word learning paradigm of Chapter 2, we could examine how people's guesses about word meanings vary based on the words they have already learned. Based on the extent to which people avoid generalizing multiple words to the same category, we can model the strength of people's preference to avoid synonymous labels (see Clark's principle of contrast (1987)). Additionally, teaching people about a new domain and modeling the development of a basic-level bias within that domain can teach us more about the developmental trajectory of this bias of category selection.

We now have a modeling approach that allows us to investigate how gradable adjectives are learned – Are they learned piecemeal and independently at first, and eventually generalized into meaningful word classes like "gradable adjectives"? Or does the learning happen in a more top-down manner – do children assume that a class of statistical functions exist that select subcategories along a dimension, and then learn the proper dimension for and parameters for each gradable adjective? Does learning the parameters governing one gradable adjective affect expectations about gradable adjectives as a whole? By modeling gradable adjective development in young children, and by teaching adults novel gradable adjectives and modeling their usage, we can learn more about how people learn and use gradable adjectives. This, in turn, can provide insights about modifiers and compositionality more broadly. In the area of semantic compositionality, there is a rich area to be explored in modeling modifiers other than gradable adjectives, such as intensifiers and quantifiers.

The model selection work presented in Chapter 4 gives us the tools to determine whether the M constraint could be learned from real-world data. Gathering a data set more representative of children's knowledge of what occurs in the world would allow us to answer with more certainty whether the M constraint could be learned or whether it must be innate. Additionally, such a data set would reveal the extent to which the M constraint is a reflection of real-world structure as opposed to a cognitive tool that simplifies the world to greatly improve our inference power.

The work done in this thesis provides a method for predicting which categories people may be referring to when using language, and the probability with which they are referring to specific categories. Additionally, it allows us to predict what humans will infer from different sets of evidence. This work could be applied toward making better language-related tools in a number of ways:

- In understanding phrases (or bag-of-words search requests) that can be interpreted in multiple ways, knowing which objects and predicates are sensible together and which are not can reduce ambiguity.

- In automatically learning about a new object or property in the world, the M constraint model can allow us to make rapid inferences and learn as much as possible from even a single piece of data.

- In building artificial agents that can interact with or interpret people's actions in a real-world context, there is often ambiguity about speech referents. Knowing how people use combinations of words like *the tall Xes* to pick out categories of items is essential in identifying groups of things that people are referring to. E.g., "Give me the bucket that's under the tall trees".
- In a similar situation, a person may instead identify a category by giving an example or two of the category members – e.g., "I want all the ones that are like those two". Knowing how people generalize to categories based on limited evidence also helps us understand what kind of category they are likely to be picking out by specifying a certain set of evidence.

Understanding how people use language and generalize about the meanings of words and phrases from limited evidence can help us better understand and predict humans. It can also allow us to build better tools to interact with humans.

Conclusion

Within this thesis, I have introduced a Bayesian approach to meaning and meaningfulness. This framework allows us to rigorously define components of word and phrase meaning – both the underlying categories and constraints, and the probabilistic reasoning that applies to those structures. This work allows us to better understand and predict how people learn, use, and understand language across different real-world contexts.

Appendix A: model details

Model details from Chapter 3

Threshold-based models

The mathematical descriptions of the threshold $T(C)$ for each of the threshold models is as follows:

- *Absolute height (AH)*: $T(C) = k$, with k a fixed parameter.
- *Absolute number (AN)*: Let $x_1, ..., x_N$ be an ordering of C by height (i.e. $h(x_i) \leq h(x_j)$ if $i < j$). Then: $T(C) = h(x_k)$.
- *Unique percent number (UAN)*: as in AN, but computed based on only one object of each height.
- *Percent number (PN)*: $T(C) = h(x_{\lfloor k \cdot N \rfloor})$.
- *Unique percent number (UPN)*: as in PN, but computed based on only one object of each height.
- *Relative height by range (RH-R)*: If we write $Mx = max_{x \in C} h(x)$ and $Mn = min_{x \in C} h(x)$, then: $T(C) = Mx - k \cdot (Mx - Mn)$.
- *Relative height by standard deviation (RH-SD)*: If we write $\bar{x}$ and σ for the mean and standard deviation of heights of object in C, then: $T(C) = \bar{x} + k \cdot \sigma$.

The CLUS model

The posterior probability of the assignment of objects to clusters, Q, given data in the context, C, could be calculated as follows:

$$
\begin{aligned}
P(Q|C) &\propto P(C|Q)P(Q) \\
&= P(Q)\left(\prod_{x\in C} P(x|\mu_{q_x},\sigma_{q_x})\right)\left(\prod_{q\in Q} P(\mu_q)P(\sigma_q)\right)
\end{aligned} \tag{1}
$$

For each item $x \in C$, μ_{q_x} and σ_{q_x} are the mean and variance of the cluster that produced x; this calculation multiplies the likelihood of each data point x given its cluster parameters by the prior probability of each of those cluster parameters given the model. However, the problem with calculating this directly is that there are an infinite number of underlying Gaussians that could have generated each of the clusters of data. In order to calculate $P(C|Q)$, I therefore instead integrate out the parameters μ and σ, based on the hyperparameters governing their conjugate distributions[1]:

$$
\begin{aligned}
\mu &\sim N(\mu_0, (\kappa_0\sigma)^{-1}) \\
\sigma &\sim \Gamma(\alpha_0, \text{rate} = \beta_0)
\end{aligned}
$$

