2010

Semi-Supervised Learning for Connectionist Networks

Rebecca Robare
The Graduate Center, City University of New York

SEMI-SUPERVISED LEARNING IN CONNECTIONIST NETWORKS

by

Rebecca J. Robare

A dissertation submitted to the Graduate Faculty in Psychology in partial fulfillment of the requirements for the degree of Doctor of Philosophy
The City University of New York
2010

This manuscript has been read and accepted for the Graduate Faculty in Psychology in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

8/27/08 Robert D. Melara

8/27/08 Maureen O'Connor

Vivien C. Tartter
James B. Marshall
Martin Chorodow
Heng Ji
Supervisory Committee

Abstract

SEMI-SUPERVISED LEARNING IN CONNECTIONIST NETWORKS

by

Rebecca J. Robare

Advisor: Robert Melara

At the computational level, language is often assumed to require both supervised and unsupervised learning. Although we have a certain understanding of these computational processes both biologically and behaviorally, our understanding of the environmental conditions under which language learning takes place falls short. I examine the semi-supervised learning paradigm as the most accurate computational description of the environmental conditions of lexical acquisition during language development. This paradigm is assessed for task learning and generalization and I argue that its real ecological validity and occasional improvements in performance over supervised learning make it an ideal candidate for modeling of language acquisition and other learning problems.

Table of Contents

SEMI-SUPERVISED LEARNING FOR CONNECTIONIST NETWORKS

Tables

Figures

SEMI-SUPERVISED LEARNING FOR CONNECTIONIST NETWORKS

Introduction

Imagine that you are an infant. The world is full of shapes, colors, and sounds. Slowly, you learn to identify these shapes as objects, to separate one word from another, to understand another's language and finally, to produce it. About all of these processes, we know very little. We understand them to be interdependent. Yet for the sake of simplicity and control, we study one at a time. The difficulty in accessing the cognitive processes of a prelingual child is daunting, though decades of clever research have given us some insight. The computational models of the past 25 years have been useful as well. They allow us to explicate theories of early learning and cognition and make experimental hypotheses more detailed and precise. The usefulness of these models is based in the assumptions they make about the workings of the brain: distributed representations, recurrent networks, etc. However, the utility of these models has been severely curtailed by their strictly dichotomous views of learning.

An infant, in this world of shapes and colors and sounds, must learn to detect the correlations among features of that world. Things with eyes move of their own intention. Faces with delicate features (relative to those with less fine features) make higher-pitched sounds. In a computational framework, this kind of learning and those similar to it are modeled through a paradigm called *unsupervised learning*.

There are other things to learn in this world as well. Sometimes an object is looked at or picked up at the same time as a series of stressed sounds are heard. Or the sounds are heard first, or are subsequent to the handling of the object. As the infant begins to interact with the world, there are sequences of sound that require a response: “What does the cat say? Can you find the RED one?” Computationally, these associative and predictive abilities are modeled through the paradigm called *supervised learning*.

The difference between unsupervised and supervised learning is that unsupervised learning is used when items of information are concurrent in the environment, whereas supervised learning is used when items of information are separated in time or space. These categories are not entirely separable and different authors may use them to refer to the same example of learning. However, this definition will be suitable for the current work. (For more information about supervised and unsupervised learning, see Appendix A.)

It is acknowledged that both kinds of learning are needed to recognize and produce fluent language (Bloom, 2000; Carpenter, Nagell, & Thomasello, 1998*, inter alia*). But our models, with very few exceptions, separate these paradigms, implementing one or the other exclusively. Those few exceptions, most notably the SUSTAIN model by Love, Medin, and Gureckis (2004), do so by alternating between paradigms rather than by combining them.

If our intent in building networks is to capture the problems of human language acquisition from available data, the current system is inadequate. Both kinds of learning processes exist in the language acquisition of real infants. To systematically exclude one or the other from a model of learning, or to suggest that the infant's brain sometimes decides to learn from one kind of data and sometimes from the other, is to blind oneself to the possibility of a more useful model, one that captures several forms of data in a single environment (just like the world) and allows the brain the flexibility to learn from such a paradigm without an artificial switching process. Here we will explore a paradigm known as *semi-supervised learning*, the use of which may rewrite our ideas about the way we model learning. In Chapter One, current knowledge about lexical acquisition will briefly be discussed. In Chapter Two, we will explore semi-supervised learning, and how it can be useful to cognitive modelers. The subsequent chapters will describe the results of specific simulations that use semi-supervised learning to emulate the processes involved in human lexical acquisition, and integrate semi-supervised modeling with the current state of our knowledge about lexical acquisition in a proposal for modeling lexical acquisition with the semi-supervised paradigm.

Chapter 1

Lexical Acquisition

The problem of acquiring mapping

The process of learning a first language (L1) is complex and multifaceted. The learner must acquire a lexicon (vocabulary), and apply the newly learned words to the correct real-world objects ideas, and circumstances. The rules of syntax must be learned, deduced, or activated, and applied correctly despite a relative paucity of example and total lack of counterexample. Language is also highly social, and research increasingly indicates that social behaviors, such as gaze following, are necessary for proper language acquisition. Cultural rules and norms also influence the speech of the language learner – who may be spoken to, and in what manner, and under what circumstances. Although these domains depend on each other to a great extent (Snedeker & Gleitman, 2004), researchers often study them separately, as a way of making such a complex cognitive process more manageable. For example, lexical acquisition, the process of learning and correctly using a vocabulary, is complex enough even without consideration of syntax and socialization.

To acquire the meaning of a word, the learner needs three things: access to the form of the word (i.e., its phonological token), access to the concept (i.e., what the word means), and access to the mapping between the form and the concept (i.e., some link between the two must exist) (Bloom, 2000). Here we will limit discussion to concept and map between form and concept. First of all, what is meaning? For the purposes of this discussion, word meaning is related to categorization, the ability to determine which *things* are identified by that word. This is most useful despite the limits to this idea, which largely have to do with philosophical considerations of reference, as opposed to the psychological consideration of mental representation (Bloom, 2000). In other words, what a *dog* is in the world is less important than what a native speaker understands a *dog* to be. This is not to suggest that philosophical considerations of meaning are irrelevant, but merely that for the current discussion, the operational definition of *meaning* will apply to the learner's mental model of the world, and not explore the potentially complex relationship between mental representation and reference. For present purposes, it is enough to assume that the mental model the learner creates is accurate, or eventually accurate, to the world as the learner experiences (including what can be inferred or generalized from that experience.) What kind of information is contained in the meaning of a word? Rogers and McClelland (2004) usefully describe semantic information as the information about an object that is not available directly from the percept of an object (excluding here verbs and function words for simplicity's sake). With all the above taken in combination, then, *meaning* is the information, not limited to that accessible from a percept, that is entailed by correct understanding of what things in the word a speaker intends to reference, "intent to reference" here being a psychological representation on the part of the utterer, and not synonymous with the more philosophical "reference" above. This is, of course, a gloss; volumes can be (and have been) written on what, exactly, meaning means. However, for the present purposes, it will suffice to a simple definition of meaning that can be operationalized in experimental and observational studies on word learning.

Objects and object names

Having an acceptable definition of meaning, we turn to the question of how the mapping between this meaning and a word form is made. To acquire a lexicon of object names, the first language (L1) learner must discover the association between a word and an object in the world, and apply that word correctly to all examples of that object, while excluding other objects from that category. Often in research, a simplified view of this process is taken, one that excludes verbs and function words, and considers the process of associating a word and its meaning as separate from the process of segmenting individual words from the speech stream. The intention of this simplification is not to indicate a belief in the inherent separability of these processes. Nor is it meant to imply that all children among all cultures learn words in the same way. Rather, it is merely to take a highly complex process and constrain it in such a way that it can be usefully studied. A complete theory of lexical acquisition must of course overcome these limitations and simplifications, and rejoin the acquisition of a lexicon to the related processes of learning syntax and pragmatics, but as the contemporary understanding of lexical acquisition cannot hope to be complete, the present discussion will, except where stated, limit itself to the acquisition of common nouns among Western children. In this population, the largest proportion of the earliestlearned 100 or so words tend to be common nouns (Brown, 1957; Macnamara, 1982; 1984; all cited by Bloom, 2000; Snedeker & Gleitman, 2004).

Even in this unrealistically simplified manner, multiple cognitive processes are involved in the problem of lexical acquisition. Psychologist Paul Bloom suggests (2000) that many of the theorized constraints that work solely for the purpose of lexical acquisition are unlikely to exist (Bloom, 2000):

> "...the whole-object bias, the taxonomic bias, and the mutual-exclusivity bias (Markman, 1989), the noun-category linkage (Waxman, 1994), the shape bias (Landau, Smith, and Jones, 1988), the principles of contrast and conventionality (Clark, 1993), and the principles of reference, extendability, object scope, categorical scope, and object name-nameless category (Golinkoff, Mervis, & Hirsch-Pasek, 1994)." (p. 10)

Rather, more general properties of learning should be sufficient to explain how word meanings are learned. In general, a theory basing lexical acquisition on general cognitive abilities is more parsimonious than one that does not, and is to be preferred if the phenomenon under study can be so explained. However, the utility of some of these biases, and the evidence for their existence either as special properties of a language-acquisition system or as general principles of cognition, make them worth further consideration. Specifically, the whole-object bias (p. 5) and the shape bias (p. 7) will be discussed below, and the reader is referred to Bloom (2000) for discussion on the other constraints he mentions.

