Learning Nouns and Adjectives: A Connectionist Account

Michael Gasser*
Computer Science and Linguistics Departments
Lindley Hall 215
Indiana University
Bloomington, IN 47405, USA
gasser@indiana.edu

Linda B. Smith
Psychology Department
Psychology 332
Indiana University
Bloomington, IN 47405, USA
smith4@indiana.edu

Abstract

Why do children learn nouns such as *cup* faster than dimensional adjectives such as *big*? Most explanations of this phenomenon rely on prior knowledge of the noun-adjective distinction or on the logical priority of nouns as the arguments of predicates. In this paper we examine an alternative account, one which relies instead on properties of the semantic categories to be learned and of the word learning task itself. We isolate four such properties: the relative size, the relative compactness, and the degree of overlap of the regions in representational space associated with the categories and the presence or absence of lexical dimensions (*what color?*) in the linguistic context of a word. In a set of five experiments, we trained a simple connectionist network to label input objects in particular linguistic contexts. The network learned categories resembling nouns with respect to the four properties faster than it learned categories resembling adjectives.

Young children learn nouns more rapidly and less errorfully than they learn adjectives. The nouns that children so readily learn typically label concrete things such as BLOCK[1] and DOG. The adjectives that young children learn with greater difficulty label the perceptible properties of these same objects, for example, RED and WET. Why are concrete nouns easier for young children to learn than dimensional adjectives?

---

*To whom correspondence should be addressed

[1]We will use uppercase for concepts, italics for linguistic forms, and double quotes for utterances.

It is common in the study of cognitive development to explain such differences in learning by positing domain-specific mechanisms dedicated to that learning. Thus one might explain the noun advantage by looking for conceptual structures that specifically constrain or promote the learning of nouns and the lack of such specific structures for adjectives. In this paper, we pursue an alternate idea. We propose that common nouns and dimensional adjectives are initially acquired by the very same processes in the very same way. But, we argue, many mundane factors conspire to make names for common things more easily learned than labels for the properties of those things. We test our account by examining how a general category learning device, a multi-layer feedforward connectionist network, learns concrete nouns and dimensional adjectives.

1 The Phenomenon

Three kinds of evidence point to the initial priority of names for things over labels for the attributes of those same things. The first concerns the kinds of words that comprise early productive vocabularies. Nouns dominate; dimensional adjectives are rare or non-existent. For example, in Stern's diary study of the acquisition of English (Gentner, 1978), 78% of the words produced at 20 months were nouns while none were adjectives. Similarly, in Nelson's (1973) study of 18 children learning English, fewer than 7% of the first 50 words were adjectives. The priority of nouns over adjectives in early vocabularies is evident in other languages as well. In Dromi's (1987) study of one child learning Hebrew, only 4 of the first 337 words were adjectives. In a longitudinal study of the acquisition of Spanish by 328 children, Jackson-Maldonado et al. (1993) found only one adjective among the 88 most common words. The finding that adjectives are infrequent in early vocabularies is remarkable given that common dimensional adjectives such as size and color terms are among the most frequently used words in adult language.

The second class of evidence concerns studies of artificial word learning. In this commonly used method, experimenters present a novel object to a child and label it with a novel word (e.g., "this is a dax"). Children's interpretation of the word is measured by the kinds of other objects to which they generalize the newly learned label. Considerable evidence indicates that by 18 months (and quite possibly before), children interpret novel nouns as referring to taxonomic categories (Markman, 1989; Waxman, 1994). Further, the evidence suggests that children remember what they have learned over several days and weeks (Woodward, Markman, & Fitzsimmons, 1994). There have been a number of attempts to use these methods to teach novel adjectives. In these studies, the novel word is placed in an adjectival context (e.g., "this is a daxy one") or is explicitly contrasted with a known adjective (e.g., "this is ecru, not red"). Learning in these instances has proved modest at best, even in children as old as 36 months (Au & Laframboise, 1990; Au & Markman, 1987; Carey, 1978; Smith, Jones, & Landau, 1992; Taylor & Gelman, 1988). Cross-linguistic studies of artificial word learning also suggest that names for concrete things are special in early language learning (Imai & Gentner, 1993; Waxman, 1994) in that there are considerable similarities in the nature of children's noun extensions across languages and considerable variability across (and within) languages in young children's interpretation of novel adjectives. Other evidence from children learning English suggests that the initial meanings of dimensional terms may be highly context specific (Keil & Carroll, 1980). In sum, whereas names for things appear to be "fast mapped" (Carey, 1982) to potential categories, the extension of a novel adjective appears more slowly and more variably determined.

The third class of evidence concerns children's errors with nominal and adjectival meanings. There are extensive literatures in both areas although they are difficult to compare because of vastly different methods, ages of subjects, and empirical questions asked. These differences derive directly from the noun advantage over adjectives. The key question for researchers who study early noun acquisition is how it is that children learn so many nouns so rapidly and with so few errors. The only errors consistently studied in this literature are the overextension errors typically noticed at about the time productive vocabulary first begins to accelerate. However, there is a debate as to whether these errors are category errors. Instead, these overextensions (for example, calling a zebra "doggy") may reflect pragmatic strategies or retrieval errors (Gershkoff-Stowe & Smith, 1996; Huttenlocher, 1974). Consistent with this idea is the rarity of overextensions in comprehension (see, for example, Naigles & Gelman, 1995).

In contrast, the key question for researchers who study the acquisition of dimensional adjectives is why they are so difficult to learn. The central phenomena are comprehension errors. Long after children begin to use dimensional words, when they are as old as 3, 4, or even 5 years, their interpretations of dimensional adjectives are still errorful. This literature is replete with examples of both within- and between-dimension errors, interpreting *big* to mean TALL (Maratsos, 1988), *big* to mean BRIGHT (Carey, 1978, 1982), *dark* to mean LOUD (Smith & Sera, 1992), and *blue* to mean GREEN (Backscheider & Shatz, 1993). Although plentiful, these errors are constrained. They consist of confusions within the semantic domain of dimensional terms. That is, children may confuse *dark* and *loud* but they do not confuse *dark* and *room*. The category specificity of these errors means that at the same time children are rapidly learning nouns and commonly misinterpreting adjectives, they have some idea that nouns and adjectives span different categories of meaning.

In sum, the phenomena to be explained are (1) why common nouns are acquired by young children earlier, more rapidly, and with fewer errors than are dimensional adjectives and (2) how, during the protracted course of learning dimensional adjectives, young children seem to recognize that the dimensional adjectives comprise a class.

2 Rationale for a Similarity-Based Approach

One way of construing the problem is in terms of category learning. Why are common noun categories more easily learned than common adjective categories?

Several proposals have been offered suggesting a foundational conceptual distinction between objects and their attributes. For example, Gentner (1978), Maratsos (1988), and Macnamara (1982) have all suggested that nouns are logically prior. They point out that predicates presuppose arguments but that the reverse is not true. The suggestion, then, is that children need not understand *shaggy* to figure out what *dog* means from examples like *the dog is shaggy* but must know *dog* to figure out *shaggy* from the same sentence. Similarly, Markman (1989; see also, Carey, 1994) proposed that children's initial hypotheses about word meanings adhere to a "whole-object principle" — that children assume that novel labels refer to individual whole objects rather than to their component properties or to collections of objects. Thus, by this account, children's initial hypotheses about meanings are noun-like. Although these proposals are probably somewhat correct, they seriously underspecify the processes through which knowledge about the differences between nouns and adjectives is instantiated or acquired.

We seek such specification in a similarity-based account. Our idea is that the noun advantage and an initial segregation of nouns and adjectives as distinct classes of words is the result of the most general and ordinary processes of associative learning. There are two arguments for this approach which we find compelling. First, whatever else children know or believe, similarity-based associative learning is part of their biology and thus a good place to begin looking for a mechanistic account. Second, similarity-based learning would seem crucial at the front-end when children know no language. At this point, children learn many words by ostensive definition (Mervis, 1987). Parents point to an object and say, for example, "that's a dog" or "that's big." This associative task of mapping words to perceptible properties would seem to be the very same for the learning of dimensional adjectives as for the learning of nouns. Even if the child possessed some pre-existing conceptual distinction between objects and their properties, the child could not use that knowledge at this stage because the child has no words and thus no knowledge of the syntactic frames that would distinguish whether a novel word is a noun or an adjective. In the beginning, the young child can only associate novel labels with the properties of things so labeled. Doing so will yield a representation of *dog* as things with DOG properties and a representation of *wet* as things with WET properties. While incomplete, such meanings are in fact on the right track.

Given these assumptions, we ask: Why are common nouns learned more readily than common adjectives?

2.1 Differences in Similarity Structure between Nouns and Adjectives

Previous researchers have pointed to three kinds of difference between common noun and dimensional adjective categories.

2.1.1 Many vs. Few Similarities

Gentner & Rattermann (1991), Markman (1989), Medin & Ortony (1989), and Rosch (1973a) have all argued that common nouns label objects similar across many interrelated and correlated properties. In contrast, dimensional adjectives label objects that are alike on only one property. This difference between nouns and adjectives has important conceptual consequences (see especially Markman, 1989). For example, knowing that an object is a bird allows predictions about many different properties of the object but knowing that an object is a member of the category WHITE-THINGS supports only predictions about the object's color.

This difference also has important implications for similarity-based learning, as illustrated in Figure 1. This figure represents the extensions of idealized nouns and adjectives as regions in a multidimensional space of all possible objects. The relevant spaces are hyperspaces of many dimensions, all of those along which noun and adjective meanings vary, but for ease of illustration we confine ourselves to three dimensions. For example, the dimensions shown could represent SIZE, SMOOTHNESS, and SHININESS. Each of the outlined regions within the large cube represents a hypothetical category associated with a single word, and instances of the category would be points within the region. As can be seen in the figure, categories organized by many dimensional similarities (cubes with thick outlines) are small and compactly shaped relative to those that are organized by similarity on just one property. Thus, the idealized noun is uniformly and closely bounded in all directions. It is a hypercube or hypersphere. In contrast, members of an adjective category are tightly constrained in only one direction (the relevant dimension) but extend indefinitely in all others. The idealized dimensional-adjective category thus may be thought of as a "hyperslab." Further, the volume of idealized noun categories, compact in all dimensional directions, is relatively small whereas the volume of adjective categories, extending indefinitely in all directions but one, is great.