I define a few more variables, based on the hyperparameters and also n, the number of items in the context, and $\bar{x}$, the mean of the data in the context:

$$
\begin{gathered}
\mu_n = \frac{\kappa_0\mu_0 + n\bar{x}}{\kappa_0 + n} \\
\kappa_n = \kappa_0 + n \\
\alpha_n = \alpha_0 + n/2 \\
\beta_n = \beta_0 + (1/2)\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{\kappa_0 n(\bar{x} - \mu_0)^2}{2(\kappa_0 + n)}
\end{gathered}
$$

Given these hyperparameters, we integrate out μ and σ, instead calculating the *marginal likelihood*:

[1] The equations for the marginal likelihood calculation and the hyperparameter definitions are taken in part from (K. Murphy, 2007).

$$P(C|Q) = P(C|\alpha_0, \beta_0, \kappa_0, \alpha_n, \beta_n, \kappa_n)$$
$$= \frac{\Gamma(\alpha_n)}{\Gamma(\alpha_0)} \frac{\beta_0^{\alpha_0}}{\beta_n^{\alpha_n}} (\frac{\kappa_0}{\kappa_n})^{1/2} (2\pi)^{-1/2} \qquad (2)$$

For the purposes of the model presented in this chapter, $\mu_0 = 0$ in all cases, and α_0, $beta_0$, and κ_0 are free parameters.

The prior probability of an assignment, $P(Q)$, is induced by the Chinese Restaurant Process (Aldous, 1985):

$$P(q_i = c|q_1, ..., q_{i-1}) = \begin{cases} \frac{n^c}{i-1+\gamma} & n^c > 0 \\ \frac{\gamma}{i-1+\gamma} & n^c = 0 \end{cases}$$

where q_i is the cluster assignment of item i in the context, c is a given cluster, and n^c is the number of items already assigned to that cluster. This process prefers to assign new items to larger clusters – thus, there is a rich-get-richer effect to the clustering. The extent of this preference is governed by the free parameter γ.

Model details from Chapter 4

The M constraint model

This section describes in more detail the factors that go into calculating the joint probability of a truth matrix T and predicability matrix R given the data, D:

$$P(T, R|D) \propto P(D|T, R)P(T|R)P(R) \qquad (3)$$

The predicability matrix R can be parameterized based on the skeletal structure of the corresponding tree (B) and the partitions over predicates (z_p) and objects (z_o) found in the tree. Formally, then, the prior probability of R is expressed as:

$$P(R) = P(z_o, B, z_p) = P(z_o|z_p)P(B|z_p)P(z_p) \qquad (4)$$

The model assumes that the partition over predicates is generated by a Chinese restaurant process with concentration parameter γ. [2] This prior assigns higher probability to partitions with fewer clusters. This contributes to the preference for smaller trees, as the number of predicate clusters is equal to the number of nodes within the tree.

B is the tree structure build from the clusters in z_p. Given that there are $|z_p|$ predicate clusters, our model has no prior preference for any particular tree structure with $|z_p|$ nodes. I assume that B is drawn uniformly from all such possible trees. Then the probability of a particular tree structure B given a particular partition z_p is inversely proportional to the number of trees with as many nodes as there are partitions:

$$P(B|z_p) = \frac{1}{|z_p|^{|z_p|-1}} \tag{5}$$

Lacking any prior expectations about the arrangements of the objects given number of nodes, z_o is generated by dropping the objects randomly onto the nodes of the tree. Each object thus has a $1/|z_p$— chance of landing in a particular node, and the probability of a partition z_o given the partition z_p is:

$$P(z_o|z_p) = \frac{1}{|z_p|^n} \tag{6}$$

where n is the number of objects.

The second of the three probability distributions that is required in order to apply Equation 3 is $P(T|R)$, the probability of the truth matrix given the predicability matrix. I assume that the truth value of each object-predicate pair which is predicable in R is determined by flipping a coin with bias η:

$$P(T|R) = \begin{cases} \eta^{|T|}(1-\eta)^{|R|-|T|}, & \text{if} T \subset R \\ 0, & \text{otherwise} \end{cases}$$

where $|R|$ is the number of predicable pairs and $|T|$ is the number of true pairs.

[2] This process was described previously in Appendix A in the section describing the CLUS model.

The third distribution to be defined is $P(D|T,R)$, the probability of the data matrix (consisting of observed object-predicate frequencies) given the truth and predicability matrices. I define v to be a single event observed in the data, that is, a single occurrence of an object-predicate pair (o,p). I assume that events are drawn from a distribution given by

$$P(o,p) \propto e^{\pi_p+\pi_o+\lambda T_{po}} \tag{7}$$

where π_o and π_p are parameters representing the popularity of the object and predicate, respectively. T_{po} is the truth value of the object-predicate pair, and λ represents the extent to which we penalize violations of the truth matrix. In the trivial cases where T is uniformly 1 (every object-predicate pair is both sensible and true) or λ is 0 (object-predicate pairs are not penalized at all for occurring when they are not true), then this distribution reduces to a simple joint distribution where the probability of an object-predicate pair occurring relies only on the popularities of o and p:

$$P(o,p) = P(o)P(p) = (e^{\pi_o})(e^{\pi_p}) \tag{8}$$

The popularities of o and p are taken here to be independent as a simplifying assumption; in actual language data, the relative popularities of phrases cannot always be explained simply by invoking the popularity of the object and predicate independently (for example, the popularity of object-predicate pair *big business* is influenced by the phrase having a particular meaning of its own beyond that of the words.