Mappings between a word and an object in the environment are arbitrary. That is, often the form of the word cannot be predicted on the basis of the perceptual properties of the object. There is statistical learning from the systematic pairing of words and objects, but there is disagreement as to whether this is sufficient for learning the earliest words (Gogate & Bahrick, 2001; Snedeker & Gleitman, 2004; Werker et al., 1998). The ability to easily make arbitrary mappings is seen in infants of 14 months, but not in those of 9 months (Werker et al, 1998, cited by Woodward, 2004), but infants as young as 7 months can make some mappings if enough stimulus redundancy is available in repeated pairings of word and object (Gogate & Bahrick, 2001, cited by Woodward, 2004). It has also been found that statistical learning is used by infants at 8 months to learn to recognize word boundaries in the speech stream (Saffran, Aslin, & Newport, 1996), which fits into the time frame for statistical learning of word meanings found by Gogate and Bahrick (2001). This suggests either that the methods of Werker and colleagues (1998) may not be sensitive enough to detect statistical learning in the youngest learners, or that statistical learning is an ability that begins early, and improves over the course of development. In either case, statistical learning is a powerful mechanism that cannot be discounted in the study of lexical acquisition.

Children are biased, in certain contexts, to interpreting novel words as object names. Object names proportionally make up more of children's early vocabularies than adults' vocabularies (Bloom 2000, following Brown, 1957; Macnamara, 1982; Pinker, 1984). Bloom considers the importance of object names to have been overstated in the literature, but overstated or not, given that object names are a crucial component of early word learning, it is important to clearly define what an object is, in order to understand what it is that children are learning the names of, and why this particular aspect of lexical acquisition is so important.

Bloom's (2000) definition of *object* closely follows Spelke's (1994; Spelke, Phillips, & Woodward, 1995) in that objects have several properties, the most important being cohesion. Objects move as single units; if someone pulls on part of an object the rest of the object also moves (Markman, 1992, cited by Echols & Marti, 2004; Pinker, 1997). Whereas other object properties appear to describe how objects behave, cohesion seems to define what it is to be an object – the other properties may occasionally be violated, or may be accidental properties of objects, but cohesion always obtains and is never accidental. Carey and Spelke (1994, cited by Bloom, 2000) argue that the knowledge that objects are cohesive, as well as continuous (all the parts of a single object connect), solid (objects cannot pass through each other), and contacting, is thought by some researchers to be innate, part of the cognitive biases with which babies come into the world (Carey & Spelke, 1994, cited by Bloom, 2000). However, that objects move as cohesive, continuous units may also be learned through experience of objects early in life, if the innate bias is to discover the properties of objects (Vivien Tartter, 2010, personal communication). Whether innate or not, the whole-object bias seems to be an important feature of lexical learning.

The primacy of objects is supported by the relative infrequency of part naming in early word learning. The part names that do appear are primarily body parts (Andersen, 1975, Smith et al., 2002, cited by Landau, 2004), which may indicate access to non-object concepts through a kinesthetic, rather than a visual mapping, in that infants are learning to move the separate parts of their own bodies. Visual experience may also play a role, however, in observation of the separable movement vectors of body parts when the whole body is in motion. The combination of whole-object movement vectors and part-level movement vectors may be unusual in wholes vs. parts, and therefore support body-part names prior to the names of other parts, in-as-much as body parts are, in some ways, like whole objects themselves.

It is not known whether the bias that results in earlier rapid learning of object names compared to other classes of words is a general cognitive bias or a mechanism specific to word learning (Baillargeon, 1993, cited by Echols & Marti, 2004; Bloom, 2000; Hall, 1996 and Shipley & Shepperson, 1990, cited by Bloom, 2000; Spelke, 1991, cited by Echols & Marti, 2004). In a study designed to mimic the characteristics of infant word learning in college students, 45% of nouns but only 15% of verbs could be inferred correctly from observation of a caregiver's behavior (Snedeker & Gleitman, 2004). This is said to suggest the possibility that the oft-observed “bias” is merely ease of induction from observation as a function of concreteness or imageability (Gillette et al., 1999, cited by Snedeker & Gleitman, 2004). Even if this is correct, however, it does not diminish the relative importance of nouns compared to verbs in early word learning (at least for the aforementioned Western children). The view that lexical acquisition is driven not by word learning modules but general characteristics of cognition is a hallmark of Bloom's presentation of the topic and, from a theoretical standpoint, is not only more parsimonious than theories that require a number of specialized mechanisms but lends itself well to neural-network models for explication and the development of hypotheses for empirical testing.

That object names seem to be special in the word learning of Western children is unsurprising, as Western parents often name objects for their children, using pointing and eye gaze to direct their children's attention to the objects and their names. The environments of Western children, therefore, are rich in salient object names isolated from other speech. This method of interacting with infants is culturally specific, however, and does not hold for the speakers of all languages. Yet the object-name bias appears cross-culturally and cross-linguistically (Clark, 1993, cited by Bloom, 2000), in those non-Western languages that have been studied. Bloom (2000) theorizes that all cultures may use some kind of referential noun phrase in isolation, including those that do not treat object names as a special class. This seems very hypothetical, however, and it is difficult to determine whether counterexamples exist. It is not known whether such isolated noun phrases are necessary for language learning, or whether the status of such phrases in English and other Western languages are an unimportant cultural artifact. This is not to say that other classes of words are unimportant in lexical acquisition (for they are) or that non-Western languages are learned differently (for they well may be). Rather, the available evidence stresses the relative importance of nouns in early vocabulary across languages, and provides justification for the current discussion, which largely omits verbs,function words, adjectives, and adverbs in favor of the relatively more-studied and more easily studied nouns.

Category formation and reorganization: Generalizing object names

Lexical acquisition is often related to category learning, the critical question being whether young children who are learning words are simultaneously learning about the categories to which they refer, or whether they already know the relevant categories and are merely learning the mapping between the category and the lexical symbol (Bloom, 2000; Mandler, 1996). This raises other questions, about how young children form categories, how these categories are reorganized during development, and how learning word meanings impacts this reorganization (Bloom, 2000; Rogers & McClelland, 2004).

The first categories that seem to be learned are those at what is called the *basic level* (Bloom, 2000; Keil, 1989; Rogers & McClelland, 2004; Rosch & Mervis, 1975; Rosch et al., 1976). This is usually an intermediate level of description that is thought to be the most informative for distinguishing members of that category from members of other categories. For example, *dog* and *cat* are basic level categories, because it is important to distinguish dogs from cats in everyday life; young word learners are likely to encounter both. (Bloom, 2000; Keil, 1989; Rogers & McClelland, 2004; Rosch & Mervis, 1975; Rosch et al., 1976).

The above briefly describes the nature of basic categories, but not how individual items are classified into one category or another. Some researchers have appealed to the importance of shape, which appears to be a primary feature in early categorization. Children will classify items into the same category if they have a similar shape by 4 months of age, and if they have a similar texture by 12 months, (Behl-Chadha, 1996, cited by Landau, 2004; Smith, Jones, & Landau, 1992). After the child has learned about 50 words, the shape bias alters subtly, and is found in naming but not in non-naming contexts (Smith et al., 2002, cited by Landau, 2004; Smith, Jones, & Landau, 1992). As Bloom (2000) rightfully points out, knowing that shape is important in categorization does not reveal whether there is an innate cognitive bias to group similarly-shaped items together (the brute shape theory), or whether shape merely serves as a cue to category membership insofar as shape results from the nature of the item (this kind of essentialism is part of the "theory theory", the hypothesis that category discriminations are based on causal theories about why objects go together; Murphy & Medin, 1985). (Note that conclusions about the importance of shape are limited to studies in the Western cultural context.)

There is not a lot of evidence for either position, and while some researchers have argued for the superiority of the shape-as-cue theory because it fits a general cognitive theory, whereas the brute shape theory applies only to words (Bloom, 1996, and Keil, 1994, cited by Bloom, 2000), this does not necessarily follow. There is nothing in the development of cognition to inherently preclude shape as an important general factor – certainly shape is crucial in some obvious areas, such as vision – and even an innate bias can be weak, rather than strong, and subject to violation in the presence of other cues. The results of Landau, Smith, and Jones (1988) suggest that shape is important innately, perhaps in the sense of a perceptual bias, but also becomes more important through the learner's experience with named objects. The entire argument may be moot, however, as more recent research has indicated that shape is not found to be a basis of overgeneralization of category names more than other category bases, and that a number of kinds of perceptual properties can serve as a basis for category judgments (Gelman & Markman, 1987, cited by Bloom, 2000). In sum, categorization based on shape versus categorization based on other properties, such as function, depends on context cues (Bloom, 2000).

Categories are not only learned early in development, but they are reorganized as other learning takes place. Category boundaries may be established by the learning of words for those categories

(Bowerman, 1996, Gopnik & Meltzoff, 1987, 1997, Waxman & Markow, 1995, Waxman & Thompson, 1998, all cited by Bloom, 2000; Landau & Shipley, 2000, cited by Landau, 2004). However, this leaves unanswered the question of whether such categories are induced, or whether it is merely the word-category mappings that are so learned, that the category already exists and it is now merely acquiring a label (Bloom, 2000; Mandler, 1996).

The importance of caregiver behavior and other environmental factors in language learning is being increasingly acknowledged by researchers. Recently, the researchers of the Human Speechome Project (Roy, 2006) have attempted to record, to the greatest extent possible, the entire linguistic environment of an L1 learner. The house of the lead researcher, at the time expecting his first child, was wired for audio and video recording in every room. With appropriate provisions for personal privacy, all waking hours of the house's three residents were recorded, thereby providing the most complete record to date of the linguistic environment and slowly growing comprehension and production of the child. Though the generality of any findings from this project will be limited because of the single subject (raised in an English-speaking household of the upper-middle socioeconomic class in the northeastern United States), the ambitious attempt to discover the total environment of a language learner is valuable. The massive data generated by this project are still being analyzed, but the level of attention devoted to the language environment by this project's researchers showcases the growing understanding of the importance of environmental factors in language acquisition. The level of data gathering enabled by contemporary technology dwarfs previous corpora such as CHILDES, the Child Language Data Exchange System,[1] (MacWhinney & Snow, 1985) and makes it clear that such corpora are inadequate compendia of the language environments of individual children.

The “theory theory” (see p. 9 for a brief description) seems supported by the observation that conceptual knowledge is reorganized over the course of development (Carey, 1985, Keil, 1985;

[1] CHILDES is a database of child-directed and child-uttered language ...