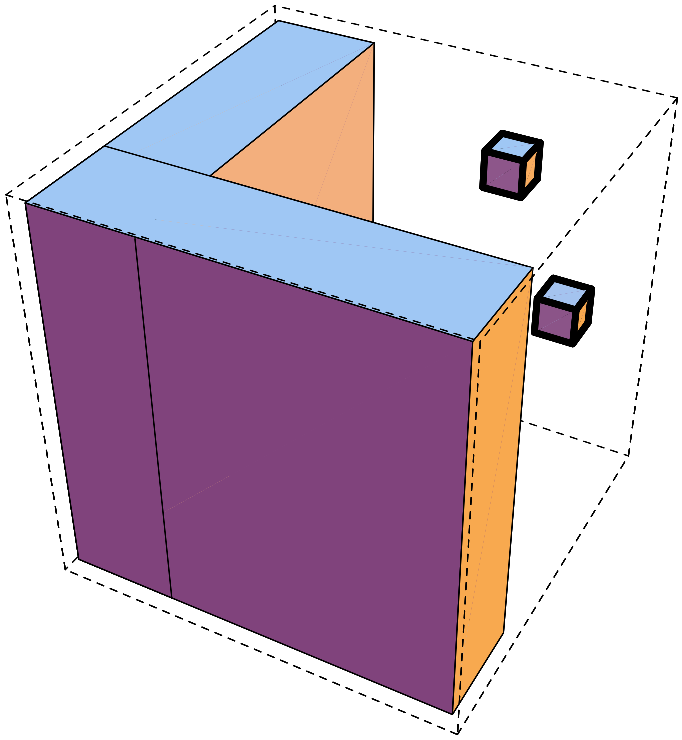

Figure 1: **Typical Noun and Adjective Categories**. Only three dimensions from the set of dimensions distinguishing the categories are shown. Noun categories appear in thick outline, adjective categories in thin outline.

Given ordinary ideas about similarity and generalization, these differences clearly favor nouns. The within-category similarity is greater for the nouns than the adjectives in Figure 2. Further for nouns, generalization can be non-selective in all directions but for adjectives generalization must be selectively inhibited in one direction. Learning about adjectives but not nouns thus requires discovering and selectively attending to one relevant direction in the multi-dimensional space.

2.1.2 Category Overlap

Nouns and adjectives also differ in the relatedness of one category to another. Common nouns all classify objects at one level (Rosch, 1973a). An object is a dog or a house or a watch or a car or a leaf. Thus the question *what is it?* is answerable by one basic noun. Markman (1989) incorporated this notion in her proposal that children adhere to a mutual exclusivity assumption in early word learning. Although this idea of a one-object, one-name rule is imperfect and complicated by a hierarchical taxonomy and synonyms, it also captures something quite real about the way common nouns are commonly used (Clark, 1973; Markman, 1989; Markman & Hutchinson, 1984; Mervis, 1987; Mervis, Mervis, Johnson, & Bertand, 1992; Rosch, 1973a). Dimensional adjectives present a markedly different structure. They are (typically) mutually exclusive within a dimension but overlap completely across dimensions. Objects in the category BIG may also be in the categories WET and FURRY.

An idealization of this difference between common nouns and dimensional adjectives is depicted in Figures 2 and 3. Relatively small noun categories fill all reaches of the space but rarely overlap with one another. In contrast, the extensions of dimensional adjectives create a dense grid-work of overlapping slabs that cut through the space in multiple directions as illustrated. Again, under the ordinary assumptions of similarity-based learning, these differences in category structure favors nouns: between-category similarity among nouns is minimal but between category similarity among adjectives is great.

2.1.3 Linguistic Associations

Nouns and adjectives also differ in their association with the linguistic form of questions about objects. Different words, for example *what is it* versus *what color is it?* are used to ask about object categories and object properties. Dimensional adjectives also differ among themselves in this regard: *what color is it* asks for a color word as an answer; *how does it feel?* asks for a description of texture. Backscheider & Shatz (1993) have shown that young children are sensitive to these associations between questions and the class of possible answers prior to their understanding of the meanings of the individual words. Thus in learning common nouns and adjectives, learners do not just map objects to words but they also map linguistic inputs to linguistic outputs.

It is not immediately clear whether these word-to-word associations favor nouns or adjectives. However, given the overlap among the to-be-learned categories, we can be certain that they are crucial to learning. A big, red, furry dog is a member of the category BIG, the category RED, the category FURRY, and the category DOG. It is the linguistic input, the question "what is it?" or "what color is it?," that specifies the relevant class of linguistic outputs. These word-to-word maps partition all the categories that the child is learning into larger proto-syntactic categories — into "noun categories," "color categories," "size categories," and "texture categories." In stages of incomplete learning,

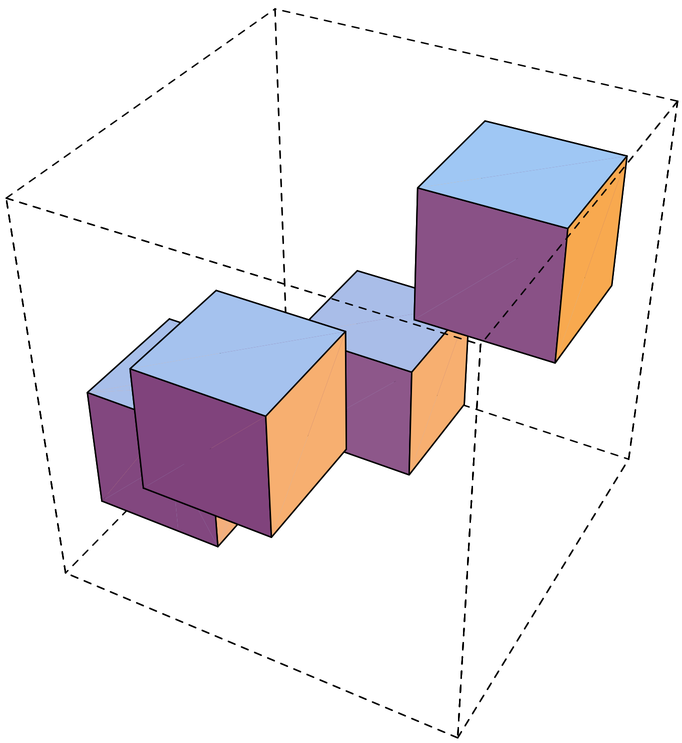

Figure 2: **Noun Categories.** Only three dimensions from the hyperspace of possible dimensions are shown. Noun categories tend to be small and compact and not to overlap with one another.

do these word-to-word maps also create a distinction between nouns and adjectives such that adjectives are confused across dimensions but are not confused with nouns?

In what follows, we demonstrate that a simple associative device that approaches the task of learning about nouns and adjectives in the very same way will nonetheless show a noun advantage and also the pattern of within-category confusions shown by children. In addition, we separately investigate the roles of category shape, volume, overlap, and word-word associations in forming this developmental trajectory.

3 A Connectionist Categorizer

To test our hypothesis that the noun advantage in early acquisition derives from the associative structure of the learning task, we used the most common similarity-based learning procedure in the literature — a three-layer connectionist network trained with back-propagation. Such a general learning device embodies no prior knowledge about differences between nouns and adjectives, and learning is purely associationist and error-driven.

As in several other recent modeling studies (Plunkett, Sinha, Møller, & Strandsby, 1992; Schyns, 1992), we investigate the behavior of a simple connectionist network which is trained to label a set of patterns representing perceptual inputs to the system. The

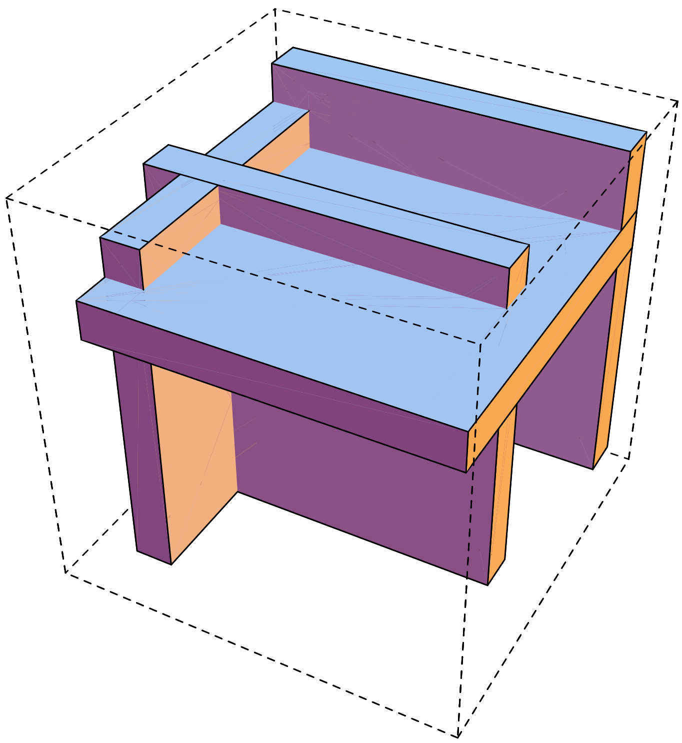

Figure 3: **Adjective Categories.** Only three dimensions are shown. Adjective categories tend to be large and elongated and to overlap with one another.

goal in these studies is to show how the facts of lexical development emerge from the interaction between the learning device and the regularities inherent in the input patterns. In our case, the relevant facts concern the relative ease of learning nouns and adjectives, and the regularities in the patterns concern differences in the way noun and adjective categories carve up the space of input dimensions and co-occur with particular linguistic contexts.

The main difference between our network and other simple connectionist models is our use of a modified form of back-propagation. Back-propagation is suitable in that early word learning in children is "supervised." Adults ask children questions about objects (e.g., "what is that?," "what color is that?") and they provide feedback (e.g., "that's not a dog; it's a horse") (Callanan, 1990; Mervis, 1987; Snow, 1977; Wood, 1980). Supervision for categorization tasks such as our word-learning task, as typically realized in connectionist networks, however, is psychologically unlikely. If separate output units represent the different category responses, standard back-propagation changes the connection weights on each learning trial in a way that encourages the correct response and discourages all other potential responses. This is like the parent saying to the child, "This is a dog, not a plate, not a cat, not an apple, not a house..." Parents do not do this but instead explicitly reinforce correct answers ("yes, that's a doggy") and provide negative feedback only when the child explicitly gives the wrong answer ("that's not a doggy; it's a horse").

This form of back-propagation is also inappropriate in the present case because in the combined task of naming objects and labeling their attributes, possible responses are not just right or wrong. There are kinds and degrees of wrongness. Consider a big, black, wet dog and the question "what color is it?" The answers "dog" and "red" are both wrong. However, it seems unlikely that parents would respond to these errors in the same way. A toddler who answers the question "what color is it?" by correctly naming the dog "dog" seems likely to hear a parental response of "yes, it's a dog, a black dog." A toddler who answers the same question by saying "red" is likely to hear, instead, a parental response of the sort "it's not red, it's black."