While the popularity parameters could potentially be learned, I fix them using the frequencies of the objects and predicates observed in the data:

$$\pi_o \propto log|\pi_o| \tag{9}$$

where π_o is the proportion of events in D that involve object o. π_p is defined similarly.

The events in the data set are independently generated, and the probability of the entire data set is taken to be the product of the probability of each individual event, v:

$$P(D|T) = \prod_{v \in D} P(v|T) \tag{10}$$

Together, these equations define the generative process by which an underlying predicability tree, truth matrix, and popularities for each object and predicate lead to a data set of observed object-predicate pairs.

Searching for the best predicability tree and truth matrix

Running an exhaustive search to find the best underlying parameters that maximize $P(R, T|D)$ is computationally infeasible for data sets of any substantial size. I therefore use a stochastic search process to identify the values of $R = (z_o, B, z_p)$, T, λ, and η that maximize the posterior probability of the truth and predicability matrices.

The search problem is difficult, since the organization of the predicability tree is such that changing the location of a predicate within the tree B is unlikely to improve the overall score unless at least some of the objects that that predicate spans also change positions within the tree (more formally, the score for a particular predicability matrix R couples the partitions z_o and z_p). This means that a search consisting of proposed search moves that change the positions of predicates and objects separately is unlikely to recover the best predicability and truth matrices. I overcome this issue by not initially assuming that the objects are placed at any particular node, and only placing the objects once I have found the best tree structure and predicate locations.

More formally, I integrate out the locations of the objects and searching for the B, z_p, and T that maximize

$$P(B, z_p|D) \propto P(D|B.z_p)P(B, z_p) \tag{11}$$

Suppose that D_j is the set of predication events in the data that involve object j (for instance, every property paired with the object *banana*, along with the frequency count of how many times each predicate-object pair has occurred). Then the probability of the overall data set D given the tree and the partition over predicates is the same as the product of D_j for each object j:

$$P(D|B, z_p) = \prod_j P(D_j|B, z_p) \tag{12}$$

Then

$$P(D_j|B, z_p) = \sum_{k=1}^{|z_p|} \sum_{t_j} P(D_j, t_j, z_o^j = k|B, z_p) \tag{13}$$

where z_o^j is the location of object j in the tree, and t_j is a vector of truth values indicating which predicates are true of object j. In other words, to integrate out the object locations exactly, it is necessary to sum over all possible locations of objects in the tree (the $|z_p|$ nodes of the tree) and also over all possible truth assignments for that particular object. Computing this sum exactly is intractable, so I approximate it as follows:

$$P(D_j|B, z_p) \approx \sum_{k=1}^{|z_p|} P(D_j|t_j^*(k))P(z_o^j = k|B, z_p) \tag{14}$$

where $t_j^*(k)$ is the truth vector that maximizes $P(t_j|z_o^j = k, B, z_p)$. In computing $t^*(z_o^j = k)$, I sort the predicates based on how many times each one appears with o_j in the data. Let p_i and p_{i+1} be a sequential pair of sorted predicates, where p_{i+1} has occurred with o_j more than p_i. I consider drawing a line between p_i and p_{i+1} and calling all the predicates on the p_i side of the line false when combined with o_j, and all predicates on the other side of the line true in combination with o_j. I compare all such possible cuts between pairs of sorted predicates, and choose the one that maximizes $P(t_j|z_o^j = k, B, z_p)$. Equation 6 implies that $P(z_o^j = k|B, z_p) = \frac{1}{|z_p|}$.

Conditioning on the number of times each object appears in the data set, it is straightforward to compute $P(D_j|t_j^*(z_o^j = k))$: in particular, this method avoids the necessity of computing the normalizing constant of the distribution in Equation 7. Let Z be the normalizing constant for Equation 7:

$$Z = \sum_{o,p} P(o,p) = \sum_{o,p} e^{\pi_p + \pi_o + \lambda T_{po}} \tag{15}$$

I avoid calculating this constant explicitly as follows. Let C be the number of times that object o_j appears in the data set. For o_j, I define

$$h(p_i) = P(o_j, p_i) \tag{16}$$

and

$$C = \sum_i c_i \tag{17}$$

where c_i is the number of times that o_j is paired with p_i in the corpus. Then

$$P(D_j | t^*(z_o^j = k)) = \frac{\prod_i h(p_i)^{c_i}}{(\sum_i h(p_i))^C} \tag{18}$$

The numerator of this term is proportional to the probability of the all the actual C predication events that involved o_j within the corpus. The denominator is proportional to the combined probability of all the possible combinations of predication events in which o_j could have occurred a total of C times. For instance, if the corpus included *green banana* once and *yellow banana* twice — and banana never occurred with the other predicates *blue* or *red* —- then the object *banana* occurred a total of 3 times. The above equation establishes the probability that it occurred in those particular 3 cases, as opposed to all other possible sets of 3 predication events (including $\{$*red banana*, *blue banana*, *yellow banana*$\}$ and $\{$*blue banana*, *blue banana*, *blue banana*$\}$). Therefore, it is possible to examine the probability of a particular set of data for a given object, and normalize on a per object basis.