Mandler, 1997, cited by Rogers & McClelland, 2004). This pattern of development is consistent with a hierarchical taxonomy of semantic knowledge, but again, this model contains no mechanism by which change in the hierarchy can occur (Rogers & McClelland, 2004). Just as in Quillian's (Collins & Quillian, 1969; Collins & Loftus, 1975) original model (see p. 16), change is assumed to be a function of experience or learning, but the change itself is undefined. In other words, where change in a connectionist model relies on a mathematical process of adjustment of synapse-like connection weights in response to specific environmental occurrences, the nature of change in the hierarchical models under review is not precisely determined to mean anything. In addition, the sorts of causal knowledge emphasized by the “theory theory” demonstrably influence judgments of category membership, but are not represented in a hierarchical model, nor does the hierarchy account for why causal knowledge should be special, and treated differently than other forms of knowledge.

“Theory theorists” argue convincingly that early learning is constrained, but do not provide the mechanism by which these constraints operate (Rogers & McClelland, 2004). This is not a failing of the “theory theory” precisely, but the point that mechanisms are important is well taken. A mechanistic account of semantic knowledge should either explain how the “theory theory” works, or provide an alternative account of constraints on the semantic-learning system. Rogers and McClelland mention Sloman and Rips's (1998) as category theorists who have tried to do this, but did not review their article.

Rogers and McClelland (2004) believe that category approaches are too narrow to capture the subtlety of a quasi-regular domain such as semantics. A connectionist approach is justified because of other things we know about the brain, such as its biological structure, and because of the parsimony offered by an account that attempts to use general properties of cognition to explore a specific cognitive domain rather than relying on specialized modules, but to base an argument on the insufficiencies of a mechanism that has not yet been explicated is, perhaps, assuming too much. However, it may well be the case that the emergent properties of a connectionist network will account for the subtle features of the human semantic system.

Environmental influences

At the environmental level, language learning is increasingly being shown to be a highly social process. According to some researchers, social cues to word meaning begin to be used between 10 and 19 months of age, supplanting perceptual salience, which is a more dominant cue in infants aged 10 to 12 months. The changes appear to be gradual; 12-month-olds are better at making inferences from social cues than 10-month-olds, though they are not yet improved enough to learn mappings from such cues (Hollich, Hirsh-Pasek, & Golinkoff, 2000, cited by Hirsh-Pasek et al., 2004). At about the same age, 13 to 25 months, children are increasingly willing to extend words to perceptually dissimilar exemplars, again showing the decreasing importance of perceptual salience (and, perhaps, the shape bias?) at this time.

An excellent demonstration of the importance of social-environmental cues is seen in studies of the relationship between caregiver-child joint attention and lexical acquisition. In one longitudinal study (Carpenter, Nagell, & Tomasello, 1998), joint engagement of attention was measured in time duration and number of episodes of looking over a ten-minute observational period. Based on the amount of time the infants spent looking at the same objects as their caregivers in the experimental setting, infants were divided into four groups: early, middle, and late (referring to the month at which joint engagement was observed), and never. Mothers made utterances of three kinds: *follow*, in which they named an object to which the child was already attending; *lead*, in which they used an object name to draw the child's attention to an object; and *other*, which were all other kinds of utterances. *Follow* and *lead* utterances did not differ in frequency, although Nelson (1988) had found that adults infer the object of children's attention in naming, and not the other way around. Vocabulary comprehension and production were measured with the Communicative Development Inventory (CDI), which surveys parents on the words their children understand and produce at various ages (Dale & Fenson, 1996; Fenson et al., 1993), at several intervals. There were significant positive correlations of joint attention at 11, 12, and 13 months of age with word comprehension between 9 and 15 months, echoing the finding that social cues to word meaning begin to be used between 10 to 19 months (Hollich, Hirsh-Pasek, & Golinkoff, 2000, cited by Hirsh-Pasek et al., 2004). The early attention group comprehended significantly more words, and had a significantly greater rate of vocabulary increase, than the late attention group (Carpenter et al., 1998). Lexical acquisition is thus shown to be highly related to the caregiver's behavior, which can be considered part of the language learner's environment.

It is unlikely, however, that words used during periods of joint attention account for all word learning. Joint attention is relatively infrequent (Hoff & Naigles, 2002, cited by Woodward, 2004), so it is likely that word learners also use non-joint-attention contexts to interpret words (Akhtar, Jipson, & Callanan, 2001, cited by Woodward, 2004). (One example of this may be mutual attention, in which the mother and child pay attention to each other.) Word learning when caregiver and child do not share attention also fits cross-linguistic accounts of cultures in which the Western practice of directly giving an infant the name of an object does not occur.

The frequency with which a caregiver uses a word – in other words, the frequency of that word in the learner's environment -- also impacts the learning of that word (Goodman, Dale, & Li, 2008). A direct test of the hypothesis that more frequent exposure to a word leads directly to an earlier age of acquisition for that word used the Communicative Development Inventory (CDI; an inventory by which parents can record which words their child understands at which ages; Fenson et al., 1993) and the CHILDES database (MacWhinney, 2000), larger databases than had previously been used to look at the relationship between frequency of caregiver use and child learning (Goodman, Dale, & Li, 2008). The authors additionally assert a conservative measure of input frequency, because of the variability in frequency across caregiver-infant pairs. Finally, the large databases eliminate the effects of particular pairs focusing on particular categories or types of language, for example, referential (using words to name objects and actions) versus social (using words such as “hi” and “bye”). The CDI was used to determine the age of acquisition for various words, and the CHILDES corpus to estimate the frequency with which caregivers used those words. The study of these databases showed that overall, noun labels were used least frequently by parents but learned earliest by children, as opposed to closed-class words (e.g., function words and pronouns), which were used most frequently and learned the latest. Note that there are prosodic differences between open and closed-class words as well as frequency differences; the authors' point is that words used more commonly by caregivers appear to be learned more easily, and the only exception to this is in the relatively unstressed closed-class words. The only words used frequently by caregivers but not learned by the age of two (the oldest age examined in this study) fell into this closed-class category. No words learned by age two were absent from the measure of caregiver use. In other words, according to these data, children did not learn words their caregivers did not use. Within lexical categories, however, the expected effect held true, with children learning earliest the words which occurred most frequently. These correlations were stronger using frequencies derived from CHILDES than from comparative databases such as the Francis-Kucera frequency norms, which are derived from written texts (Kucera & Francis, 1967). Goodman, Dale, and Li (2008) additionally suggest that frequency becomes more important for nouns but less important for function words over the course of development, but this is not confirmed and parsing this change in the importance of frequency for various word classes requires additional study (Goodman, Dale, & Li, 2008).

Of course, the influences of the environment interact with object and category properties during word learning. Labeling can serve as a cue for the learner to attend to an object that has appeared consistently while the label was repeated. This finding does not hold true for verbs, reinforcing the special status of object names in word learning. (Echols, 2002, cited by Echols & Marti, 2004).

Models and mechanisms

From a computational perspective, our drive is to understand how lexical acquisition arises through the interaction of simple processing neurons. In other words, we seek an explanation of how brains do semantics. Rogers and McClelland (2004) devote a book to this question, and go a considerable distance toward answering it. Rogers and McClelland (2004) agree with Bloom (2000) about the fragility of the argument that categories hang together on the basis of causal theories. It should be possible to account for category cohesion without basing that account on high degrees of knowledge about various domains (Rogers & McClelland, 2004). However, it may be possible to argue that a person theorizes that the properties shared by birds all have something critical to do with what it means to be a bird, without making claims as to what that *thing* is. In this case, Rogers and McClelland's argument is something of a straw man; theories requiring domain knowledge are arguably impossible to learn in early (prelingual) childhood, but a general theory that properties cohere for a reason could conceivably be an innate bias.

A common failing of accounts of lexical acquisition is the lack of a mechanism by which the observed developmental changes in language can take place, a failing cited by Bloom (2000), Rogers and McClelland (2004), and Landau (2004). Hirsh-Pasek and her colleages (2004) developed the *emergentist coalition model* to address some of these mechanistic issues. Their model stipulates that “children mine a coalition of cues on their way to word learning;” that “the cues for word learning *change their weights* over developmental time;” and that the principles for word learning are emergent and not given a priori” (p. 181). This model accounts for the changing valence of cues for, for example, categorization, as discussed above, but though it serves to explicate the developmental process of these cues, a mechanism to account for the observed changes is absent. These missing mechanisms are best accounted for by cognitive scientists engaged in connectionist modeling.

Rogers and McClelland (2004) focus on a specific type of model often called the Rumelhart model. This model uses two forms of input representation, one for the input itself (as a visual input) and one to identify a connection between the input and the properties to be produced in the output. For example, the total input could specify that a robin IS, with the properties to be output *small* and *brown*, or that it CAN *fly* and *sing* . The original Rumelhart model (Rumelhart and Todd, 1993), from which Rogers and McClelland (2004) draw their model, was derived from Quillian's Collins & Quillian, 1969) hierarchical model of semantic structure. Collins and Quillian (1969) tested reaction time predictions of Quillian's (1967, 1969) semantic network model. They suggested that the time to retrieve semantic information from a node and the time to move up a level in the hierarchy are additive, and would be reflected in the time subjects needed to determine the truth of falsity of subset-superset or characteristic statements. Although Collins and Quillian's (1969) reaction time predictions were in many instances borne out, they did note violations of those predictions, such that a subset-superset relation such as “a dog is an animal” resulted in a faster reaction time than a relation with a more intermediate step, such as “a dog is a mammal.” It is these violated predictions that Rogers and McClelland (2004) specifically tried to account for.

In the Rumelhart model, (Rogers & McClelland, 2004) the input projects to an internal, or hidden, layer, with this hidden layer projecting to a second hidden layer that also receives from the connection input. This second hidden layer projects to the output, where the properties of the input are named (Rogers & McClelland, 2004; Rumelhart & Todd, 1993).