Accordingly, we modified the back-propagation algorithm to fit these assumptions about the kinds of feedback provided by parents. Briefly, we provided targets only for a limited number of output words, and we distinguished the kinds of incorrect errors by using distinct targets for them. In the next two sections, we provide a detailed description of the network and the learning rule.

3.1 The Network Architecture

Figure 4 shows the network architecture. Each thin arrow represents complete connectivity between two layers of processing units. The network is designed to take objects and a linguistic context as inputs and to produce a noun or adjective as output.

Inputs to the network are presented to two layers of processing units, one for the representation of the object itself and one for a linguistic context corresponding to a question the network is asked. Input objects consist of patterns of activation representing a perceptually present object in terms of a set of sensory dimensions. For the simulations discussed in this paper, the inputs are specified in terms of four or five dimensions. We require that the network learn to associate points along each dimension with particular words, so the simplest possible representation of a dimension, that is, a single unit, is excluded because it would only permit the association to different degrees of the dimension as a whole with each word. Therefore each dimension takes the form of a group of units in the input layer of the network. That is, input to the network along a given dimension consists of a vector of numbers, each between the minimum and maximum activation values of the units in the input layer of the network. There are several ways to represent dimensional input in the form of a vector, varying in the extent to which they make explicit the ordering of points along the dimension. At one extreme is a completely localized encoding, in which each dimensional vector contains one maximum value and the remainder of the numbers take on the minimum value. This form of encoding completely obscures ordering along the dimension because there is no correlation between the numbers in different positions in the vector (or the activations of units in each dimension group). At the other extreme is a "thermometer" encoding (Harnad, Hanson, & Lubin, 1991). In a thermometer representation, each of the positions in the vector corresponds to a point along a scale, and the value to be encoded normally falls between two of the positions. All of those positions to the "right" of this point take on their minimum values, the first position to the "left" of this point takes on an intermediate value, and all of the other leftward positions take on their maximum values.

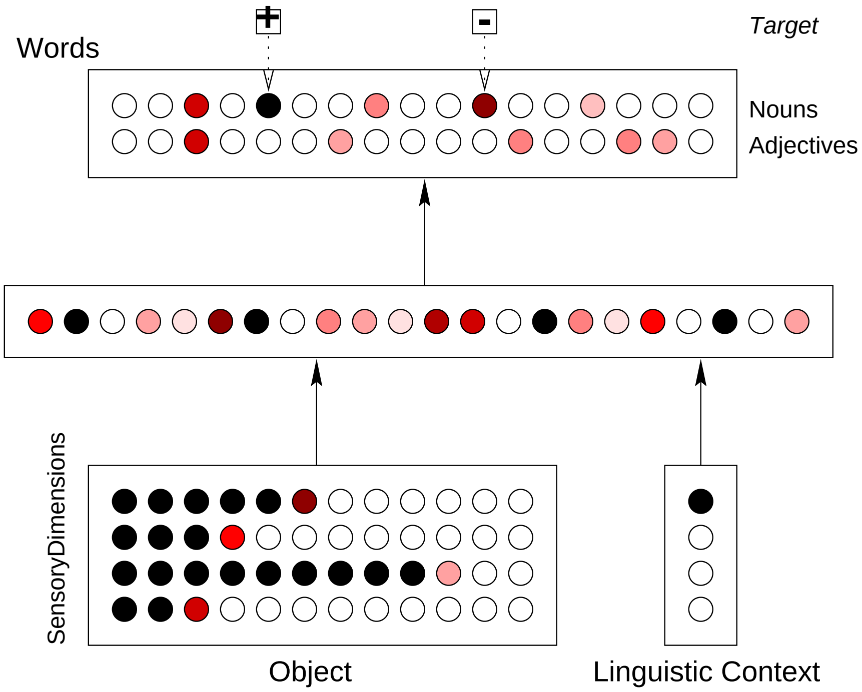

Figure 4: **The Network.** Each small circle is a processing unit, and each rectangle a layer of processing units, unconnected to each other. An arrow represents complete connectivity between the units in two layers. A possible input pattern and network response to it are shown, the degree of shading of each unit representing its activation level. The small squares at the top of the figure indicate the two targets which the network receives for this pattern, one for the correct response and one for an incorrect output above the network's response threshold.

In this paper, we confine ourselves to thermometer representations.[2] In the networks used in the experiments reported here, each dimension is represented by 12 units which have maximum activations of 1 and minimum activations of 0. So in the network, dimensional values of 3.3 and 8.8 along the scale with maximum value of 12 would be represented as the patterns [1, 1, 1, .3, 0, 0, 0, 0, 0, 0, 0, 0] and [1, 1, 1, 1, 1, 1, 1, 1, .8, 0, 0, 0]. The figure illustrates a possible set of activations along each of the four sensory dimensions for an input object.

The linguistic context input consists of a question of the form *what size is it?*, *what color is it?*, or *what is it?*, each question represented by a separate unit in the linguistic context layer of the network. (Four units are shown in the figure.) It is important to

[2]We have no reason to believe, however, that the conclusions we reach will not generalize to other representational schemes. An alternative, for example, is a variant of localized encoding in which units on either side of the most highly activated unit are also activated, in inverse proportion to their distance from the activated unit. A version of the present network using such a scheme trained on the data generated for Experiment 3 below exhibited the same advantage for compact over elongated categories as was found with thermometer encoding.

note that, because the network is given no actual syntactic context, the noun context (*what is it?*) is indistinguishable from the adjective contexts (*what color is it?*, etc.) at the start of training. In terms of the network's architecture, there are just several equally different linguistic context inputs that might be viewed as corresponding to noun, color, size, and texture. There is no hierarchical organization of the adjective terms in the architecture; that is, there is nothing that groups the adjectives as a class in opposition to the nouns.

Critically, from the perspective of the network, there is also no distinction between the input activation that corresponds to the object and that which corresponds to the question. From the network's point of view, there is just one input vector of 66 numbers jointly specifying an event in the world in terms of the five perceptual dimensions and the linguistic context input that co-occurs with the presentation of the object.

The hidden layer of the network compresses the input patterns into a smaller set of units, 15 to 24 units in the experiments we report here.[3] Thus at this level, the system no longer has direct access to the input dimensions. This is an important aspect of the architecture and an important theoretical claim. It means that input dimensions that are distinct at input are not (at least not without learning) represented separately. This aspect of the architecture is based on considerable research indicating that young children have difficulty attending selectively to individual dimensions (Aslin & Smith, 1988) and on our past use of this architecture to model developmental changes in selective attention to dimensions (Gasser & Smith, 1991; Smith, 1993). We will discuss more fully the wider implications of this aspect of the network in the General Discussion.

The output layer consists of a single unit for each adjective and noun. A +1 activation on an output unit represents the network's labeling the input object with the corresponding word. A -1 activation represents the network's decision that the corresponding word is inappropriate for the input object, and a 0 activation represents an intermediate response, one that might be made if an object is described by the category but that is not an appropriate answer to the linguistic input question, for example, if "red" were the response to the question "what is it?" for a red dog.

3.2 The Learning Rule

The specific learning rule used operates as follows. During training, a target is associated with each input pattern; this target represents the appropriate response to the input. In ordinary back-propagation, each output unit receives a target on each trial. But, as noted above, this is an implausible procedure, as it means that all possible responses which are not appropriate are punished. Further, as noted above, not all wrong answers are wrong in the same way and unlikely to be responded to the same way by parents. Accordingly, we give the network feedback for only two sorts of words, the correct word and any incorrect words to which the network has made a significant response. We defined a "response threshold" for the word units, 0.05 in all of the experiments reported on here; only activations above this threshold are treated as overt responses for which feedback

[3] Increasing the number of units in the hidden layer of the network both speeds up performance and leads to improvement in the asymptotic level of performance.

is possible. Further, the target for these explicit errors depends on the input as follows.

1. The target for a correct response is +1.
2. For a response which is not a correct label for the input object under any circumstances (e.g., "small" for a large, red object), the target for the corresponding output unit is -1.
3. For a response which would be a correct label for the input object if it matched the lexical dimension input (e.g., "large" for a large, red object when the input question is "what color is it?"), the target for the corresponding output unit is 0.

4 Experiments

4.1 Experiment 1: Nouns vs. Adjectives in General

In Experiment 1, we investigate how this simple three-layer network simultaneously learns many categories organized to be like nouns and to be like adjectives with respect to the properties of shape, volume, overlap, and number of different categories. The central question is whether their will be a noun advantage early in learning and whether, prior to complete learning, the network will show partial knowledge that nouns and adjectives are distinct classes of words.

4.1.1 Stimuli

The input to the network consisted of an object described on five perceptual dimensions and the question accompanying the object. The input objects were instances of 30 possible categories. Each input object had a value for each of the five perceptual dimensions, and each category was defined in terms of the range of values that its instances could take along each of the dimensions. Twenty of these categories were organized to be noun-like and 10 were organized to be adjective-like. Each noun was defined in terms of a range of 1/10 of the possible values along each of the five input sensory dimensions. Each adjective category was defined in terms of a range of 1/5 of the possible values along one of the input dimensions and any value along the other four. Thus each noun spanned $1/10 \times 1/10 \times 1/10 \times 1/10 \times 1/10 = 0.00001$ of the multi-dimensional space of all possible categories whereas each adjective spanned 1/5 of the space. Table 1 shows ranges of possible values on the five dimensions for two of the noun and three of the adjective categories. Note that the noun categories may overlap on one or more dimensions (dimensions 2 and 5 in the example categories). No noun categories overlap completely, however. This is not so for the adjective categories. In Table 1, adjective 1 overlaps with both adjective 2 and 3 because it is possible to create an object which is an instance of both adjective 1 and adjective 2 or both adjective 1 and adjective 3.

The ten adjective categories were organized into five lexical dimensions by association with the specific input dimension whose values were constrained within the adjective category and by association with a specific linguistic context input, e.g., "what size

	Perceptual Dimensions				
Noun 1	$0.9 < v_1 < 1$	$0 < v_2 < 0.1$	$0 < v_3 < 0.1$	$0 < v_4 < 0.1$	$0 < v_5 < 0.1$
Noun 2	$0 < v_1 < 0.1$	$0 < v_2 < 0.1$	$0.4 < v_3 < 0.5$	$0.4 < v_4 < 0.5$	$0 < v_5 < 0.1$
Adj 1	*any*	*any*	*any*	$0.8 < v_4 < 1$	*any*
Adj 2	$0 < v_1 < 0.2$	*any*	*any*	*any*	*any*
Adj 3	$0.8 < v_1 < 1$	*any*	*any*	*any*	*any*

Table 1: **Experiment 1: Ranges of Values on Perceptual Dimensions for 5 Input Objects.** v_1, etc. represent the values on the five dimensions. Each range is expressed in terms of proportions of the distance from the minimum to the maximum value.

is it?" Thus the ten adjectives were structured into five dimensions each with two contrasting terms.[4] In Table 1, adjectives 2 and 3 belong to the same lexical dimension.