Having now discussed how to calculate the best object locations, I move on to the search process. (B, z_p) is an incomplete tree: that is, a predicability tree without the objects attached. Using Equation 11, I run a search by starting from a random incomplete tree and considering proposals that move the predicates around the tree, possibly creating new nodes in the process. For each incomplete tree, I use an approximation similar to the idea behind Equation 18 to compute a candidate pair (T, R), where T and R are the matrices that maximize $P(T, R|B, z_p, D)$. At the end of the search, I return the best candidate pair encountered, where each pair is scored according to Equation 3.

The modified hierarchical clustering model

To score a predicate placement at a particular location within the hierarchical clustering tree, I compared the predicted data that we would expect given that predicate placement with the actual data seen in the data set. Data were predicted as follows: any object not spanned by the predicate when placed at that location was not predicted to occur with that predicate at all. Any object that was spanned by the predicate was predicted to occur in proportion to the product of the popularity of the predicate and the popularity of the object (each of which was proportional to the frequency with which the item in the data set). In other words, the predicted data vector q_i for predicate p_i was as follows:

$$q_{ij} \propto \begin{cases} \pi_{o_j} & \text{if } (p_i, o_j) \text{ is predicable} \\ 0 & \text{otherwise} \end{cases}$$

This predicted data vector for the predicate given its location was then compared to the actual data vector d'_i (based on a row rather than a column of the data matrix, as opposed to the object data vector discussed above). The location for each predicate was chosen to maximize the inner product of the normalized vectors $\frac{d'_i}{|d'_i|}$ and $\frac{q_i}{|q_i|}$.

The flat model

The flat model is mathematically identical to the M constraint model, except that, because it is not limited to predicability matrices that correspond to tree structures, there are many more possible predicability matrices. Therefore, the prior over predicability matrix, R, is different.

I parameterize each predicability matrix as a triple (z_o, C, z_p), where, as before, z_o and z_p are partitions over objects and predicates respectively. Here, C is the graph structure built from these clusters, governing which predicate clusters span which object clusters.

The M constraint model *a priori* preferred simpler tree structures with fewer nodes (corresponding to the number of predicate clusters). Here I implement an identical preference for fewer predicate clusters. Before, the tree structure was drawn uniformly from amongst all possible trees that could be built using $|z_p|$ nodes. Here, again, there is no prior preference for a particular structure and therefore again the structure is drawn uniformly. However, the number of structures that can be built from $|z_p|$ predicate nodes has greatly increased now that there is no constraint that the graph must be a strict hierarchy.

C is a binary matrix of size $|z_p|x|z_p|$, with each row representing a predicate cluster, and each column a binary vector corresponding to a possible behavior of an object — e.g., a column [1 1 0 1 0] in a $5x5$ matrix C would correspond to the behavior of objects that were spanned by the predicates in clusters 1, 2, and 4. The tree structure B in the M constraint model can also be described using such a binary matrix, with the columns of the matrix constrained to be tree-compatible object behaviors. In describing a tree B as such a matrix, then, the column [1 1 1 0 0] would not be allowed if the column [0 1 1 1 0] was also in the matrix, as this would create a non-hierarchical structure. Matrix C lacks any similar constraint, and all binary matrices of size $|z_p|x|z_p|$ are equally probable *a priori*. z_o is a random assignment of objects to rows of the matrix. This is equivalent to placing each object at a random node in a tree B previously; there, the object was spanned by all predicates in nodes above that node in the tree. Here, the object is spanned by all predicates in the graph nodes specified by the assigned column of the matrix C. Given these assignments, then

$$P(R) = P(z_o, C, z_p) = P(z_o|z_p)P(C|z_p)P(z_p) \tag{19}$$

While $P(z_p)$ and $P(z_o|z_p)$ are computed as before, the other component of the prior now differs:

$$P(C|z_p) = \frac{1}{2^{|z_p|}} \tag{20}$$

Each triple (z_o, C, z_p) uniquely determines a predicability matrix R where R_{ij} takes the same value as the entry in C corresponding to predicate i and object j (the row and column in C corresponding to the assignments of z_p^i and z_o^j respectively). The resulting hypothesis space of matrices R contains the subset of predicability matrices that do follow the M constraint, but also contains many more matrices that do not. This model captures the idea that predicates and objects cluster, and that predicates span objects, but contains no other structural biases or constraints.

The flat model represents the truth matrix T as a subset of R in exactly the same way as the M constraint model, and the probability of a data set D given T and R is calculated as before:

$$P(T, R|D) \propto P(D|T, R)P(T|R)P(R) \tag{21}$$

Appendix B: stimuli details

In this appendix I describe in more detail the distributions used in Chapter 3. For each of the distributions the object values (heights or lengths) are listed or described. A list of the experiments that included the distribution as a stimulus is also shown. (Note: the distribution labels were not used in the text of the thesis.)