The Rumelhart model emphasizes slow, interleaved learning. The Rumelhart model itself does not incorporate fast mapping, the ability of a child to learn the meaning, or at least the partial meaning, of a word on a single hearing, but Rogers and McClelland believe such fast mapping can be accounted for when the Rumelhart model is paired with a model of the hippocampus, where they localize this sort of single-trial learning (for information on the hippocampal model, see McClelland, McNaughton, & O'Reilly, 1995; Rogers & McClelland, 2004). In contrast to Quillian's model (Collins & Quillian, 1969; Collins & Loftus, 1975), the Rumelhart model does not require an explicit hierarchy search, because all information can be accessed through a specific representation. In other words, semantic information can be accessed from any level in the network. It is not necessary to move up a level from *sparrow* to *bird* to determine that sparrows have wings. This does not necessarily mean that semantic information is not organized hierarchically but merely that a such an organizational hierarchy is not necessary. The Rumelhart model also does not require specific category information to be provided during learning, as the correct category can be induced from other item information. The Rumelhart model also shows property inheritance because it uses distributed representations and small learning changes (Rogers & McClelland, 2004; Rumelhart & Todd, 1993). The model in general complies with the criterion Bloom (2000) specifies for lexical acquisition; that is, it provides a general cognitive basis for word learning rather than a dedicated module.

The Rumelhart model allows for categories to emerge, gradually differentiating from each other, much as is seen in infants even before they learn to speak. The Rumelhart model provides a mechanism for this gradual change, as the hidden layer representations of different items slowly become distinct from each other under the changes driven by feedback (Rumelhart & Todd, 1993). Rogers and McClelland (2004) argue that children are sensitive to coherent covariation of object properties – that is, they do statistical learning – learning which properties are relevant and should therefore be attended to. They do not demonstrate that this salience is learned so much as that it could plausibly be learned, though admittedly it is more parsimonious at this time to say that salience can be learned, rather than salience is learned. In response to researchers who ask how a child can learn which correlations are important (Ahn et al., 1995; Keil et al., 1998, cited by Rogers & McClelland, 2004), Rogers and McClelland point out that this learning is not done with correlation alone but with *coherent* *covariation*; important properties correlate in multiple objects, over multiple contexts, and it is this coherence that allows children to learn which properties are important, and use them to drive category development. This is demonstrated in the Rumelhart network, where it is possible to learn, for example, that color is an identifying characteristic of flowers but not of cars, or that things that can move and have eyes are living things, but things that can move but do not have eyes are non-living things (Rumelhart & Todd, 1993; Rogers & McClelland, 2004). An empirical demonstration that children do, in fact, learn in this manner is lacking, but two important features set this account apart from that of the “theory theory”; the mechanism is apparent, and it is biologically plausible (that is, it is explicitly biologically plausible; the plausibility of the “theory theory” is indeterminate).

Coherent covariation seems important for cross-context learning, that it is possible to learn different things about objects over different contexts. Each unique situation will drive learning about a different aspect of the object or context. Rogers and McClelland (2004) suggest that the Rumelhart model is still too simple to account for this kind of learning. In this case, however, they may be underestimating the power of the model. In a Rumelhart model, (see pp.16-18) partial patterns can activate complete ones (Rumelhart & Todd, 1993), and there is no inherent difference between what constitutes a “partial” pattern and what constitutes a complete one. So even if different contexts activate different partial patterns, it may still be possible for the Rumelhart model to learn new things about the object represented by the complete pattern, and no limit on this is discernable from the structure of the model.

In the Rumelhart model, feedback and nonconceptual content – that is, information that can be perceived from the environment – drive learning, and targets drive the differentiation of the concepts that are being acquired (Rogers & McClelland, 2004). In their look at environmental influences on learning, Rogers and McClelland (2004) seem to be too hasty to draw conclusions about the influence of environmental word frequency, following Merris and Merris (1982, cited by Rogers & McClelland, 2004). Of chief concern is that they chose words to study and looked for those words in caretakers' speech, rather than choosing words from caretakers' speech, a procedure that may omit from study the most frequent words in child-directed speech. Rogers and McClelland (2004) presented patterns representing these words to their model each with its own frequency, and showedthat “the structure of the output properties across items in the network's environment determines the similarity of the model's internal representations” (p. 208). In other words, the items to which the model was exposed most frequently have the most distinct internal representations. Word frequency may also contribute to the aforementioned basic level effect in category learning (Rogers &k McClelland, 2004), but the basic-level effect may impact word frequency by making basic level terms more “accessible” to caregivers in a cyclical sort of effect; typicality may also have the same impact.

In general, Rogers and McClelland (2004) focus admirably on what they term “ecological validity” (p. 211), the principle that what is in the world affects what is in the brain. To this end, overgeneralization in the Rumelhart network is shown to be influenced by the category differentiation already extant at the time an item is first named. The effect of familiarity or frequency of some words over others lessens as differentiation continues[2]. Additional frequency manipulations created expertise effects in the model, and frequency and similarity are shown to interact (Rogers & McClelland, 2004). Rogers and McClelland (2004) explicitly make the case that the environment shapes cognition in models and humans. However, they omit Hebbian learning from the model, and item supervision as an environmental variable. In short, they use a great deal of co-occurrence, but not the Hebbian learning that learns from such co-occurrence more directly than backpropagation and error-driven learning (see Appendix A). However, while they consider that error-driven learning must

[2] This predicts the familiarity effect seen in semantic dementia, in which items cannot be named nor their functions described, but the patient knows that he or she has used that item at some time in the past.

be at work in word learning, Rogers and McClelland are not committed to the notion that only error-driven learning is at work (p. 350), which opens the door to combining a Rumelhart-type model with broader learning paradigms (see Appendix A for information on computational learning mechanisms such as those referred to here).

In discussing the extension of names to novel exemplars, a trait called inductive projection or more usually, generalization, Rogers & McClelland (2004) criticize authors, including Collins and Quillian (1969) on the grounds of lacking mechanisms – previous researchers have failed to show how accumulating information causes the reorganization of conceptual knowledge, including incorporating new exemplars into existing categories. The development of inductive projection in the Rumelhart model is not very different from the development of other features of the Rumelhart model in that the same patterns of development hold (Rogers & McCelland, 2004). However, it is not reassuring that properties are said to be handled "quite" and "somewhat" differently from each other (p. 276); some statistical information on the obtained results would be "quite" nice here.

In contrast to the supervised Rumelhart model, some models learn on the basis of unsupervised learning, learning without feedback. One unsupervised model of meaning is the latent semantic analysis (LSA) model, which learned word meaning from a body of text (Landauer & Dumais, 1997). This model used word co-occurrences to develop representations in a semantic vector space on the basis of their similarity (i.e., the smaller the angle between two word vectors, the more similar the words). The model was quite sophisticated; it not only developed similar representations for two words on the basis of those two words appearing together in text, but also on the basis of those words appearing independently but in similar contexts. Thus, if the word *gear* co-occurs with *cars* and *brakes* also co-occurs with *cars*, *gears* and *brakes* will have a degree of similarity in their representations even if *gears* and *breaks* themselves do not co-occur (Landauer & Dumais, 1997, p. 215).

The semantic space postulated by this model (and others) is multidimensional, but it is an open question as to the number of dimensions that define the word vectors (Jordan, 1986; Landauer & Dumais, 1997). Landauer and Dumais (1997) varied the dimensionality captured by LSA in a range from 1 to 10,000 dimensions. The model's best performance occurred at around 300 dimensions. Model performance was measured by the synonym test of the *Test of English as a Foreign Language* (TOEFL); at the model's best performance, it matched the average score of human test-takers (Landauer & Dumais, 1997). The issue of dimensionality is an important one, but it is difficult to compare this aspect of the model's functioning to mechanisms for human semantic behavior. How do we measure the dimensionality of a human speaker's semantic space? Unfortunately such a thing has yet to be accomplished. The transparency that is one of the benefits of a connectionist model is that its cognitive mechanisms are transparent. Unfortunately, this transparency does not always extend to human mechanisms! Much remains to be learned before the comparability of LSA to human mechanisms can ultimately be established.

Landauer and Dumais (1997) point out that reference and usage both are components of word meaning. The LSA model primarily handles issues of usage by organizing words in semantic space on the basis of their covariation with the other words with which they are used. However, other researchers have demonstrated how another unsupervised system can learn semantic reference (Roy, 2000; Roy & Pentland, 2002).

The SUSTAIN model (Love, Medin, & Gureckis, 2004) is similar to the Rumelhart model in the importance it places on the roles of environmental feedback and slow learning rates in the discovery of category structure. Also like Rogers and McClelland (2004), some importance is placed on the notion of ecological validity. However, SUSTAIN adds innovative features, namely, the incorporation of unsupervised learning in the form of unlabeled exemplars (analogous to the unnamed but similarly-shaped items that can drive category formation in 4-month-olds; Behl-Chadha, 1996) and the goals of the learner. Putting aside the latter, the inclusion of unlabeled exemplars adds to the ecological validity of the learning environment and subtly but critically alters the shape of the learning task.

SUSTAIN derives categories on the basis of dimensions similar to the features of the Rumelhart network (Love, Medin, & Gureckis, 2004; Rogers & McClelland, 2004). The model tends toward simple solutions (though this can be violated if exemplars are presented in an unfavorable order), clusters similar stimulus items, incorporates both supervised and unsupervised learning, predicts differential solutions on the basis of different patterns of feedback, and uses competition among clusters of nodes to drive category differentiation. This contrasts with the Rumelhart network in that 1) Rogers and McClelland (2004) suggest that the order of presentation ideally should not influence the performance of the trained network, and 2) the Rumelhart network does not incorporate unsupervised learning (though see O'Reilly & Munakata, 2000, for a version that incorporates the Hebbian learning algorithm (see Appendix A) though does not utilize unlabeled exemplars).