For each training instance, the inputs were generated as follows. First an output category was selected at random from the set of 30 possible outputs (the 20 nouns and the 10 adjectives). The selection of the relevant output determined the linguistic context input. Then for each of the five perceptual dimensions, a possible value was picked at random consistent with the selected output.

The linguistic context input consisted of the pattern representing a question that would be appropriate for the selected category, each question corresponding to a lexical dimension. For example, if the category was *big*, the input unit representing *what size it is?* was turned on (that is, its output was set to 1.0), and the other linguistic context units were turned off. If the category was *dog*, the input unit representing *what is it?* was turned on, and the other linguistic context units were turned off.

Because there was randomness in the selection of output categories and corresponding input objects, because the input objects varied continuously, and because the targets depended in part on the network's response, the network was never trained more than once on a particular input-target pair.

4.1.2 Method

On each training trial, the network was presented with an input (object plus linguistic context), generated as just described, and an appropriate target on the output. The weights in the network, other than those feeding output units for which no targets were available, were then adjusted according to the back-propagation algorithm.

Following each presentation of 1000 input patterns the network was tested on 500 novel inputs generated in the same fashion as the training patterns. There are several options for evaluating the network's performance. We chose to look only at the output unit with the highest activation, unless this unit's activation was not above the response threshold, in which case the network was viewed as not making any overt response at all. Our assumption was that production processes not modeled in our network would

[4] As we will see in subsequent experiments, the noun advantage in the network does not depend on there being only two terms for each adjective dimension.

force the system to select one word over all of the candidates which might be activated. Thus only the most highly activated output unit was relevant. For each test input, following activation of the network it was determined whether the output unit with the highest activation was above the response threshold and whether that unit corresponded to the appropriate word. Performance for each category of word was measured as the proportion of test trials for which this was true.

4.1.3 Results

Figure 5 shows the learning rates for adjectives and nouns in this experiment. The data shown are averages over 10 runs with different initial random weights on the network's connections. The smaller and more compactly shaped noun categories are learned much faster than the larger and more slab-like adjective categories ($p < .001^5$). Performance on the nouns is close to perfect by the 2000th training trial. Performance on the adjectives continues to improve, but never reaches the level of the nouns.[6]

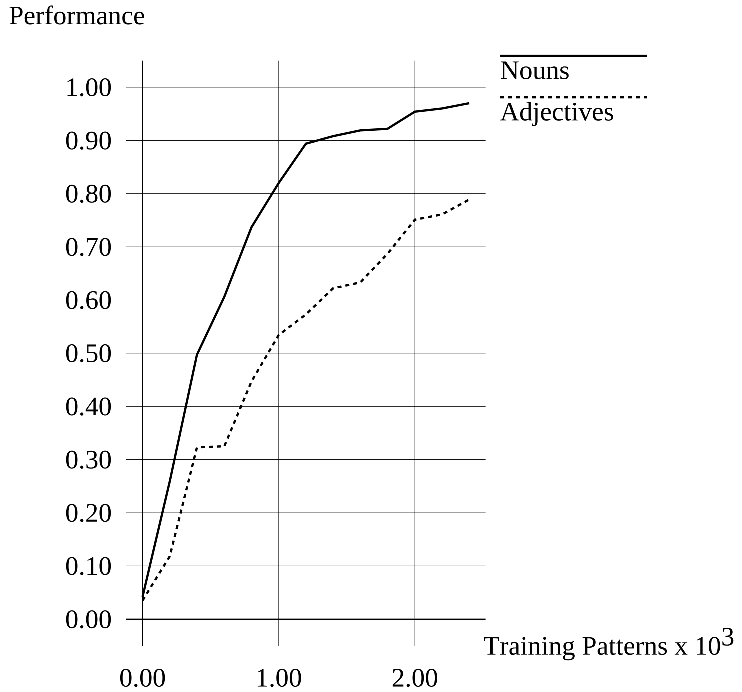

Figure 5: **Experiment 1: Nouns vs. Adjectives.** Performance is the proportion of test items for which the highest overt response was correct. Responses are averaged over 10 separate runs of the network.

We also asked whether in learning these categories, the network showed any implicit knowledge of lexical categories. First, does the network develop a distinction between nouns and adjectives as a class? Second, does the network develop a distinction between

[5]For statistical tests here and in Experiments 2–5, we treated each run of the network as a separate subject.

[6]An initial difference in learning but ultimately equal and near perfect learning of both nouns and adjectives is achieved with larger hidden layers.

Incorrect output	0 Training Patterns		1000 Training Patterns	
	Noun Context	Adj Context	Noun Context	Adj Context
Nouns	**.66**	.34	**.65**	.35
Adjectives	.70	**.30**	.37	**.63**

Table 2: **Experiment 1: Within- and Between-Part-of-Speech Errors.** Figures represent the proportion of incorrect overt responses in different part-of-speech categories.

different dimensional terms, analogous to knowing, for example, that wet and dry are attributes of one kind and that rough and smooth are attributes of another kind? These are important questions because children show clear evidence of the first distinction in their early errors but not the second distinction (see Carey, 1994; Smith, 1984; Smith & Sera, 1992; but see Backscheider & Shatz, 1993).

To answer the first question, we defined "within-part-of-speech errors" as the proportion of cases with an incorrect response (above threshold) for which the response was the correct "part of speech" (adjective or noun). Table 2 shows the proportion of within- and between-part-of-speech errors at the start of learning and after 1000 training trials. At the start of learning when the network knows nothing, the relative frequency of noun and adjective responses (2:1) corresponds to the relative number of noun and adjective output units (2:1) and is unrelated to the linguistic context input. However, as learning progresses, the character of the error becomes associated with the linguistic input that specifies the class of possible answers. After 1000 training trials, when the network still has not yet fully acquired the adjective terms, the network shows implicit knowledge that all the adjectives form a class.

To answer the second question, we defined "within-dimension errors" as the proportion of cases in which adjective questions received incorrect adjective responses and the response was on the right dimension. Noun questions and noun responses to adjective questions did not contribute to this measure. At the start of training, such within-dimension errors were rare, occurring .08 of the time. The frequency of within-category errors increased with training, reaching a maximum of .23 of the time after 2000 trials. Thus the network shows little implicit knowledge of which terms refer to attributes on the same dimension.

4.1.4 Discussion

The central result of this simulation is that a simple connectionist network when simultaneously trained on adjective-like and noun-like categories learns the nouns faster, just as children do. Yet this difference is not due to any built-in preferences on the part of the network nor to any pre-training representation of a difference between nouns and adjectives. It is due entirely to the similarity structure inherent in the learning task — that is, to the nature of the categories which the network learns and the linguistic input which specifies which of several classes of overlapping categories is the relevant one. In brief, a learner can show a marked advantage for the learning of one kind of category over another without any built-in distinction between them. The developmental precedence of nouns over adjectives in children thus need not derive from a priori conceptual distinctions, as commonly assumed, but rather from quite general similarity-based learning mechanisms.

During the course of learning, the network, like young children, also exhibits a structured pattern of errors — dimensional terms are confused with each other and not with nouns. This distinction emerges as a consequence of simultaneously learning not a single adjective class but several different adjective categories. The most likely possibility is that this is accomplished by the rapid learning of noun categories. That is, what the network "really knows" may essentially be that adjectives are "not nouns." The implication is that this may be all that young children know too (see Smith, 1995 for a similar suggestion based on empirical evidence from children). The network did not show strong learning of the connection between pairs of terms on a single dimension. This is also consistent with the evidence from children. With the exception of color terms, between-dimension rather than within-dimension confusions characterize children's initial errors (Backscheider & Shatz, 1993; Carey, 1994; Smith & Sera, 1992).

This experiment thus demonstrates the viability of a similarity-based approach to the noun advantage in children's early lexical acquisitions. In the following experiments, we examine the specific contributions of the volume and shape of category extensions, overlap and word-word associations in creating the noun advantage by examining unnaturally structured classes of categories that differ only in their volume, shape, overlap, or associations between linguistic context inputs and outputs.

4.2 Experiment 2: Category Volume

In this experiment, we investigate the role of volume differences. We create small categories and large categories that are both like nouns in being defined by similarities on many dimensions. We ask whether smaller categories of this kind have an advantage over larger ones.

4.2.1 Stimuli and method

Stimuli for this experiment were generated analogously to those in Experiment 1. There were two types of categories, those which spanned relatively wide regions of the space of all possible input objects and those which spanned relatively narrow regions. Both the Small set and the Large set contained 18 words. In the Small set, each word was defined in terms of a range of 1/6 of the possible values along each input dimension. Thus the extension of each of these categories covered $1/6 \times 1/6 \times 1/6 \times 1/6 = 0.00077$ of the space of possible inputs. In the Large set, each word was defined in terms of a range of 1/3 of the possible values along each input dimension, a total of $1/3 \times 1/3 \times 1/3 \times 1/3 = 0.012$ of the space of possible objects, that is, 16 times the size of the region occupied by the extension of each of the categories in the Small set. Note that the volumes of the two sets are closer than in the first experiment. The Large and Small categories overlapped in the space of all possible categories. Two linguistic context inputs were used to signal the relevant kind of category, one for which the Large-volume words were appropriate responses, the other for which the Small-volume words were appropriate responses. Given the relatively simpler learning task with fewer overlapping categories, we tested the network after every 500 training trials.

4.2.2 Results and discussion

Figure 6 shows the mean correct responses over 10 separate runs of the network. As can be seen, outputs referring to Small categories are learned faster than the ones referring to Large categories ($p < .001$). The difference is smaller than in Experiment 1 probably because the ratio of Large-to-Small volume is smaller: 16 to 1 in this experiment, but 216 to 1 in Experiment 1.