label	values	experiments
Uniform	1 2 3 4 5 6 7 8 9	1a, 1b, 1c, 3
Exponential	0.33 0.33 0.6 1 1.6 2.6 3.5 5.5 9	1a, 1c
AddShort	0.5 1 1 1.5 2 2 3 4 5 6 7 8 9	1a, 1c
AddTall	1 2 3 4 5 6 7 8 8 8.5 9 9 9.5	1a, 1c
Clusters1	1 1 2 6 6.5 7 7.5 8 8.5 9	1a, 1c
Clusters2	1 1 2 6 6.5 7 7 7 7 7 7.5 8 8.5 9	1a, 1c
Clusters3	1 1 1 4 4 4 7 7 7	1a, 1c
PositiveSkew	1 2 3 4 2x{5}, 3x{6}, 4x{7}, 6x{8}, 8x{9}	1a, 1c
UniformSample1	20 items drawn from a uniform distribution ($\mu = 4, \sigma = 1$)	1a
UniformSample2	20 items drawn from a uniform distribution ($\mu = 4, \sigma = 1$)	1a
GaussianSample1	20 items drawn from a Gaussian distribution ($\mu = 4, \sigma = 1$)	1a, 3
GaussianSample2	20 items drawn from a Gaussian distribution ($\mu = 4, \sigma = 1$)	1a
TwoGaussE	40 items drawn from two Gaussians ($\mu = 0.95, 1.1; \sigma = 0.04, 0.06$)	1a
TwoGaussH	40 items drawn from two Gaussians ($\mu = 0.79, 1.32; \sigma = 0.13, 0.19$)	1a, 3
LVHHV	40 items drawn from two Gaussians ($\mu = 10, 120; \sigma = 1, 35$)	1a, 3
ExpInv	1 4.5 6.5 7.4 8.4 9 9.4 9.67 9.67	1a

label	values	experiments
TwoGaussAB	48 items from two pseudo-Gaussians (num = 24, 24; $\mu = 6.5, 16.5; \sigma = 1.1, 1.1$)	1a
TwoGaussGH	48 items from two pseudo-Gaussians (num = 12, 36; $\mu = 6.5, 16.5; \sigma = 1, 1.1$)	1a
TwoGaussIJ	48 items from two pseudo-Gaussians (num = 36, 12; $\mu = 6.5, 16.5; \sigma = 1.1, 1$)	1a
TwoGaussKL	48 items from two pseudo-Gaussians (num = 24, 24; $\mu = 8.5, 20.5; \sigma = 2.14, 1.1$)	1a
Short1Tall1	54 items drawn from 2 Gaussians (num = 36, 18; $\mu = 30, 73; \sigma = 7.7, 7.2$)	2
Short1Tall2	54 items from 3 Gauss. (num = 36, 9, 9; $\mu = 30, 66, 80; \sigma = 7.7, 0.5, 0.5$)	2, 3
Short2Tall1	54 items from 3 Gauss. (num = 18, 18, 18; $\mu = 22.4, 37.6, 73; \sigma = 0.5, 0.5, 7.2$)	2
Short2Tall2	54 items from 4 Gauss. (num = 18, 18, 9, 9; $\mu = 22.4, 37.6, 66, 80; \sigma = 0.5, 0.5, 0.5, 0.5$)	2
Short3Tall1	54 items from 4 Gauss. (num = 12x3, 18; $\mu = 20.7, 30, 39, 73; \sigma =$0.62x3, 7.2)	2
Short3Tall2	54 from 5 Gauss. (num = 12x3, 9x2; $\mu = 20.7, 30, 39, 66, 80; \sigma =$0.62x3, 0.5x2)	2

References

Aldous, D. (1985). Exchangeability and related topics. In *Ecole d'ete de probabilities de saint-flour xiii-1983* (pp. 1–198). Springer.

Anderson, J. (1991). The adaptive nature of human categorization. *Psychology Review*, *98*(3), 409–429.

Aristotle. (1963). *Categories.* London: Oxford University Press. (translated with notes by J.L. Ackrill)

Atran, S. (1998). Folkbiology and the anthropology of science: Cognitive universals and cultural particulars. *Behavioral and Brain Sciences*, *21*, 547–609.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet project. *Proceedings of the COLING-ACL.*

Barner, D., & Snedeker, J. (2008). Compositionality and statistics in adjective acquisition: 4-year-olds interpret tall and short based on the size distributions of novel noun referents. *Child Development*, *79*(3), 594–608.

Becker, J. (1975). The phrasal lexicon. In B. Nash-Webber & R. Schank (Eds.), *Theoretical issues in natural language processing 1.* Cambridge, MA: Bolt, Beranek, and Newman.

Berwick, R. (1985). *The acquisition of syntactic knowledge.* Cambridge, MA: MIT Press.

Berwick, R. (1986). Learning from positive-only examples: The subset principle and three case studies. *Machine Learning*, *2*, 625–645.

Bierwisch, M. (1989). The semantics of gradation. In M. Bierwisch & E. Lang (Eds.), *Dimensional adjectives* (pp. 71–261). Berlin, Germany: Springer-Verlag.

Bloom, P., & Keil, F. C. (2001). Thinking through language. *Mind and Language*, *16*(4), 351–367.

Boroditsky, L., Schmidt, L. A., & Phillips, W. (2003). Sex, syntax, and semantics. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: advances in the study of language and thought.* Cambridge, MA: MIT Press.

Bowerman, M., & Choi, S. (2003). Space under construction: Language-specific spatial categorization in first language acquisition. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: advances in the study of language and thought.* Cambridge, MA: MIT Press.

The British National Corpus, version 3 (BNC XML edition). (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/.

Brown, P. (2001). Learning to talk about motion UP and DOWN in Tzeltal: is there a language-specific bias for verb learning? In M. Bowerman & S. C. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 512–543). Cambridge, England: Cambridge University Press.