The SUSTAIN model incorporates feedback on classification trials. In unsupervised trials, all stimuli are assumed to be members of the same category. However, rather than being truly unsupervised in the Hebbian sense, SUSTAIN self-supervises; that is, it recruits a cluster of nodes when the stimulus is insufficiently similar to the internally represented pattern it most resembles. While this is not feedback from the environment, it does rely on expectation violation and can be considered internally, rather than externally, supervised (Love, Medin, & Gureckis, 2004). It is also significant that SUSTAIN fits several sets of human data, either supervised or unsupervised, but never both supervised and unsupervised learning over the same set of data. While SUSTAIN is a good match for human classification experiments, it is not a good choice for expanding to considerations of lexical acquisition, in which learners must be sensitive to conditions in which objects are named and in which they are not named, in which learning will be reinforced by interaction with a caregiver or in which it is motivated by observation of extralinguistic information or by overhearing.

The importance of the environment is also underscored by the cross-channel early lexical learning (CELL) model (Roy & Pentland, 2002). The CELL model learned word referents from un-annotated raw data. Samples of child-directed speech were presented concurrently with static object images. The CELL model learned to identify the words in the speech stream and associate them with the correct sets of images. This result demonstrates that unsupervised learning can underlie learning of reference, a point emphasized by Bloom (2000) in his discussion of supervised and unsupervised learning in the acquisition of word meaning. At its best, the CELL model performed with 57% semantic accuracy – that is, of correctly segmented single words extracted by CELL, 57% matched the visual prototype with which CELL associated it (Roy & Pentland, 2002).

As remarkable as this result is – a four-fold increase over the performance of a model that received no visual input – it still does not approach the eventual performance of human language learners. At first glance, the CELL model appears to accomplish both categorization and naming. However, as a model that uses an unsupervised learning paradigm, CELL merely segments its input space on the basis of similarity – in this case, similarity of verbal input correlated with visual input. The complexity and success of CELL come from its use of multiple-modality rather than single-modality inputs (akin to graphing in two dimensions rather than just one).

CELL approaches the problem of category based on naming – CELL's object categories are entirely built around names; it does not induce categories for which it has no names. Surely this is an inadequate view of human semantic learning. Many studies – exactly those reviewed by Bloom (2000) and Rogers & McClelland (2004) – have examined category formation in preverbal children. Of course, what it means to be preverbal could be debated; the inability to produce a category name is not necessarily indicative of not knowing the category name.

While existing models capture some important aspects of learning, they are not completely satisfactory for lexical acquisition. This is most obvious as it pertains to the ecological validity of a model. Both Rumelhart and SUSTAIN, while paying attention to the learning environment, do not go far enough in capturing its complexity. To solve this problem, it is necessary to look into additional learning paradigms that can support environments containing more diverse types of data.

Chapter Two

Semi-Supervised Learning

[Note: The reader will require familiarity with connectionist modeling in order to understand this chapter. For reference, a primer on connectionist modeling has been provided in Appendix A.]

Semi-supervised learning refers to any learning paradigm that uses both labeled and unlabeled data to train a model or classifier (Anagnostopoulos et al., 2003; Chen, Wang, & Dong, 2003; Nigam & Ghani, 2000; Nigam et al., 2000). In a neural network, therefore, semi-supervised learning must make use of both supervised and unsupervised learning simultaneously. However, not all combinations of supervised and unsupervised learning are semi-supervised (O'Reilly, 2001; O'Reilly & Munakata, 2000); the use of both types of data is essential to this paradigm.

Data that are 'labeled' have a target output or class label associated with them; 'unlabeled' data have no such target (Yarowsky, 1995). These data types are most commonly used in supervised and unsupervised learning paradigms, respectively, with a prominent exception in the Leabra algorithm, which implements an unsupervised, Hebbian learning procedure on labeled data during supervised learning (O'Reilly, 2001; O'Reilly & Munakata, 2000).

Semi-supervised learning may be implemented in any number of algorithms. Two commonly used are expectation maximization algorithms (Collins & Singer, 1999; Gharamani & Jordan, 1994; Muslea, Minton, & Knoblock, 2002; Nigam et al., 2000) and support vector machines (Bennet & Demiriz, 1998; Chen, Wang, & Dong, 2003; Li et al., 2003; Tong & Koller, 2001; Vapnik, 1998). In their classifiers, Anagnostopoulos and his colleagues (Anagnostopoulos et al., 2003, see also

Anagnostopoulos et al., 2002; Anagnostopoulos & Georgiopoulos, 2001) use semi-supervision in conjunction with adaptive resonance theory (ART). Another method uses a self-organizing output with a conventional supervised learning paradigm (Sarrukai, 1997), and a fourth, adaptive learning, implements a network which queries a human user for labels during training, when it encounters unlabeled exemplars that are detected as being particularly informative (Muslea, Minton, & Knoblock, 2002; Tong & Koller, 2001). These latter two methods are semi-supervised in the sense of using both labeled and unlabeled data, but are not truly comparable to the prior methods because of the explicit separation of the two kinds of learning.

Expectation-Maximization Algorithms

Expectation-maximization algorithms take their name from their two steps, the expectation, or E-step, and the maximization, or M-step. These algorithms are commonly used to find the parameters of a set of labeled and unlabeled data or to classify such data into two or more sets (Collins & Singer, 1999; Dempster, Laird, & Rubin, 1977; Gharamani & Jordan 1994; Nigam et al., 2000). It is assumed that however the data are distributed, there is some probability that a given label is assigned to an exemplar (Collins & Singer, 1999). In the E-step, the log likelihood of the unlabeled or missing data -- the probability that the estimated classifiers are correct, or that the predicted data would actually be observed -- is estimated; in other words, the data are classified (Dempster, Laird, & Rubin, 1977; Gharamani & Jordan, 1994; Nigam et al., 2000; Weisstein, 1999). In the M-step, parameters are computed that maximize this likelihood; in other words, a new classifier is constructed (Dempster, Laird, & Rubin, 1977; Gharamani & Jordan, 1994; Nigam et al., 2000). The algorithm iterates until it reaches convergence, resulting in class labels or an estimation of the parameters for all the data, whether they are labeled, unlabeled, missing, or complete (Collins & Singer, 1999; Dempster, Laird, & Rubin, 1977; Gharamani & Jordan, 1994; Nigam et al., 2000). The expectation-maximization iteration process can be described as hill-climbing (or, conversely, gradient descent), as it finds the local maximum of the likelihood of the data (or the local minimum of the error) (Collins & Singer, 1999; Nigam et al., 2000). Expectation-maximization algorithms are semi-supervised in that they rely on unlabeled data to construct the classifier. They are most successful when labeled data are scarce, as it is in this situation that the use of unlabeled data has been found to give the greatest benefit to classifier accuracy (Nigam et al., 2000). This principle applies to all implementations of semi-supervised learning (Chapelle, Schölkopf, & Zien, 2006).

Support Vector Machines

Support vector machines (SVMs) are used to establish parameters for data when a binary classification is needed, though the principle can be expanded to a greater number of classifications (Tong & Koller, 2001; Vapnik, 1998). In the binary-classification SVM, the goal is to find a hyperplane that accurately divides the data into two maximally distant groups (Gat, 2001; Tong & Koller, 2001; Vapnik, 1998). The "support vectors" refer to the training examples that come the closest to the hyperplane (Tong & Koller, 2001). SVMs can use a variety of algorithms, including the perceptron algorithm, in which case the support vectors need not be explicitly represented (Gat, 2001). SVMs may be inductive or transductive. An inductive SVM is not semi-supervised; it computes from labeled exemplars a classifier that is meant to have good performance on all possible exemplars (Tong & Koller, 2001). A transductive SVM is semi-supervised; it computes from both labeled and unlabeled data a classifier that is meant to assign labels to the unlabeled data as accurately as possible (Chen, Wang, & Dong, 2003; Tong & Koller, 2001). Its performance is tested only on the unlabeled data that are already present, not on a new set of test exemplars (Tong & Koller, 2001). Transductive SVMs tend to show more accurate classification of novel exemplars because of the inclusion of unlabeled data during training (Chen, Wang, & Dong, 2003; Tong & Koller, 2001). However, they can have a high computational cost, particularly when there is a large amount of labeled training data (Tong & Koller, 2001).

Other Methods for Semi-Supervised Learning

Adaptive resonance theory (ART) is a self-organizing pattern classifier that is meant to balance the tradeoff between plasticity, the network's ability to learn new patterns, and stability, the network's ability to remember the patterns it has already learned (Gurney, 1997). Anagnostopoulos and his colleagues (2003) apply semi-supervision to a version of ART known as fuzzy ARTMAP, a supervised form in which patterns may have graded membership in sets (Carpenter et al., 1992; Gurney, 1997). The addition of semi-supervised learning to fuzzy ARTMAP was intended to retain the stability of ART architectures but prevent the problem of overfitting the data (Anagnostopoulos et al., 2003).

Sarrukai's (1997) method involves supervised learning in the ordinary sense, but with the caveat that output class labels not be provided explicitly. In this case, supervision provides only the information about whether inputs belong or do not belong to the same class, without specifying the class to which each input belongs. This process tends to preserve the differences among classes that are apparent in the training data set; inputs from the same class should have outputs that are as similar as possible, and inputs from different classes should have outputs as dissimilar as possible. There is a biological justification for this method; in some cases, labels may be available for data learned by a biological system, but in other cases they may not be (Sarrukai, 1997). This point is important for the application of semi-supervised learning to neural networks designed to model human cognition.

Finally, the paradigm of active learning can be used with EM algorithms, inductive and transductive SVMs, and perhaps with other types of learning algorithms as well (Nigam et al., 2000; Tong & Koller, 2001). In active learning, the classifier being trained selects the most informative unlabeled exemplars and asks a human user to label them. The classifier is then trained on the newly labeled data, and then queries the user again (Muslea, Minton, & Knoblock, 2002; Nigam et al., 2000; Tong & Koller, 2001). Using active learning allows for a considerable reduction in the amount of labeled data necessary for the classifier to learn the task without a loss in performance, and the more unlabeled data are available, the better performance will be (Tong & Koller, 2001). This paradigm is not truly semi-supervised in that it does not learn directly from unlabeled exemplars, but it does make use of unlabeled information in a way that improves classification accuracy.