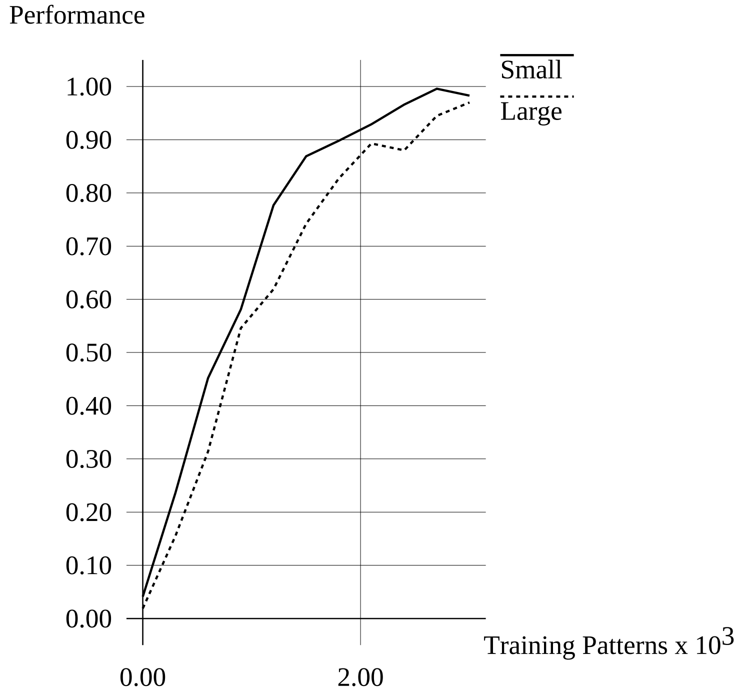

Figure 6: **Experiment 2: Category Volume.** Performance is the proportion of test items for which the highest overt response was correct. Responses are averaged over 10 separate runs of the network.

The network also readily learned the association between one linguistic context input and the class of Large-volume outputs and between the other linguistic input and the class of Small-volume outputs. As in Experiment 1, we examined "within-part-of-speech errors;" here the Small-volume and Large-volume categories represented the two parts of speech. At the start of learning, "within-part-of-speech" errors comprised (as expected by chance) about half the errors for both Small-volume and Large- volume targets (.50 of the errors given a Small-volume targets and .45 of the errors given a Large-volume target). After 1000 training trials, however, within part-of-speech errors predominated, .88 of all errors given a Large-volume target and .81 of the errors given a Small-volume target. These results again demonstrate the role of word-word associations in the network's learning.

In sum, this experiment shows that differences in the volume of a category, one of the differences that exists between common nouns and dimensional adjectives is sufficient to create an advantage in learning. This is not an unexpected result, given all that is known about the importance of within-category similarity to similarity-based learning. But it is a result that is consistent with the idea that developmental differences between the early acquisition of nouns and adjectives could derive from processes no more complex than those embodied by a three-layer connectionist network.

4.3 Experiment 3: Category Compactness

In this experiment, we investigate the effect of differences in the shape of category extensions on learning when the volumes of the to-be-learned categories do not vary. Recall that the shape or compactness of the category concerns the number of dimensions (or directions in the space of all possible objects) on which there is a restricted range of values within the category. In order to determine how important compactness, independent of volume, is for learning by a simple associative device, we contrasted compact noun-like categories with less compact adjective-like categories of the same volume.

4.3.1 Stimuli and method

The stimuli were generated as in Experiments 1 and 2. The 16 less compact "adjective-like" categories were defined in terms of ranges of 2/3, 2/3, 1/3, and 1/12 of the possible values along the four input dimensions used in this experiment. That is, one input dimension, the one for which the possible within-category range was 1/12th of the input dimension, was much more relevant than the other three in defining the category. Each of the four dimensions played this role for four of the adjectives. Each of the more compact noun-like categories was defined in terms of a range of 1/3 of the possible values along each input dimension. The extensions of both the noun-like and adjectives-like categories encompassed the same volume (1/81 of the space). The noun-like and adjective-like categories overlapped in the space. As in Experiment 2, "noun-like" categories were associated with a linguistic context input specifying noun targets and all the adjective-like categories were associated with one linguistic input specifying adjective targets.

4.3.2 Results

Figure 7 shows the results of Experiment 3 over 10 runs of the network. The noun-like categories that were organized by an equally restricted range of variation on all four sensory dimensions were learned more rapidly than the adjective-like categories in which the range of variation on some dimensions was wide and on others narrow ($p < .001$). In other words, evenly compact categories are more rapidly learned than elongated ones, a difference which again favors the basic-level nouns children learn early over the dimensional adjectives that they learn later. We also assessed the association of noun and adjective outputs with the two different linguistic inputs by measuring within- and between-category errors. At the start of learning, within category errors were at chance; the proportions of all errors (above threshold responses) that were within syntactic category were .45 and .53 for nouns and adjective respectively. After 2000 trials, the proportions of within-category errors were .85 for both nouns and adjectives. Given that the input specified two categories, this result is not surprising but it does demonstrate again the learning of word-word associations and their potential role in generating structured patterns of errors.

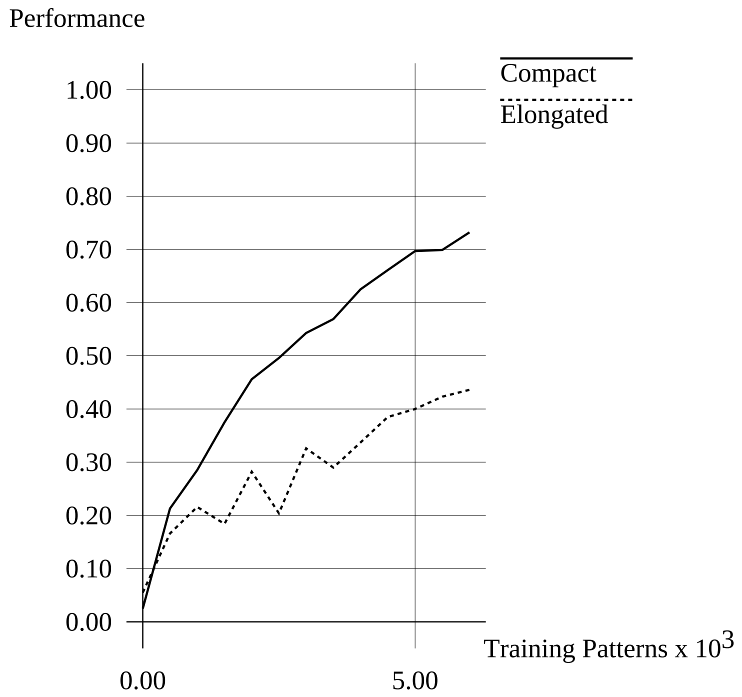

Figure 7: **Experiment 3: Category Compactness.** Performance is the proportion of test items for which the highest overt response was correct. Responses are averaged over 10 separate runs of the network.

4.4 Experiment 4: Linguistic Associations

In Experiment 1, and we believe in the labeling tasks faced by young children in world, noun and adjective categories differ in their volume, compactness, and in their association with with specific linguistic contexts. In this fourth experiment, we ask how the association between lexical dimensions in the input and the specific adjectives that comprise the output contribute to the noun advantage. We do this by creating two classes of words whose extensions did not differ in volume nor shape. Each category was organized principally by variation along one input dimension. Four adjective-like categories were defined by associating all categories organized by one input dimension (e.g., color or size) with the same linguistic context unit. Thus there were four adjective categories associated with four linguistic inputs specifying the relevant object dimension. The "noun" categories were defined by taking the very same categories (each organized by one input dimension) and associating them with a single linguistic context input. Thus we ask whether it helps or hurts in learning the very same categories to have linguistic inputs specifying subsets of outputs or to have no linguistic inputs that specify subclasses of outputs. Because the linguistic context inputs in the first case also specify the relevant dimension, we call them “lexical dimensions.”

4.4.1 Stimuli and method

As before, stimuli for this experiment were generated randomly, given the constraints which defined each of the categories. As in Experiment 1, adjectives were organized along lexical dimensions, specified by the most relevant input dimension and the linguistic context input. In this case, there were four lexical dimensions, one each for the four input dimensions that specify the presented objects.

Unlike in Experiment 1, however, the adjective and noun categories were identical in every other way; in fact, the same set of 16 categories was used for the 16 nouns as well as the 16 adjectives. For all categories a single sensory dimension was most relevant; that is, the range of variation possible along that dimension was considerably narrower than on the other three dimensions. For example, one adjective category was defined in terms of ranges spanning 2/3, 2/3, and 1/3 of three of the input dimensions and 1/12 of the relevant dimension, and one of the noun categories was defined in exactly the same way. Whereas the noun and adjective categories overlapped completely (since they were identical categories), there was no overlap within the noun and adjective classes. This is necessary for the condition with no linguistic context specifying the lexical dimensions; without such linguistic input, it would be impossible to learn overlapping categories. Thus in this experiment, the only factor distinguishing the two classes of outputs is the presence of linguistic contexts associated with subsets of words and specific perceptual dimensions.

4.4.2 Results

Figure 8 shows the results of this experiment over 10 separate runs. There is an advantage for words associated with specific lexical dimensions ($p < .001$). Thus, rather than adding complexity to the learning task, linguistic input dimensions, in the absence of category overlap, provide redundant information about category identity that aids learning.

The network again readily formed two “syntactic” categories presumably by associating the class of words for which there were no lexical dimensions in the linguistic context with the one linguistic context specifying that class. At the start of learning, the network's errors were distributed equally among the noun-like set and adjective-like set of outputs; the proportion of within class (above threshold) errors were .47 and .52 respectively. After 4000 trials, however, errors were predominantly from within the proper “part of speech”; when the correct output was from the noun-like set, the network erred by responding with another item from that set .82 of the time and when the correct output was from adjective-like set, the network erred by responding with another item from that set .84 of the time. With these non-overlapping categories, the network also made within-dimension errors for the adjectives. These were .18 at the start of learning and .86 after 4000 trials.