Buitelaar, P., Cimiano, P., & Magnini, B. (Eds.). (2005). *Ontology learning from text: Methods, evaluation and applications.* (Volume 123, Frontiers in Artificial Intelligence and Applications)

Carey, S. (1983). Constraints on the meanings of natural kind terms. In T. Seller & W. Wannenmacher (Eds.), *Concept development and the development of word meaning.* Berlin: Springer-Verlag.

Carey, S. (1985). *Conceptual change in childhood.* Bradford Books, MIT Press.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development*, *15*, 17–29.

Chierchia, G., & McConnell-Ginet, S. (1990). *Meaning and grammar: an introduction to semantics.* Cambridge, MA: MIT Press.

Chklovski, T., & Pantel, P. (2004). VerbOcean: Mining the Web for fine-grained semantic verb relations. *Proceedings of EMNLP-2004*.

Choi, S., McDonough, L., Bowerman, M., & Mandler, J. (1999). Early sensitivity to language-specific spatial categories in English and Korean. *Cognitive Develop-*

ment, *14*, 241–268.

Chouinard, M., & Clark, E. (2003). Adult reformulations of child errors as negative evidence. *Journal of Child Language*, *30*, 637-669.

Clark, E. (1987). The principle of contrast: A constraint on acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 1–33). Hillsdale, NJ: Erlbaum.

Collins, M. (1999). *Head-driven statistical models for natural language parsing.* Unpublished doctoral dissertation, University of Pennsylvania.

Colunga, E., & Smith, L. (2005). From the lexicon to expectations about kinds: A role for associative learning. *Psychological Review*, *112*(2), 347–382.

Cresswell, M. J. (1976). The semantics of degree. In B. H. Partee (Ed.), *Montague grammar* (pp. 261–292). New York, NY: Academic Press.

Dilkina, K., McClelland, J., & Boroditsky, L. (2007). How language affects thought in a connectionist model. *Proceedings of the 29th Annual Cognitive Science Society.*

Drange, T. (1966). *Type crossings.* The Hague: Mouton.

Duda, R., & Hart, P. (1973). *Pattern recognition and scene analysis.* New York, NY: Wiley-Interscience.

Ebeling, K. S., & Gelman, S. A. (1988). Coordination of size standards by young children. *Child Development*, *59*(4), 888–896.

Ebeling, K. S., & Gelman, S. A. (1994). Children's use of context in interpreting "big" and "little". *Child Development*, *65*(4), 1178–1192.

Elman, J. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Everett, D. (2005). Cultural constraints on grammar and cognition in Pirahã: Another look at the design features of human language. *Current Anthropology*, *46*(4), 621-646.

Fellbaum, C. (Ed.). (1998). *WordNet: an electronic lexical database.* MIT Press.

Frank, M. C., Everett, D. L., Federenko, E., & Gibson, E. (2008). Number as a cognitive technology: Evidence from pirahã language and cognition. *Cognition*, *108*, 819–824.

Frank, M. C., Goodman, N. D., Lai, P., & Tenenbaum, J. B. (2009). Informative communication in word production and word learning. *Proceedings of the 31st Annual Conference of the Cognitive Science Society.*

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (in press). Using speakers referential intentions to model early cross-situational word learning. *Psychological Science.*

Gelman, S. A., & Ebeling, K. S. (1989). Children's use of nonegocentric standards in judgments of functional size. *Child Development*, *60*(4), 920–932.

Gentner, D., & Boroditsky, L. (2001). Individuation, relational relativity and early word learning. In M. Bowerman & S. Levinson (Eds.), *Language acquisition and conceptual development.* Cambridge, England: Cambridge University Press.

Gentner, D., & Goldin-Meadow, S. (Eds.). (2003). *Language in mind: advances in the study of language and thought.* Cambridge, MA: MIT Press.

Gibson, E., & Pearlmutter, N. (1998). Constraints on sentence comprehension. *Trends in cognitive sciences*, *2*(7), 262–268.

Gold, K. (under review). Learning continuous-valued, compositional meanings for adjectives from sensor data. *IEEE Transactions on Autonomous Mental Development*, *1*(1).

Gordon, P. (2004). Numerical cognition without words: evidence from Amazonia. *Science*, *306*, 496–499.

Griffiths, T., & Tenenbaum, J. (2007). From mere coincidences to meaningful discoveries. *Cognition*, *103*(2), 180–226.

Halberda, J., Taing, L., & Lidz, J. (2008). The development of "most" comprehension and its potential dependence on counting ability in preschoolers. *Learning and Development*, *4*(2), 99–121.

Hare, R. M. (1952). *The language of morals.* Oxford, England: Oxford University Press.

Horn, L. (1989). *A natural history of negation.* Chicago, IL: Chicago University Press.

Huttenlocher, J., & Hedges, L. V. (1994). Combining graded categories: Membership and typicality. *Psychological Review*, *101*(1), 157–165.

Imai, M., Gentner, D., & Uchida, N. (1994). Children's theories of word meaning: The role of shape similarity in early acquisition. *Cognitive Development*, *9*, 45–75.

Inagaki, K., & Hatano, G. (2003). Conceptual and linguistic factors in inductive projection: how do young children recognize commonalities between animals and plants? In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: advances in the study of language and thought.* Cambridge, MA: MIT Press.

Justeson, J., & Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, *1*, 9–27.