It is most useful to think of semi-supervised learning as being halfway between supervised and unsupervised learning. (We should be careful to realize that it is not always clear which paradigm is used in a model, as there are differences in usage of the terminology among researchers and problem types; Fletcher, 2000.) Remember that supervised (or task) learning is a learning paradigm that is used on labeled data (Gharamani & Jordan, 1994; Hanson, 1995; Jordan & Rumelhart, 1992; O'Reilly, 1996, 1998, 2001; Rumelhart, Hinton, & McClelland, 1986; Rumelhart, Hinton, & Williams, 1986; Stone, 1986). In most cases, the network is fed an input and computes an output representation. The difference between the output that the network produces and the desired target, or label, is used to derive an error signal, which is fed back into the network for the purpose of altering the connection weights such that on successive iterations, the network will generate an output closer to the desired target (Hanson, 1995; Jordan & Rumelhart, 1992; O'Reilly, 1996, 1998; O'Reilly & Munakata, 2000; Rumelhart, Hinton, & McClelland, 1986; Rumelhart, Hinton, & Williams, 1986; Rumelhart & Todd, 1993; Sarrukai, 1997; Stone, 1986). It is also possible to use supervision only to cluster the input data, not to apply output labels to it (Sarrukai, 1997). The exception to the above is the Hopfield network, in which the desired output is clamped and the network settles into a state that will generate that outcome (Hopfield, 1982). In a fully trained network, it is possible to use the input to predict what the output will be (Gat, 2001; Gharamani & Jordan, 1994).

Unlabeled data are sufficient in and of themselves for dividing a data set into classes or clusters (O'Reilly, 2001; Roy, 2000). However, because all the information used to make weight changes in Hebbian learning is local, there is no way to drive the solution of a more global task (O'Reilly, 2001). Supervised learning is necessary for assigning labels to the newly classified sets of data (Nigam et al., 2000). The word sense disambiguation study of Yarowsky (1995), which uses a few previously labeled "seed" instances to label clusters of words divided by an unsupervised paradigm, is dependent on this, and it is nearly opposite to the formulation of Sarrukai (1997), which uses supervised learning but no class labels. In any case, it is another pointer to the utility of learning from labeled and unlabeled data in the same framework.

Connecting the Paradigms

There is a gap between supervised and unsupervised learning; the promise of semi-supervised learning is to connect these two paradigms. However, semi-supervised learning is not the only way of doing so, as indicated in the discussion of mixed-model learning in the previous chapter. The great benefit of semi-supervised learning is that in combining supervised and unsupervised learning within a single learning paradigm, we can model not only brain states with as much accuracy as possible, but we can model the input to those brain states in as realistic a way as possible.

Task learning.

Researchers who have investigated semi-supervised learning have found that in many cases, semi-supervised learning betters the uncombined learning paradigms (Nigam et al., 2000). Adding some supervision to a network using an unsupervised learning paradigm improves task learning (O'Reilly, 1998, 2001). This is not a great concern of most researchers in semi-supervised learning, who are usually comparing semi-supervised learning to supervised learning paradigms (but see Yarowsky, 1995). Their motivation for pursuing semi-supervised learning is to be able to include unlabeled data in the training sets for classifier networks that will carry out the task of sorting or classifying data such as documents in a database or genetic information.

Availability of labeled data.

Semi-supervised learning tends to show better performance than supervised learning in situations where there is much more unlabeled data than labeled data (Guo et al., 2008; Nigam & Ghani, 2000; Nigam et al., 2000; Singh et al., 2008). Adding unlabeled data to a training set considerably reduces the amount of supervised training data necessary for appropriate classification performance (Collins & Singer, 1999). However, in situations where there is a great amount of labeled data or a small amount of both labeled and unlabeled data, inclusion of unlabeled data is not only superfluous, but tends to be detrimental to the classification ability of the network (Nigam et al., 2000; Singh et al., 2008). Semi-supervised learning can also be detrimental in certain situations where the class labels do not match the actual distributions of the unlabeled data (Nigam et al., 2000), though this depends on the method used to implement semi-supervised learning and the actual distributions of the data.

However, semi-supervised learning has usually been applied in situations where there is very little labeled data, either because the true labels are unknown or because training data must be labeled by a human, which is an expensive and time-consuming process (Nigam et al., 2000; Roy, 2000) and prone to errors (Pathak-Pal & Pal, 1987). Semi-supervised learning is also useful in cases where data are misclassified, as some formulations allow labels to be changed after the initial classification of exemplars (Yarowsky, 1995). In these situations, using semi-supervised learning allows for faster and more accurate classification compared to using the available amount of labeled data alone (Nigam & Ghani, 2000; Nigam et al., 2000). In addition, using semi-supervised learning tends to increase post-training error compared to supervised learning (Anagnostopoulos et al., 2003); however, the modest increase in error is what permits better generalization to new data (Anagnostopoulos et al., 2003; Nigam & Ghani, 2000).

Speed.

One interesting feature of combining error-driven and Hebbian learning, as demonstrated in experiments using Leabra, is that error-driven learning is fast, while Hebbian learning is slow. This means that the network learns its task long before it acquires the ability to generalize to new sets of data (O'Reilly, 2001). This may be an important consideration in deciding whether to use semi-supervised learning for a given task. If speed is of primary concern, semi-supervised learning may not be worthwhile. Usually, however, speed in and of itself will be of less importance than biological plausibility and the ability of the network to behave similarly to, as opposed to better than, a human.

Human Learning and Semi-Supervised Learning

To the best of my knowledge, only one study has directly asked the question of whether human learning is semi-supervised. A study by Zhu and colleagues (2007) showed that exposure to unlabeled data improved human performance in a classification task. This result was predicted by the performance of a semi-supervised Bayesian classifier (Zhu et al., 2007).

This is promising evidence for exploring semi-supervised learning in human domains. In considering semi-supervised learning in humans, we should also consider four additional areas: biological plausibility, environmental inconsistencies, ecological validity, and privileged inputs.

Biological Plausibility

First, it is recognized that the biological plausibility of an algorithm is always important in psychology or neuroscience (Fletcher, 2000; O'Reilly, 1998). However, semi-supervised learning is a learning paradigm, not a learning algorithm. This suggests that semi-supervised learning is biologically plausible when implemented with a biologically plausible algorithm. Most research using semi-supervised learning has not utilized biology as a constraint upon its algorithms. However, semi-supervised learning can be implemented in the Leabra learning algorithm. This algorithm, which was specifically designed to be biologically plausible (O'Reilly, 1996, 2001; O'Reilly & Munakata, 2000), also demonstrates how both supervised and unsupervised learning can operate simultaneously over the same synapses. When semi-supervised learning is carried out in a Leabra framework, semi-supervised learning is as plausible as Leabra itself.

Environmental Inconsistencies

Second, there is evidence that humans must learn some things from the environment that are supervised at some times and not at others (Bloom, 2000). One potential partially supervised input to a human brain is language. Natural language provides ambiguous training data, some of which may be expected and some not; in production, some may be corrected and some not; and no incorrect exemplars are ever presented to an early language learner as an example of how *not* to use language (Bloom, 2000). At the environmental level, this seems very much like a good candidate for the kind of data used in a semi-supervised learning paradigm.

Ecological Validity

Semi-supervised learning is more about environmental plausibility rather than biological plausibility. It is not environmentally plausible to assume that all data presented to a human (or other adaptive) learner are labeled (Hanson, 1995; Roy, 2000; Sarrukai, 1997), and some data may be mislabeled (Pathak-Pal & Pal, 1987). Environments are imperfect, and so are the systems that take in data from the environment (Gharamani & Jordan, 1994). In addition, the environment may change as the result of an action, giving the learner information about whether the action was correct or not (Jordan & Rumelhart, 1992). But this will not be true in every case. Semi-supervised learning allows us to model accurately the imperfect and dynamic environments from which human neural systems learn.

Privileged Inputs

In some cases, we may also wish to account for the varying salience of stimuli, either in our models of the environment, in which some stimuli are, for example, louder or more repetitive than others, or in our models of the brain, which may, for example, be primed to process some stimuli. This may be represented by assigning varying weights to unlabeled versus labeled examples. Nigam and colleagues (2000) add to their expectation-maximization algorithm a variable λ, which weights the influence of the unlabeled training instances. When $\lambda = 0$, the unlabeled instances have no influence on learning; when $\lambda = 1$, the unlabeled instances are given the same weight as the labeled instances. Pathak-Pal and Pal (1987) also suggest a way of restricting which training exemplars are used for weight updating, depending on whether any particular exemplar modifies a weight in a way to previous modifications of that weight. There may, in fact, be reason to suppose that this sort of weighting occurs in the brain; at least, there is evidence that precludes its automatic rejection. It is known that synapses may be modifiable only under certain circumstances, such as being active within a particular range of time. This may bias the biological neural network to make use of unlabeled (or indeed, any) data at some times and not at others.

Another approach to changeable weighting of data might come in the form of windowed momentum, in which a partial history of modification at a synapse helps to determine the size of the weight change based on the current exemplar (Istook & Martinez, 2002). The “window” referred to is the number of previous modifications used to bias the current modification. As each new exemplar is presented, the window moves forward, and the oldest exemplar is discarded and the newest is added. A window size longer than 100 tends to increase the amount of time needed to train the network. However, an increased window also allows for increased accuracy because each weight update is smaller, and therefore has less of an effect by itself. This indicates the need for an appropriate balance between speed and accuracy in network training (Istook & Martinez, 2002).

Applying Semi-Supervised Learning to Connectionist Models

Semi-supervised learning has been used for the most part in computer science, with the occasional crossover to another field where an automated classifier is needed (see Li et al., 2003). To date, semi-supervised learning has only been used in psychology in preliminary forms (Robare, 2004, 2005). Groundwork has been laid, in the form of research that has combined error-driven and Hebbian learning in the same networks, but until now the use of both labeled and unlabeled data to train a network has been outside the province of psychology.