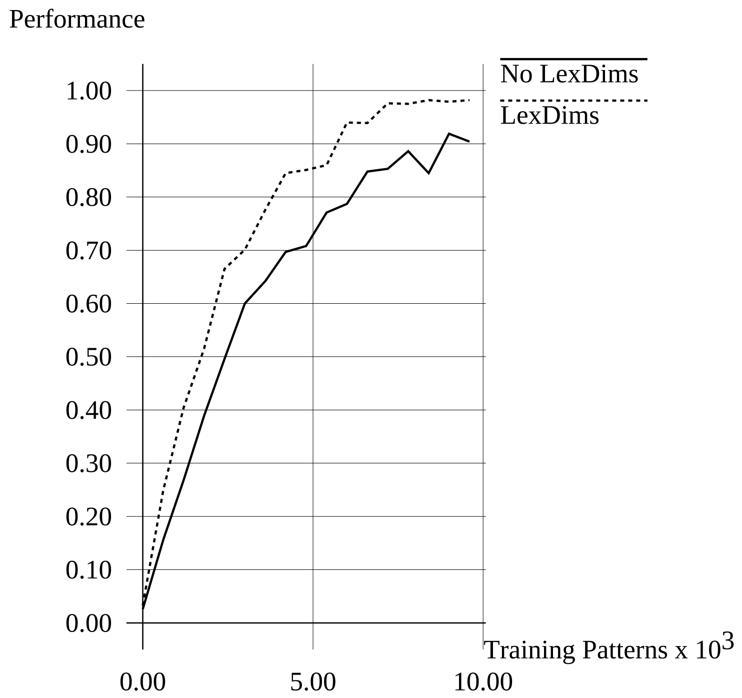

Figure 8: **Experiment 4: Lexical Dimensions.** Performance is the proportion of test items for which the highest overt response was correct. Responses are averaged over 10 separate runs of the network.

The principal result from this simulation is that, all other things being equal, learning subcategories of associated questions and responses provides an advantage.

4.5 Experiment 5: Category Overlap

One aspect of Experiment 4 is highly artificial, however. In the world, lexical dimensions are tied closely to the massive overlap of adjective categories. Far from providing redundant information about category identity, lexical input ("what color is it?") functions to pick out one label true of the specific object (e.g., *red*) from a large set of other labels also true of that same object (*big*, *furry*, *wet*, *dog*). In Experiment 5, we investigated the effect of category overlap in the context of lexical dimensions.

4.5.1 Stimuli and method

We defined categories in this experiment analogously to those in Experiment 4. The extension of each category encompassed 1/64 (1/2 x 1/2 x 1/2 x 1/8) of the representational space and thus was constrained principally on one of the four object input dimensions. Sixteen overlapping categories and 16 non-overlapping categories were defined. Four categories within each set were restricted in their range of variation principally on one of the four input dimensions. We trained separate networks to learn the overlapping and non-overlapping categories. For the overlapping categories, four linguistic context inputs specified the relevant input dimension and the subclass of outputs. In the non-overlapping case, four linguistic inputs provided redundant information about subclasses of outputs and thus were not necessary to distinguish a correct from an incorrect category.

4.5.2 Results

As can be seen in Figure 9, the non-overlapping categories were learned considerably faster than the overlapping categories ($p < .001$). Even in the context of disambiguating lexical dimension inputs, overlapping categories are more difficult to learn than non-overlapping ones. Since lexical dimensions in the linguistic context favor adjectives, but overlap (along with volume and compactness) favors nouns, these results are consistent with the idea that the developmental trajectory observed in children may arise from a consortium of differences between the associative structure of nouns and adjectives that jointly but not necessarily singly favor nouns.

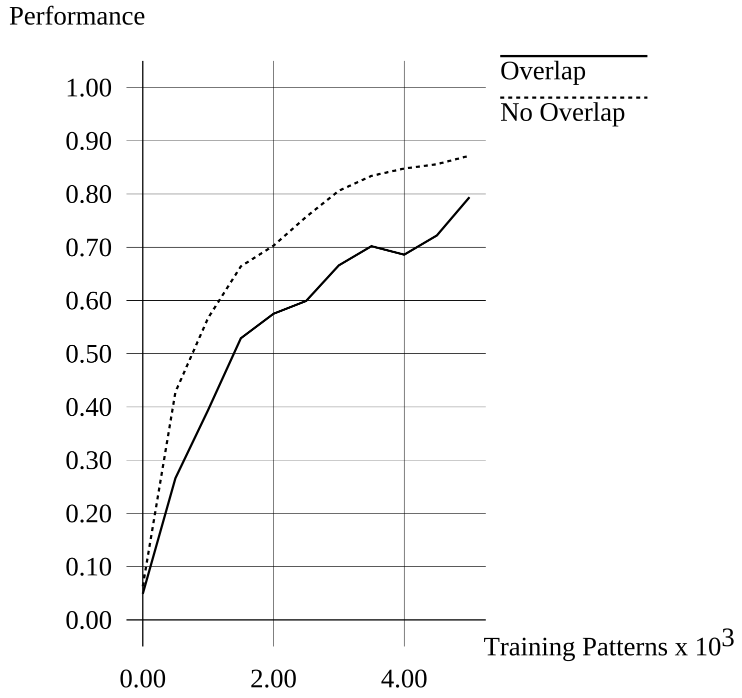

Figure 9: **Experiment 5: Category Overlap.** Performance is the proportion of test items for which the highest overt response was correct. There were two separate runs of the network, one for each condition.

4.6 Experiment 6: Emergent Syntactic Categories

In all of the experiments we have described, there are two classes of categories to be learned, nouns and adjectives, differing in one or more ways. The task of the network is to learn the categories, and we have shown how certain differences between classes of categories can affect the rate of and ultimate level of learning. The network's task is *not*, however, to learn that there are two classes of categories and to discover how these classes are distinguished. Ultimately children do learn to make this distinction. Does our simple model have anything to say about how this is accomplished?

While the network starts the task without the knowledge that there are two classes of categories, it does have access to a much more direct indication of the distinction: the linguistic contexts associated with the two classes of words. More precisely, what these inputs tell the network is simply that there is a distinction to be made. But does the network use the linguistic context inputs in this way? The explicit task of the network is to map input objects, accompanied by linguistic contexts, onto one label or another. However, if the linguistic context is informative for this task, then we would expect the network to also learn to associate particular contexts with particular words. These associations, in a sense, would constitute the beginnings of syntactic categories. In this final experiment, we ask what the network can learn when the meta-categories associated with specific linguistic inputs, that is, noun and adjective, are more arbitrarily defined than the classes of categories thus far examined. If noun and adjective are just arbitrary collections of categories, the network will have to rely on the linguistic context input if it is to learn anything about these meta-categories.

4.6.1 Stimuli and method

As before, stimuli for this experiment were generated randomly, given the constraints which defined each of the categories. Two classes of categories (noun and adjective) were defined that were identical with respect to all of the variables of interest (volume, compactness, lexical dimensions, overlap). They differed only in terms of where the member categories were located in the representational space. The categories, 18 in each class, were defined in such a way that in the representational space, each noun category was surrounded by adjective categories and vice versa. The pattern of noun and adjective categories resembled a multi-dimensional checkerboard. Thus at the level of the meta-categories, there was no generalization whatsoever to be made about the nature of the member categories or the particular regions associated with nouns or adjectives. In a sense, the meta-categories had no semantics associated with them. Each category took up .003 of the space; this left uncategorized regions of representational space separating adjacent categories. There was no overlap between categories. As in experiments other than Experiment 1, there were four input dimensions defining the perceptual properties of the object, but in this case, there were only two linguistic context inputs, one for one class of words and the other for the second class.

As in all of the experiments, the network was trained on randomly generated instances of the categories. In this case, the network was *tested*, however, on a set of 18 pre-defined object input patterns which did not belong to any of the noun or adjective categories; that is, these inputs fell in the gaps between the categories which the network had been trained on. Each of these 18 patterns was tested once together with a noun linguistic context and once with an adjective linguistic context. The relevant dependent variable in each case is the relative activation over the noun and the adjective output units. If the network has begun to divide the words into meta-categories on the basis of the linguistic context, we should see higher mean activations on the adjective units when the adjective

Word Output	Input Linguistic Context	
	Noun	Adjective
Nouns	**-.103**	-.165
Adjectives	-.162	**-.113**

Table 3: **Experiment 6: Noun and Adjective Response to Noun and Adjective Linguistic Contexts.** Figures show the mean activation of noun and adjective output units in response to 18 object input patterns which belong to neither meta-category and which are presented together with either noun or adjective linguistic contexts.

linguistic context is presented and higher activations on the noun units when the noun linguistic context is presented.

4.6.2 Results

Table 3 shows mean output activations for the four cases.[7] There is a strong interaction ($p < .001$[8]): output activations are higher for words in the meta-category corresponding to the linguistic context than for words in the other meta-category. In other words, even though the network cannot have generalized about what constitutes an adjective and what constitutes a noun — there is no generalization to be made, after all — it has made a distinction between the two meta-categories. The associations between linguistic inputs (the two linguistic context units) and linguistic outputs (the 36 word units) are sufficient to create two classes of words. We do not believe that the picture is this simple for word learning in children because there *are* semantic generalizations to be made concerning part-of-speech categories. In a more realistic setting, the straightforward learning demonstrated in this experiment might serve to *bootstrap* the learning of the relatively abstract semantic differences between the meta-categories. At any rate, the implication is that the patterns of errors made by children that implicate distinct noun and adjective categories could arise only from form-to-form associations.

5 General Discussion

We discuss the results of these experiments on two levels: First, we consider the network and why it learned as it did. Second, we consider the implications of the present results for our understanding of the origins of the noun advantage in children and for the nature of children's knowledge about the differences between nouns and adjectives.

[7] All of the mean activations are negative because for this experiment, the network learns to strongly inhibit all but the right response for each training instance, and for the test patterns, there is no "right" response from among the trained categories.

[8] For the analysis of variance, there were two factors, input linguistic context (noun or adjective) and average activation over output units by meta-category (noun or adjective). There was only one "subject" (network run) in this experiment, but there were 18 instances of each of the four combinations of the factors.

5.1 The Network

We defined the categories on which the network was trained in terms of the properties of the categories' extensions (volume, shape, overlap) and in terms of the presence of form-to-form associations between a linguistic context specifying the question asked of the network and the linguistic outputs that were possible answers to those questions. The network of course does not have direct access to any of these global properties of the learning task. It simply receives one category example at a time and for each modifies its weights in such a way that it has stored a composite record of the instances of each category. The network in no sense stores category boundaries or anything like the representations of category extensions we have used throughout this paper to visualize the differences between nouns and adjectives.

Why then do factors such as shape and volume and overlap matter as they do? Two factors are fundamental to the network's performance: (1) the distance between members of the same category relative to the distance between members of different categories and (2) the degree of redundancy in the input.