Kamp, H. (1975). Two theories of adjectives. In E. Keenan (Ed.), *Formal semantics of natural language* (pp. 123–155). Cambridge: Cambridge University Press.

Kamp, H., & Partee, B. (1995). Prototype theory and compositionality. *Cognition*, *57*, 129–191.

Kay, P., & Kempton, W. (1984). What is the Sapir-Whorf hypothesis? *American Anthropologist*, *86*(1), 65–79.

Kay, P., & Regier, T. (2007). Color naming universals: The case of Berinmo. *Cognition*, *102*, 289–298.

Keil, F. (1979). *Semantic and conceptual development: an ontological perspective.* Cambridge, MA: Harvard University Press.

Keil, F. (1983). Semantic inferences and the acquisition of word meaning. In T. Seiler & W. Wannenmacher (Eds.), *Concept development and the development of word meaning.* Berlin: Springer-Verlag.

Keil, F., & Carroll, J. (1980). The child's acquisition of "tall": Implications for an alternative view of semantic development. *Papers and Reports on Child Language Development*, *19*, 21–28.

Keil, F., & Kelly, M. (1985). The more things change . . . Metamorphoses and conceptual structure. *Cognitive Science*, *9*, 403–416.

Keil, F., & Kelly, M. (1986). Theories of constraints and contstraints on theories. In

W. Demopoulous & A. Marras (Eds.), *Language learning and concept acquisition* (pp. 173–183). Norwood, NJ: Ablex.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2004). Learning domain structures. *Proceedings of the 26th Annual Conference of the Cognitive Science Society.*

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Kennedy, C. (1999). *Projecting the adjective: The syntax and semantics of gradability and comparison.* New York, NY: Garland.

Kennedy, C. (2002). *The landscape of vagueness.* Manuscript presented at Philosophy and Linguistics Workshop—University of Michigan.

Kennedy, C. (2007). Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, *30*(1), 1–45.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. *Proceedings of EURALEX.*

Kipper, K., Dang, H. T., & Palmer, M. (2000). Class-based construction of a verb lexicon. *AAAI-2000 Seventeenth National Conference on Artificial Intelligence.*

Klein, E. (1980). A semantics for positive and comparative adjectives. *Linguistics and Philosophy*, *4*(1), 1–45.

Klein, E. (1991). Comparatives. In A. von Stechow & D. Wunderlich (Eds.), *Semantics: An international handbook of contemporary research* (pp. 673–691). Berlin, Germany: Walter de Gruyter.

Kutas, M., & Hillyard, S. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205.

Landau, B., Smith, L., & Jones, S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*, 299–321.

Lapata, M., McDonald, S., & Keller, F. (1999). Determinants of adjective-noun plausibility. *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, 30–36.

Lenat, D. B. (1995). Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, *38*(11).

León, L. de. (2001). Finding the richest path: language and cognition in the acquisition of verticality in Tzotzil (Mayan). In M. Bowerman & S. C. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 544–565). Cambridge, England: Cambridge University Press.

Li, P., & MacWhinney, B. (1996). Cryptotype, overgeneralization and competition: A connectionist model of the learning of English reversive prefixes. *Connection Science*, *8*(1), 3–30.

MacWhinney, B. (1989). Competition and lexical categorization. In R. Corrigan, F. Eckman, & M. Noonan (Eds.), *Linguistic categorization* (pp. 195–242). New York, NY: Benjamins.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press.

Markman, E. (1989). *Categorization and naming in children.* Cambridge, MA: MIT Press.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* Henry Holt & Company.

Merriman, W. E. (1999). Competition, attention, and young children's lexical processing. In B. MacWhinney (Ed.), *Emergence of language.* Hillsdale, NJ: Lawrence Earlbaum Associates.

Montague, R. (1970). English as a formal language. In B. Visentini et al. (Ed.), *Linguaggi nella societa e nella tecnica.* Milan: Edizioni di Comunita.

Murphy, G. L., & Smith, E. E. (1982). Basic-level superiority in picture categorization. *Journal of Verbal Learning and Verbal Behavior*, *21*, 1–20.

Murphy, K. (2007). *Conjugate Bayesian analysis of the univariate Gaussian: a tutorial.* (Unpublished technical note; retrieved from http://www.cs.ubc.ca/ murphyk/Papers/bayesGauss.pdf)

Opfer, J. E., & Siegler, R. S. (2004). Revisiting preschoolers' living things concept: A microgenetic analysis of conceptual change in basic biology. *Cognitive Psychology*, *49*(4), 301–332.

Osherson, D., & Smith, E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition*, *9*, 35–58.

Parsons, T. (1970). Some problems concerning the logic of grammatical modifiers. *Synthese*, *21*, 320–334.

Partee, B. (1995). Lexical semantica and compositionality. In D. Osherson (General Ed.), L. Gleitman, & M. Liberman (Eds.), *Invitation to cognitive science. Part I: Language.* MIT Press.

Pearlmutter, N., & MacDonald, M. (1992). Plausibility and syntactic ambiguity resolution. *Proceedings of the 14th Conference of the Cognitive Science Society*, 498-503.

Piantadosi, S. T., Goodman, N. D., Ellis, B. A., & Tenenbaum, J. B. (2008). A Bayesian model of the acquisition of compositional semantics. *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society.*

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an Amazonian indigene group. *Science*, *306*, 499–503.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure.* Cambridge, MA: MIT Press.