Algorithm Selection

Semi-supervised learning can easily be implemented in the Leabra learning algorithm by O'Reilly (2001; O'Reilly & Munakata, 2000). Leabra is ideal for the purpose because of its established biological plausibility, which is a prerequisite for the ability to use unlabeled data for learning. The great advantage of Leabra is that when a network is learning with the Leabra algorithm, error-driven and Hebbian learning are operating simultaneously over the same synapses. The relative strengths of error-driven and Hebbian learning are variable, and Hebbian learning operates more slowly to change connection weights than does error-driven learning, but both mechanisms operate over the same synapses at the same time. If we assume that segregating supervised and unsupervised learning into separate brain regions is unrealistic, and that each synapse or group of synapses is modified by exposure to both labeled and unlabeled data, then the operation of both mechanisms over the same synapses at the same time is an absolute must for the appropriate use of semi-supervised learning to address psychological questions, and it separates Leabra from its potential alternatives.

The utility of algorithms from computer science is likely to be limited. Computer science uses of semi-supervised learning have been divorced from biological considerations. This is reasonable, but limits the transfer of such algorithms to computational modeling in psychology. The mathematical formulations, particularly that of support vector machines, may be useful to psychologists as ways of understanding learning in a theoretical or mathematical sense without speaking to neural properties themselves, in much the same way as the notion of attractors has been used. However, the transfer of algorithms from computer science to psychology will be inappropriate in most cases, because of the different goals that drive the development of those algorithms.

Use of the General Guidelines for Connectionist Models

Though a new mathematical model for learning is exciting, we must be careful to retain the useful principles that have structured the development of previous connectionist models, so as not to become divorced from what we already know about how learning happens in the brain. O'Reilly (1998) has introduced six guiding principles for creating neural networks, which should guide the use of semi-supervised learning in these networks as well. These principles are biological realism, distributed representations, inhibitory competition, bidirectional activation propagation- or interactivity- error-driven learning, and Hebbian learning. The last two, of course, are more or less explicit in any formulation of semi-supervised learning. Adhering to all of these guidelines guarantees that from a theoretical (non-performance) standpoint at least, our semi-supervised models are as good as their supervised and unsupervised counterparts. In fact, semi-supervised learning takes the fullest advantage of these six principles, even more so than using the Leabra algorithm in a fully supervised learning paradigm. As O’Reilly (1998) writes, “. . . combining task-based and model-based learning enables one to account for phenomena specifically with these different types of learning” (p. 461). Semi-supervised learning uses the two types of learning explicity to do what they do best.

Utility

We can expect that semi-supervised learning will have its greatest utility in those psychological domains that are roughly comparable to those domains in which it has been used in the computer sciences. These domains are characterized by a large amount of data, much of which is unlabeled. One such domain is language; semi-supervised learning has been used for sorting documents and “tagging” natural language sentences in computer science (Collins & Singer, 1999; Nigam et al., 2000; Yarowsky, 1995), and will be useful in a model of human language learning as well. Other domains that are characterized by large, ambiguous sets of data may also find uses for semi-supervised learning, but language may be something of a test case.

The use of semi-supervised learning in computational models and the study of language seem made for each other; as we now have a learning paradigm that allows us to consider the effects of an ambiguous data set with erratic corrective feedback and a data set that demands just this sort of paradigm.

The research described below has helped explore and describe the parameters and outcomes of semi-supervised learning in connectionist networks. Despite certain ambiguities in the results, these simulations have shown the potential for semi-supervised learning as a paradigm for thinking about computational learning in organisms' brains.

Simulation 1 is a proof-of-concept, meant to check the feasibility of programming a semi-supervised network in the PDP++ software. Simulation 2 addresses the ability of semi-supervised learning to solve a simple cognitive problem in a connectionist framework, and Simulations 3 and 4 confirm and extend the results of Simulation 2. Simulation Five considers not only the question of whether a language-like network using semi-supervised learning differs in training and generalization from a fully-supervised network, but also how the amount of supervision affects training and generalization. Simulation Six asks how Hebbian learning value (k_{hebb}) affects training and generalization in semi-supervised networks with varying amounts of supervision. Finally, Simulation 7 systematically manipulates the dimensionality of the inputs to the network, to determine that the varying dimensionality of the earlier simulations did not impact the results. Conceptually, Simulations 1 and 2, 3 and 4, and 5 and 6 form units, whereas Simulation 7 is different from the other simulations and stands alone. As a group, these simulations provide a test of semi-supervised learning in humanlike cognitive domains and demonstrate the feasibility and potential superiority of semi-supervised learning over supervised learning for language and other complex data sets in human learning. Thus they tell us that human cognitive processing may be more akin to semi-supervised learning than to supervised or unsupervised learning when data are labeled and unlabeled.

Proof-of-Concept

Simulations 1 and 2

With a new paradigm such as semi-supervised learning, a number of basic, practical questions need to be answered before more substantive questions can even be asked. Can a semi-supervised connectionist network simulation be programmed appropriately? Will it learn in a way that is comparable to other networks? Does the inclusion of unlabeled data have any effect on network performance? These are simple questions, which were addressed with simple simulations, but their simplicity does not make them trivial. Failure here could indicate insufficiently complex programming of the learning algorithm, or a severe deficiency in understanding the relationship of supervised, unsupervised, and semi-supervised learning. The first two simulations therefore serve as proof-of-concept, a preliminary trial of the practicalities of semi-supervised connectionist networks.

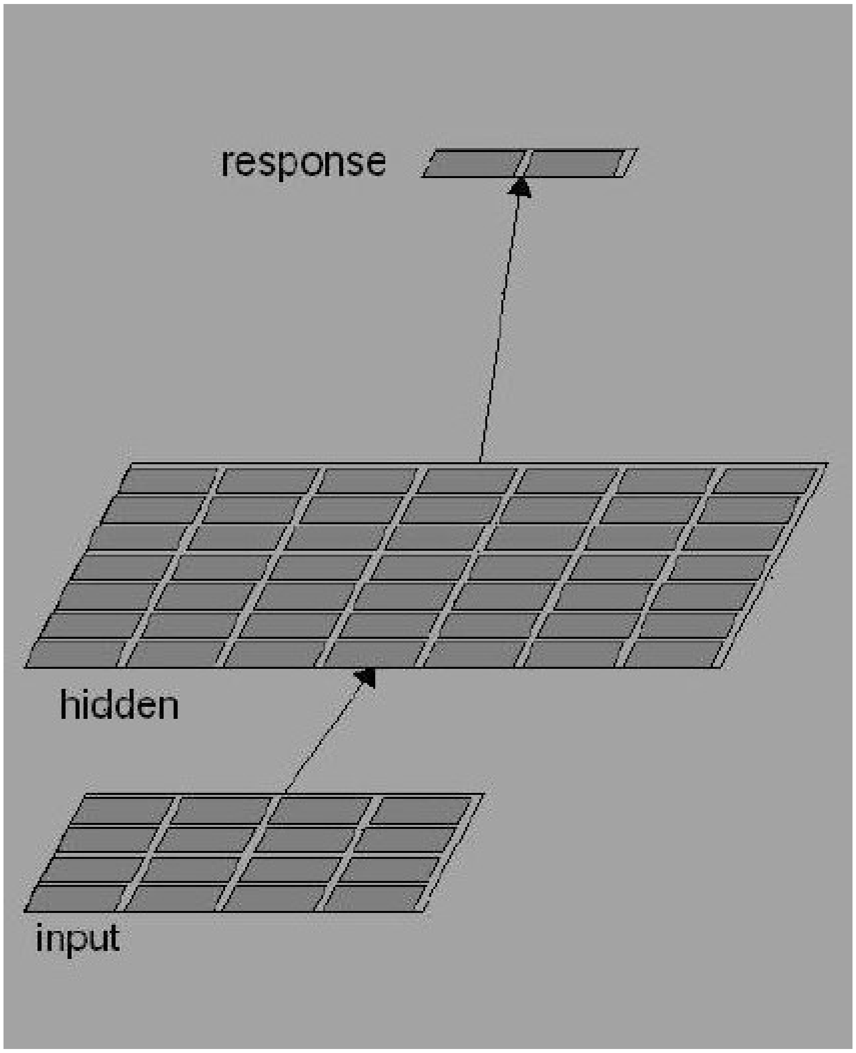
All simulations described below were implemented using the PDP++ software (Carnegie Mellon University, 1995).

Architecture

Simulations 1 and 2 used a simple, two-layer network (Figure 1). There were 16 units in the input layer, arranged in a 4 x 4 grid. The patterns that served as the input to the network were arranged on this grid. The hidden layer contained 49 units. (It was reasoned that learning would probably have been equally effective with a smaller number of hidden units, but overestimation would not hurt performance where underestimation could prevent learning.) The output layer consisted of only two units and served to indicate the binary decision of the network.

Figure 1

Architecture of Simulations 1-4



Simulation 1

The first experiment trained the network to decide whether a line was on the right- or left-hand side of the input space. Networks were trained to recognize vertical lines of three or four units on an input space of 16 square units. The lines were divided equally between the right and left sides of the input space, and 75% of the data were supervised. This simple task was easily learned by networks using both semi-supervised and fully supervised learning, with networks trained with both paradigms quickly reaching a ceiling level of performance. This experiment does not need to be further elaborated ; it served merely to demonstrate that a semi-supervised network can be run in Leabra, and that it does learn a simple task as easily as a fully supervised network.

Simulation 2

A simple two-layer network was trained on the task of classifying a line as horizontal or vertical (Please note, this simulation was intended to demonstrate the usefulness of the learning paradigm, not to make claims about the way the human visual system handles line orientation.) There were 16 units in the input layer, arranged in a 4 x 4 grid. The lines that served as input to this network were arranged on this grid. The hidden layer initially contained 49 units; a later version of the same simulation reduced that number to 16 with no statistical difference in the behavior of the network (see results below). The output layer consisted of only two units and served to indicate the binary decision of the network.