Each input the network receives represents a point in its multi-dimensional input space. Via the weights connecting the input layers and the hidden layer, the network maps this point in input space onto a point in multi-dimensional hidden-layer space. Inputs which are similar—close to each other in input space—will tend to map onto points which are close to each other in hidden-layer space. Points in hidden-layer space in turn are mapped onto points in category space via the weights connecting the hidden layer and the output layer. Before training, these mappings will be random, depending on the randomly generated initial weights. As training progresses, however, the weights in the network take on values which permit regions in input space to be associated roughly with the appropriate regions in category space. This involves some readjustment of the regions in hidden-layer space associated with inputs. In particular, inputs belonging to the same category will tend to map onto relatively compact regions in hidden-layer space (Harnad et al., 1991). Each time the network is trained on an instance of a category, the weights in the network are adjusted in such a way that that point in input space tends to get assigned to the region in output space associated with the category. When a test item is presented to the network, where it maps to in category space depends entirely on where it is in input space, in particular, how far it is from previously trained inputs. The input is implicitly *compared* to all of these inputs. Thus the network is an instance of an exemplar-based model of categorization (e.g., Nosofsky, 1986). In these models, it is the relative distance between an input and previously learned exemplars of the different categories which determines the behavior of the system.

If a given input is likely to be as close to a previous member of another category as it is to previously trained members of its own category, error will tend to be high, and learning will take longer, requiring more examples of each category. More examples result in a greater *density* of within-category examples which can compensate for the nearness to a test input of distracting examples of other categories.

Category volume and compactness both relate to this relative distance measure. As category volume increases and number of examples remains constant, density within categories decreases: the average distance between members of each category increases.

At the same time, the boundaries of different categories approach each other, so that for a given example of one category, the nearest distractor becomes nearer. Thus increasing volume leads to greater potential confusion between categories.

As category compactness decreases, we also see an increase in the average distance between members of a category. Consider two extreme cases, a set of parallel "hyperslabs" which extend across the full range of values on all dimensions but one and a set of evenly-spaced hyperspheres of the same volume as the hyperslabs. The average distance between members of the same category is greater for the hyperslabs because they may be arbitrarily far apart on all but one dimension. At the same time, the average distance between a member of one category and the nearest distractor in another category is smaller for the parallel hyperslabs, since the boundary of the nearest other category is found just across the narrow hyperslab-shaped gap separating the categories. Thus decreasing compactness, like increasing volume, means greater difficulty because of the potential confusion from examples of competing categories.

A further factor in category difficulty, though not as important in our results, is the degree of redundancy in the input. If more than one input unit conveys information about the category for an input pattern, then more network resources (weights) will be dedicated to representing the input-to-category mapping than would be the case if only one unit were relevant. In our experiments there is redundancy in all input patterns because of the use of thermometer encoding. On a given sensory dimension, all units to the "left" of a unit which is activated are redundant. However, in Experiment 4, some categories, namely, those with lexical dimension input, had the benefits of more redundancy than other categories. Recall that in this experiment, lexical dimensions were not required to categorize inputs, which on the basis of sensory input alone were unambiguous. Thus the redundant linguistic input gave the advantage to those categories for which it was available. Note, however, that while real adjective categories tend to be distinguished in part by lexical dimensions, they also tend to overlap with one another. When there is overlap, the lexical dimension is no longer redundant; rather, it, in combination with the sensory input, is necessary for determining the category of the input.

In sum, these two factors, (1) relative within- and between-category exemplar distances and (2) input redundancy, account for the results of our experiments. Interestingly, a third potential factor, the extent to which a particular input sensory dimension is relevant for a category, did not play a significant role. In Experiment 3, "adjective" categories were defined in such a way that a single dimension mattered much more than the other three. For "nouns", on the other hand, each sensory dimension was equally relevant. A learner with a propensity to selectively attend to particular sensory dimensions might find the adjectives easier. Relevance of a single dimension for a category conveys a disadvantage rather than an advantage for the network, and this result agrees with what we find for children.

5.2 The Noun Advantage in Children

These simulations were motivated by the goal of explaining two facts well-documented in the literature on children's early word learning: (1) the fact that nouns labeling concrete objects are learned faster than the dimensional adjectives that label the perceptible properties of those same objects and (2) the fact during the protracted course of learning dimensional adjectives, children seem to recognize that the dimensional adjectives comprise a class in that they confuse adjective meanings but do not confuse noun and adjective meanings.

The principal contribution of the present results is that they show that these two facts can emerge from the simple effects of similarity-based learning and thus that they do not demand an explanation in terms of prior conceptual knowledge of noun meanings or the differences between nouns and adjectives. The argument for pre-linguistic notions of the distinction between objects and their properties is often couched in terms of arguments that “one cannot get something from nothing” (see, for example, Markman, 1989). These simulations demonstrate that one can get a lot from ordinary effects of similarity and redundancy on learning — a noun advantage and proto-syntactic categories that in terms of their outward manifestations, that is, performance, look very much like the developmental trajectories of children learning common nouns and dimensional adjectives.

In the remainder of this paper, we discuss the further contributions and limitations of the research by addressing three questions: (1) Are the real nouns and adjectives that children learn like the idealized nouns and adjectives presented to the network? (2) Does the network instantiate a conceptual bias for noun-like meanings? and (3) Could the simple associative effects between linguistic inputs and linguistic outputs be the basis for a more conceptual understanding of the differences between nouns and adjectives?

5.2.1 Idealized versus Real Nouns

The nouns we sought to model are the common names for concrete objects, such nouns as *bottle*, *cup*, *mom*, *dog*, *bed*, and *cookie*. The idea that the extensions of such nouns are relatively small, compactly shaped, and non-overlapping has been offered repeatedly in the literature (e.g., Rosch, 1973; Markman, 1989, Medin & Ortony, 1989). Further, Rosch (1973b) has reported empirical evidence in support of these claims and Mervis (1987) has presented evidence that when mature usage does not fit this characterization, adults in their speech to children shift their use of nouns to keep the extensions compactly shaped and non-overlapping.

However, there are other findings in the literature that might be interpreted as showing that common names are not compact but are, rather, adjective-like in their emphasis of a single dimension. These findings concern the so-called shape bias in early noun learning (see Smith, 1995 for a review). Specifically, in novel word learning tasks, when a novel rigid object is named by a count noun, young children systematically generalize the newly learned name to novel objects by their shape ignoring such properties as color and texture. This shape bias in early noun learning fits well with Biederman's (1987) and Rosch's (1973) earlier results showing that adults recognize common objects principally by their shape. Do these results, then, suggest the real nouns learned by young children are not compact but are rather like adjectives in being constrained principally on one dimension. The evidence on the shape bias in early word learning is quite compelling, but we believe the inference from this fact about the non-compactness of noun extensions is wrong.

First, the complete evidence on the shape bias does not suggest exclusive attention to shape in children's early word learning. Rather, children attend to shape when objects are rigid but attend to texture and color when they are nonrigid, and even with rigid objects, children attend to shape and texture when the objects have eyes (see Smith, 1995 for a review). Thus, the total pattern suggests that while young children often emphasize shape in their early word generalizations, it is not to the complete exclusion of other properties. Rather, children attend to other properties and shift attention weights as a function of those properties. Second, although shape may be important to determining membership in a specific category, for real categories (as opposed to those used in artificial word learning tasks), other properties are also clearly predictive of category membership. Thus dogs do not just have a characteristic shape, they have characteristic colors, surface properties, and manners of movement. Thus, the extensions of the nouns that children encounter are relatively compact. Finally, shape is not a simple dimension but is composed of many sensory dimensions; constraints on the shape of instances will thus make for more compact category extensions than constraints on, for example, wetness or color. In sum, the extensions of the real nouns that children learn early may not be hypercubes in the space of all possible objects, but all that we know indicates that they are much more compact than dimensional adjectives.

5.2.2 A Conceptual Bias for Noun-like Meanings?

Our finding that the similarity relations within and among early-learned nouns and adjectives may create the noun advantage over adjectives contrasts with the suggestion that objects as opposed to their attributes are conceptually special (see, e.g., Gentner & Rattermann, 1991; Markman, 1989). However, one might argue that a three-layer network in which the hidden layer compresses the sensory input into one holistic representation is one instantiation of how a whole-object conceptual assumption might be implemented. From this argument, one might conclude that this network was "designed" to learn easily about categories in which all instances are globally similar to each other (and thus compact and small). Is this not, in a sense, a built-in bias for noun-like categories?

By one interpretation of this question, the answer is a clear "yes." The proposal that noun categories are more "natural" than adjective categories and the proposal that young children "assume" that words name things and not their properties are currently unspecified in terms of the processes through which the naturalness of nouns or children's assumptions might be realized. This network offers one implementation of these ideas; it shows just how nouns might be more "natural" and why very young children seem to interpret novel words as having nominal meanings. Thus the results of these simulations might be properly viewed as supportive of and an extension of proposals about young children's early biases and assumptions about word meanings.

But there is a second interpretation of the question of whether a noun-advantage was built into the network that demands a clear "no." It is true that representations at our hidden layer holistically combine the input from the separate sensory dimensions. Connectionist networks do not have to do this. For example, Kruschke's ALCOVE network

(1992) utilizes distinct dimension weights such that the network retains information about distinct attributes at the hidden layer level. Given these differences, one might expect that Kruschke's network would learn adjective categories more easily than the present one. This may be. However, the conclusion that our network is structured to make the learning of adjectives hard is not warranted. It is not warranted because our network learns single-dimension adjective categories easily, trivially fast when there is only one relevant dimension and no overlapping categories. That is, when we presented our network with the same kind of task that ALCOVE has been presented with — classifying all inputs into two-mutually exclusive categories, each constrained by variation on the same dimension (what might correspond to learning the categories BLACK versus WHITE) — the network rapidly (in less than 500 trials) converged to a set of attention weights that emphasize the solely relevant input dimension. In brief, it is not hard for this network to learn adjective-like categories.

However, it is hard for this network to learn adjective-like categories when it must, like young children, simultaneously learn noun-like categories that require attention to many dimensions and multiple overlapping adjective categories that each require attention to different dimensions. We conjecture that a similar difficulty might hold even for models like ALCOVE when the task is the simultaneous learning of multiple overlapping noun-like and adjective-like categories.

In sum, the ease with which the present network learns adjective categories on one dimension when that is all that it has to learn indicates that the noun advantage is not solely the product of the compression of dimensional information at the hidden layer. Rather, the noun advantage appears to be a product of similarity-based learning and the task of learning overlapping categories. Given this kind of learning device and this set of tasks to be learned, noun-like meanings are primary.

5.2.3 Learning the Categories "Noun" and "Adjective"

The general acceptance of the idea that young children distinguish between nouns as name for things and adjectives as labels for the properties of things is based on the facts of the noun advantage and the pattern of within-adjective confusions that characterize children's slow and errorful acquisition of dimensional terms. The simple network that we have studied distinguishes nouns and adjectives in the very same way that young children do: It learns noun categories faster than adjective categories and during the protracted course of learning adjectives, its errors consist of confusing one adjective with another and not of confusing an adjective with a noun. Thus, our network, like children, "knows" that nouns and adjectives are different.