Pinker, S. (1994). *The language instinct.* New York, NY: William Morrow and Company.

Quine, W. (1960). *Word and object.* Cambridge, MA: MIT Press.

Raffman, D. (2000). Is perceptual indiscriminability nontransitive? *Philosophical Topics*, *28*(1), 153–175.

Rasmussen, C. E. (2000). The infinite gaussian mixture model. *Advances in Neural Information Processing Systems*, *12*, 554–560.

Regier, T. (1996). *The human semantic potential: Spatial language and constrained connectionism.* Cambridge, MA: MIT Press.

Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive Science*, *29*, 819–865.

Resnick, P. (1993). *Selection and information: A class-based approach to lexical relationships* (Tech. Rep.). University of Pennsylvania Institute for Research in Cognitive Science.

Roberson, D., Davies, I., & Davidoff, J. (2000). Color categories are not universal: replications and new evidence from a stone-age culture. *Journal of experimental psychology: general*, *129*(3), 369–398.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. Lloyd (Eds.), *Cognition and categorization.* Laurence Erlbaum Associates.

Rosch, E., & Mervis, C. (1975). Family resemblance: studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573–605.

Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.

Ross, B. H., & Murphy, G. L. (1999). Food for thought: cross-classification and category organization in a complex real-world domain. *Cognitive Psychology*, *38*, 495–553.

Roy, D., & Pentland, A. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, *26*(1), 113–146.

Roy, D. M., Kemp, C., Mansinghka, V. K., & Tenenbaum, J. B. (2007). Learning annotated hierarchies from relational data. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems 19* (pp. 1185–1192). Cambridge, MA: MIT Press.

Samuelson, L. (2002). Statistical regularities in vocabulary guide language acquisition in connectionist models and 15-20 month olds. *Developmental Psychology*, *38*, 1016–1037.

Sera, M. D., Troyer, D., & Smith, L. B. (1988). What do two-year-olds know about the sizes of things? *Child Development*, *59*(6), 1489–1496.

Shafto, P., Kemp, C., Mansinghka, V., Gordon, M., & Tenenbaum, J. (2006). Learning cross-cutting systems of categories. *Proceedings of the 28th Annual Conference of the Cognitive Science Society.*

Shtulman, A., & Carey, S. (2007). Impossible or improbable? how children reason about the possibility of extraordinary events. *Child Development*, *78*, 1015–1032.

Simon, H. A. (1969). *The sciences of the artificial.* Cambridge, MA: MIT Press.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*, 39–91.

Smith, E., Osherson, D., Rips, L., & Keane, M. (1988). Combining prototypes: a selective modification model. *Cognitive Science*, *12*(4), 485–527.

Smith, L. (2005). Shape: A developmental product. In L. Carlson & E. VanderZee (Eds.), *Functional features in language and space* (pp. 235–255). Oxford University Press.

Snow, R., Jurafsky, D., & Ng, A. Y. (2006). Semantic taxonomy induction from heterogenous evidence. *Proceedings of COLING/ACL 2006*, 801–808.

Soja, N., Carey, S., & Spelke, E. (1991). Ontological categories guide young children's inductions of word meaning: object terms and substance terms. *Cognition*, *38*, 179–211.

Sommers, F. (1959). The ordinary language tree. *Mind*, *68*, 160–185.

Sommers, F. (1963). Types and ontology. *Philosophical Review*, *72*, 327–363.

Sommers, F. (1965). Predicability. In M. Black (Ed.), *Philosophy in America.* Ithaca, NY: Cornell University Press.

Sommers, F. (1971). Structural ontology. *Philosophia*, *1*, 21–42.

Stork, D. G. (1999). The Open Mind Initiative. *IEEE Expert Systems and Their Applications, May/June 1999*, 16–20.

Syrett, K., Bradley, E., Lidz, J., & Kennedy, C. (2006). Shifting standards: Childrens understanding of gradable adjectives. In K. U. Deen, J. Nomura, B. Schulz, & B. D. Schwartz (Eds.), *Proceedings of the inaugural GALANA.* Cambridge, MA: MITWPL.

Teh, Y. W., Daumé, H., III, & Roy, D. (2008). Bayesian agglomerative clustering with coalescents. *Advances in Neural Information Processing Systems*, *20*.

Whorf, B. L. (1956). *Language, thought, and reality* (J. B. Carroll, Ed.). London, England: Chapman and Hall.

Winawer, J., Witthoft, N., Frank, M., Wu, L., Wade, A., & Boroditsky, L. (2007). The Russian Blues reveal effects of language on color discrimination. *Proceedings of the national academy of science*, *104*(19), 7780–7785.

Xu, F., & Tenenbaum, J. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*, 245–272.

Yu, C., & Ballard, D. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*, 2149–2165.

Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*, 414–420.

Zadeh, L. A. (1965). Fuzzy sets. In *Information and control, vol. 8* (p. 338-353).

Zadeh, L. A. (1975). Calculus of fuzzy restrictions. In L. A. Zadeh, K. Fu, K. Tanada, & M. Shimura (Eds.), *Fuzzy sets and their applications to cognitive and decision processes.* Academic Press.

Zettlemoyer, L. S., & Collins, M. (2009). Learning context-dependent mappings from sentences to logical form. *Proceedings of the Joint Conference of the ACL-IJCNLP.*