The processes that make up this "knowing" by the network, however, are not of the kind one usually thinks of as knowledge about the different meanings of nouns and adjectives. All that appears to be known when the network in Experiment 1 makes these errors is (1) the noun categories, (2) the linguistic context that specifies nominal outputs, and (3) the fact that the linguistic contexts that specify adjective categories are not associated with nominal outputs. Apparently this is enough to get a behavioral distinction between nouns and adjectives in the course of learning. The network knows about nouns and ipso facto "knows" a class of items that are not nouns. The results remind us that the internal processes that comprise some external pattern of behavior may be simpler than the external behavior itself.

The present network is a very simple model that leaves out much of what children probably do know about nouns and adjectives. While our approach is unabashedly grounded in the semantics of nouns and adjectives, we have tried to show in Experiment 6 how purely form-to-form learning can also play a role in the emergence of syntactic categories. In fact learners appear to have access to a wealth of purely formal information to guide them in learning, and a large body of recent work has focused on the extent to which linguistic categories can be learned on the basis of distributional information (Elman, 1990; Finch & Chater, 1992) or the formal properties of the words themselves (Kelly, 1992). As in the present model, these approaches are statistical and associative; in fact, many are implemented in the form of connectionist networks. However, given the nature of the inputs and the restricted architecture, the present network obviously cannot make use of the phonology of the words or of the detailed pattern of co-occurrences with other words. We have only sought to demonstrate that syntactic categories can begin to emerge as a kind of side-effect as the system learns to label objects. Note what distinguishes these syntactic categories from the conventional ones, however; because they are directly associated with objects and their properties, they have a semantic force. Although this may not be what is usually meant by theorists who write about children's understanding of the differences between nouns and adjectives, this could be pretty much what the differences amount to in the early stages of acquisition.

6 Conclusion

What is the difference between common nouns and dimensional adjectives that allows children to acquire nouns more rapidly than adjectives? We could distinguish the two categories in purely syntactic terms, with respect to the other categories with which they co-occur. We could also distinguish them in terms of their function, as Markman (1989) does; we carve up the world in useful ways with nouns and then resort to adjectives when we need to distinguish members of the same nominal category along arbitrary dimensions. But underneath all this might be a more mundane distinction, one based on the tendencies of nouns and adjectives to delineate particular sorts of regions in multi-dimensional perceptual space. Unlike the first two kinds of distinctions, this third is one which is directly available to a relatively simple learning device, as we have shown in this paper. Of course a child must eventually learn about more abstract functions and about syntactic categories much richer than those examined here, but the distinction based on the most accessible sort of information could provide a foundation for this later learning.

References

Aslin, R. N. & Smith, L. B. (1988). Perceptual development. *Annual Review of Psychology*, *39*, 631–682.

Au, T. K. & Laframboise, D. E. (1990). Acquiring color names via linguistic contrast: the influence of contrasting terms. *Child Development*, *61*, 1808–1823.

Au, T. K. & Markman, E. M. (1987). Acquiring word meaning via linguistic contrast. *Cognitive Development*, *2*, 217–236.

Backscheider, A. G. & Shatz, M. (1993). Children's acquisition of the lexical domain of color. In *What We Think, What We Mean, and How We Say it: Papers from the Parasession on the Correspondence of Conceptual, Semantic, and Grammatical Representations*, Vol. 29, pp. 11–21. Chicago Linguistics Society.

Biedermann, I. (1985). Human image understanding. *Computer Vision, Graphics, and Image Processing*, *32*, 29–73.

Callanan, M. A. (1990). Parents' description of objects: potential data for children's inferences about category principles. *Cognitive Development*, *5*, 101–122.

Carey, S. (1978). The child as word learner. In Halle, M., Bresnan, J., & Miller, G. (Eds.), *Linguistic Theory and Psychological Reality*. MIT Press, Cambridge, MA.

Carey, S. (1982). Semantic development: the state of the art. In Wanner, G. & Gleitman, L. R. (Eds.), *Language Acquisition: The State of the Art*, pp. 139–195. Cambridge University Press, Cambridge, MA.

Carey, S. (1994). Does learning a language require the child to reconceptualize the world?. *Lingua*, *92*, 143–167.

Clark, E. V. (1973). What's in a word: on the child's acquisition of semantics in his first language. In Moore, T. E. (Ed.), *Cognitive Development and the Acquisition of Language*. Academic Press, New York.

Dromi, E. (1987). *Early Lexical Devlopment*. Cambridge University Press, New York.

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

Finch, S. & Chater, N. (1992). A hybrid approach to the automatic learning of linguistic categories. In Aleksander, I. & Taylor, J. (Eds.), *Artificial Neural Networks, 2* Amsterdam. ICANN, Elsevier.

Gasser, M. & Smith, L. B. (1991). The development of the notion of sameness: a connectionist model. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, pp. 719–723 Hillsdale, NJ. Lawrence Erlbaum.

Gentner, D. & Rattermann, M. J. (1991). Language and the career of similarity. In Gelman, S. A. & Byrnes, J. P. (Eds.), *Perspectives on Language and Thought: Interrelations in Development*, pp. 225–277. Cambridge University Press, Cambridge.

Gentner, D. (1978). On relational meaning: the acquisition of verb meaning. *Child Development*, *48*, 988–998.

Gershkoff-Stowe, L. & Smith, L. B. (1996). Naming errors and emerging retrieval processes: a study of early changes in lexical processing. Under review.

Harnad, S., Hanson, S. J., & Lubin, J. (1991). Categorical perception and the evolution of unsupervised learning in neural nets. AAAI Spring Symposium on Symbol Grounding: Problem and Practice, Stanford, CA.

Huttenlocher, J. (1974). The origins of language comprehension. In Solso, R. (Ed.), *Theories in Cognitive Psychology*. Lawrence Erlbaum, Potomac, MD.

Imai, M. & Gentner, D. (1993). What we think, what we mean, and how we say it: papers from the parasession on the correspondence of conceptual, semantic, and grammatical representations. In *Proceedings of the Chicago Linguistic Society*, Vol. 29. Chicago Linguistics Society.

Jackson-Maldonado, D., Thal, D., Marchman, V., Bates, E., & Gutierrez-Clellen, V. (1993). Early lexical development in Spanish-speaking infants and toddlers. *Journal of Child Language*, *20*, 523–549.

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.

Kelly, M. H. (1992). Using sound to solve syntactic problems: the role of phonology in grammatical category assignments. *Psychological Review*, 349–364.

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44.

Macnamara, J. (1982). *Names for Things: A Study of Human Learning.* MIT Press, Cambridge, MA.

Maratsos, M. (1988). Crosslinguistic analysis, universals, and language acquisition. In Kessel, F. (Ed.), *The Development of Language and Language Researchers: Essays in Honor of Roger Brown*, pp. 121–152. Lawrence Erlbaum, Hillsdale, NJ.

Markman, E. M. & Hutchinson, J. E. (1984). Children's sensitivity to constraints on word meaning: taxonomic vs. thematic relations. *Cognitive Psychology*, *16*, 1–27.

Markman, E. M. (1989). *Categorization and Naming in Children: Problems of Induction.* MIT Press, Cambridge, MA.

Medin, D. & Ortony, A. (1989). Psychological essentialism. In Vosniadou, S. & Ortony, A. (Eds.), *Similarity and Analogical Reasoning*, pp. 179–195. Cambridge University Press, New York.

Mervis, C. B., Mervis, C. A., Johnson, K. E., & Bertand, J. (1992). Studying early lexical development: the value of the systematic diary method. In Rovee-Collier, C. & Lippsitt, L. (Eds.), *Advances in Infancy Research, 7*, pp. 291–379. Ablex, Norwood, NJ.

Mervis, C. B. (1987). Child-basic object categories and lexical development. In Neisser, U. (Ed.), *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*. Cambridge University Press, Cambridge.

Naigles, L. G. & Gelman, S. (1995). Overextensions in comprehension and production revisited: preferential-looking in a study of *dog*, *cat*, and *cow*. *Journal of Child Language*, *22*, 19–46.

Nelson, K. (1973). *Structure and Strategy in Learning to Talk*. No. 149 in Monographs of the Society for Research in Child Development. University of Chicago Press, Chicago.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Plunkett, K., Sinha, C., Møller, M. F., & Strandsby, O. (1992). Symbol grounding or the emergence of symbols? vocabulary growth in children and a connectionist net. *Connection Science*, *4*, 293–312.

Rosch, E. (1973a). Natural categories. *Cognitive Psychology*, *7*, 573–605.

Rosch, E. (1973b). On the internal structure of perceptual and semantic categories. In Moore, T. E. (Ed.), *Cognitive Development and the Acquisition of Language*, pp. 111–144. Academic Press, New York.

Schyns, P. G. (1992). A modular neural network model of concept acquisition. *Cognitive Science*, *15*, 461–508.

Smith, L. B. & Sera, M. (1992). A developmental analysis of the polar structure of dimensions. *Cognitive Psychology*, *24*, 99–142.

Smith, L. B., Jones, S., & Landau, B. (1992). Count nouns, adjectives, and perceptual properties in children's novel word interpretations. *Developmental Psychology*, *28*, 273–286.

Smith, L. B. (1984). Young children's understanding of attributes and dimensions: a comparison of conceptual and linguistic measures. *Child Development*, *55*, 363–380.

Smith, L. B. (1993). The concept of same. In Reese, H. W. (Ed.), *Advances in Child Development and Behavior, Vol. 24*. Academic Press, New York.

Smith, L. B. (1995). Self-organizing processes in learning to learn words: development is not induction. In *Basic and Applied Perspectives on Learning, Cognition, and Development*, Vol. 28 of *The Minnesota Symposium on Child Psychology*, pp. 1–32. Lawrence Erlbaum Associates, Marwah, NJ.

Snow, C. E. (1977). The development of conversation between mothers and babies. *Journal of Child Language*, *4*, 1–22.

Taylor, M. & Gelman, S. A. (1988). Adjectives and nouns: children's strategies for learning new words. *Child Development*, *59*, 411–419.

Waxman, S. R. (1994). The development of an appreciation of specific linkages between linguistic and conceptual organization. *Lingua*, *92*, 229–250.

Wood, D. J. (1980). Teaching the young child: some relationships between social interaction, language, and thought. In Olson, D. R. (Ed.), *The Social Foundations of Language and Thought*. Norton, New York.

Woodward, A. L., Markman, E., & Fitzsimmons, C. M. (1994). Rapid word learning. *Developmental Psychology*, *30*, 553–566.