The task was presented in both fully supervised and a semi-supervised paradigms. The supervised data presented to each network were the same: of the four units in the grid that made up any horizontal or vertical line, three were active. Each network saw four lines in each orientation. The semi-supervised network saw an additional four patterns, two full lines in each orientation, which were not labeled. The networks were trained to 1500 epochs and then tested on the four patterns that had been presented, unlabeled, to the semi-supervised network. Over 50 test runs of each network, the semi-supervised network was correct on all four patterns more often than the fully-supervised network, *t* (98) = 2.003, $p < .05$.

This result demonstrates that semi-supervised learning has a tangible effect on the performance of a network; inclusion of unlabeled training data does have an impact on the performance of a network. However, this effect was seen in performance on patterns that had already been presented to the semi-supervised, but not to the fully supervised network. Therefore, without additional testing we could interpret this result as a benefit of having seen all the test patterns, rather than any effect of semi-supervised learning. To rule out this possibility, the network was tested on its generalization performance, its ability to classify novel patterns. Remember that one important motivation for using semi-supervised learning is that under some conditions it is expected to show better generalization than a fully supervised network. For a test of generalization, the networks were again trained to 1500 epochs. Over 50 runs, the semi-supervised network indeed showed better generalization to novel patterns not seen by either network prior to the test, *t(*98) = 2.449, *p* < .02. Therefore, on the basis of the first two simulations, we know that semi-supervised learning can be implemented in a biologically plausible connectionist network.

Chapter 3

Learning and Generalization: Simulations 3 and 4

Simulation 3

Simulation 3 investigated the course of learning and generalization performance in semi-supervised and fully supervised networks. Thirty runs were made of each of the orientation networks described for Simulation 2. In each run, the network was trained one epoch at a time, and the training set error and generalization performance were recorded. When the error fell to zero, and remained at zero for seven epochs (seven was chosen because occasionally the error fell to zero, remained for up to six epochs, and then rose above zero again), the network was then trained ten epochs at a time until reaching or surpassing 100 epochs. By this method, enough data points were obtained to investigate the learning curves of the networks.

These learning curves are very nearly identical. A slightly longer time was needed to reach and remain at zero error in the fully supervised network, but this is an effect of the unusually long training times needed for a couple of the runs of this network. The mean time to train to zero is not different between the fully supervised and semi-supervised networks, *t* (28)= 0.221, *p* = .826.

Interestingly, generalization in these networks is also not different, *t* (28) = 0.318, *p* = .752. There is a discrepancy between these results and those of Simulation 2, which can be explained by the amount of training given to each network.

The networks in Simulation 2 were trained to 1500 epochs, whereas those in Simulation 3 were trained only to (approximately) 100. This supports the assertion that in any network, generalization requires more training than learning the training set itself does (O’Reilly & Minakata, 2001). We may also hypothesize that the benefit of semi-supervised learning will not be seen in the early stages of training a network. At small durations of training, semi-supervised and fully supervised learning are functionally equivalent. However, at longer training periods, semi-supervised learning gives advantages in generalization to novel patterns. It would not be surprising if this effect proves to be more apparent with large numbers of training epochs and larger and more complex sets of data. I would go so far as to predict that the more complex the network task, the more useful semi-supervised learning will be.

Simulation 4

Simulation 4 confirmed that the discrepancy between Simulations 2 and 3 was due to different training durations. The networks described for Simulation 3 were trained to 3000 epochs.

Performance on the training set was tracked over 30 runs of each network separate from those used in the analysis of generalization performance. The variable is number of epochs to train to 0 error, in other words, to perfect performance on the training set. In terms of the length of time it takes for each network to learn the training set, fully supervised learning and semi-supervised learning are statistically the same, t (28)= 0.313, p = 0.76.

On the generalization set, summed squared error (SSE) and the number of correct responses to items in the set (generalization correct) were negatively correlated, r = -0.814, p < .01. This was true for fully supervised and semi-supervised networks, r = -0.776, p< .01 and r = -0.836, p< .01, respectively. At 3000 epochs, at which duration generalization was tested for the last time, the semi-supervised network showed significantly lower SSE, t (28)= 2.198, p= .032, and significantly higher generalization correct, t (28)= -2.122, p = .038 (Table 1). The prediction of better performance in the semi-supervised network holds. For generalization, accuracy is higher (i.e., the network gives a greater number of correct responses) in the semi-supervised than in the fully supervised network. There is also a greater degree of change in generalizations over time (comparing generalization at 100 and 3000 epochs of training) in the semi-supervised than in the fully-supervised network, t (28)= 2.084, p < .05 (though this is not true for SSE, t (28)= -1.243, p = 0.219) (Table 2). This latter result further suggests that the hypothesis raised by Simulation 3, that semi-supervised learning is advantageous at long but not short training durations, is correct.

The Power of Semi-Supervised Learning

Overall, the results demonstrate the power of semi-supervised learning as a paradigm. Although semi-supervised and fully supervised networks perform similarly on the training set, they perform very differently on measures of generalization to novel instances. The semi-supervised networks consistently outperform the fully supervised networks. This has important implications for models of cognitive processes.

The simulations discussed in the rest of the current work are intended to be more relevant to the question of language learning than those discussed above. Simulation Five addressed two questions. One, in a language-like data set, does semi-supervised learning perform differently from supervised learning at (a) training and (b) generalization? Two, how does the percentage of supervision in a semi-supervised model affect training and generalization performance? Simulation Five used categories and category exemplars similar to those used by Rogers and McClelland (2004) in their explorations of the Rumelhart network (see pp. 14-17). (Categories used in Simulation Five included birds, quadrupeds, insects, fruit, flowers, and trees.) Appendix B shows the list of training and generalization data used in Simulation Five, as well as selected images of the input presented to the network.

The question of supervision percentage is important to address in this initial exploration of semi-supervised learning for connectionist models of language. There are two reasons for this. First, to say that a model is learning in a “semi-supervised” fashion merely means that not all input data are labeled. In a data set of 40 inputs, for example, anywhere from one to 39 of those input items might be labeled, and it is reasonable to think that learning may be different with more or fewer labeled items.

Second, it is being theorized here that human language learning is semi-supervised in the computational sense. Therefore, there may be some particular proportion of the input to a language learner that is labeled. If the model is accurate enough, this proportion of unlabeled input may be predictable. That would be a powerful test for the model and is beyond the scope of the current research, but the question still has bearing on the present effort because making this prediction will require knowing whether the model always performs better with more labeled inputs, or whether there is a U-shaped curve for performance such that some intermediate level of supervision results in the best performance.

Simulation Six addressed the question of whether proportion of Hebbian learning and proportion of supervision interact to affect training and generalization performance in a semi-supervised model.

Recall that the Leabra equation is:

$$\Delta w_{ij} = \varepsilon[k_{hebb}(\Delta_{hebb}) + (1 - k_{hebb})(\Delta_{sberr})]$$

(O'Reilly, 2001; O'Reilly & Munakata, 20000; see Appendix A.). The term:

$$(k_{hebb})$$

represents the unsupervised learning; it gives the amount of change in the weight between two units that is driven by Hebbian learning. Unsupervised learning operates independently on unlabeled data. That is, on such data, 100% of the learning is unsupervised, and the Leabra equation reduces to:

$$\Delta w_{ij} = \varepsilon[k_{hebb}(\Delta_{hebb}) + 0] = \varepsilon[k_{hebb}(\Delta_{hebb})].$$

When unsupervised learning operates independently over a larger proportion of training data, the pattern of learning may change. Therefore, Simulation Six investigated the potential interaction between the unsupervised learning (k_{hebb}) and proportion of supervision in the training data.

The data for Simulation Six were based on those used in the CELL model (Roy & Pentland, 2002). Inputs for Simulation Six corresponded to spoken words and visual images of six categories of items commonly used as toys (cars, horses, etc.). The features (either visual or phonemic) represented in the input varied from item to item. The varying amounts of Hebbian learning and proportion of supervision in the model were crossed and the effects on training and generalization performance were measured.

Simulation Seven addressed the question of dimensionality of the training and generalization data. Because the previous simulations used data with varying dimensionality, their differing results may have been due to the differing dimensionality of the data. Simulation 7 therefore systematically manipulated dimensionality of data and overlap within and between categories. Although Simulations Five and Six presented language-like inputs to the model, Simulation Seven used more abstract data in order to allow for this systematic manipulation. These data were presented as binary patterns on a grid, with more similar patterns classified as being in the same category (except in the case of irregular exemplars, which were included in this data set). Views of some of the input items used in Simulation Seven can be seen in Figure 2.

The abstraction of Simulation Seven distances the model one step from language learning. However, the precision of the manipulations allowed by this abstraction are worth the distance. The varied data sets of Simulations 1-6 may have at least as much to do with the results of those simulations as the factors being deliberately manipulated. Simulation Seven was therefore envisioned as a check on the results of those simulations.

Together, Simulations Five, Six, and Seven are the predictive models of the current research. Their results may be used to predict the results of empirical studies of semi-supervised learning in lexical acquisition.

Chapter Four: Models of Meaning

Simulation Five

Rationale.

The horizontal/vertical discrimination task used in Simulations 1-4 is simple from a human point of view. It is an easy matter for the experimenter to determine whether learning in the network has been successful. However, from a computational standpoint it is a more difficult problem. Consider the horizontal and vertical lines superimposed on the 4 x 4 input grid (Figure 1). On this grid, each individual input unit carries no information about whether a line is horizontal or vertical because every unit carries an equal probability of appearing in each kind of line[3]. A more regular domain in which the individual units carry more information may make semi-supervised learning more or less useful for learning. A category learning task affords a quasi-regular domain and a more "semantic" flavor. Simulation 5 provided the opportunity to not only examine the central question of generalization in semi-supervised versus supervised networks in this category learning task, but also to define some of the parameters important for this different modeling paradigm.

With Simulation Five, we attempted to answer two questions: One, does a semi-supervised model perform differently from a supervised model at training and at generalization? Two, how does the percentage of supervision in a semi-supervised model affect the model's training and generalization performance? These basic questions were explored in a model which performed category learning in a manner similar to that in the Rumelhart network (Rogers & McClelland, 2004; Rumelhart & Todd, 1993; see pp. 14-17).

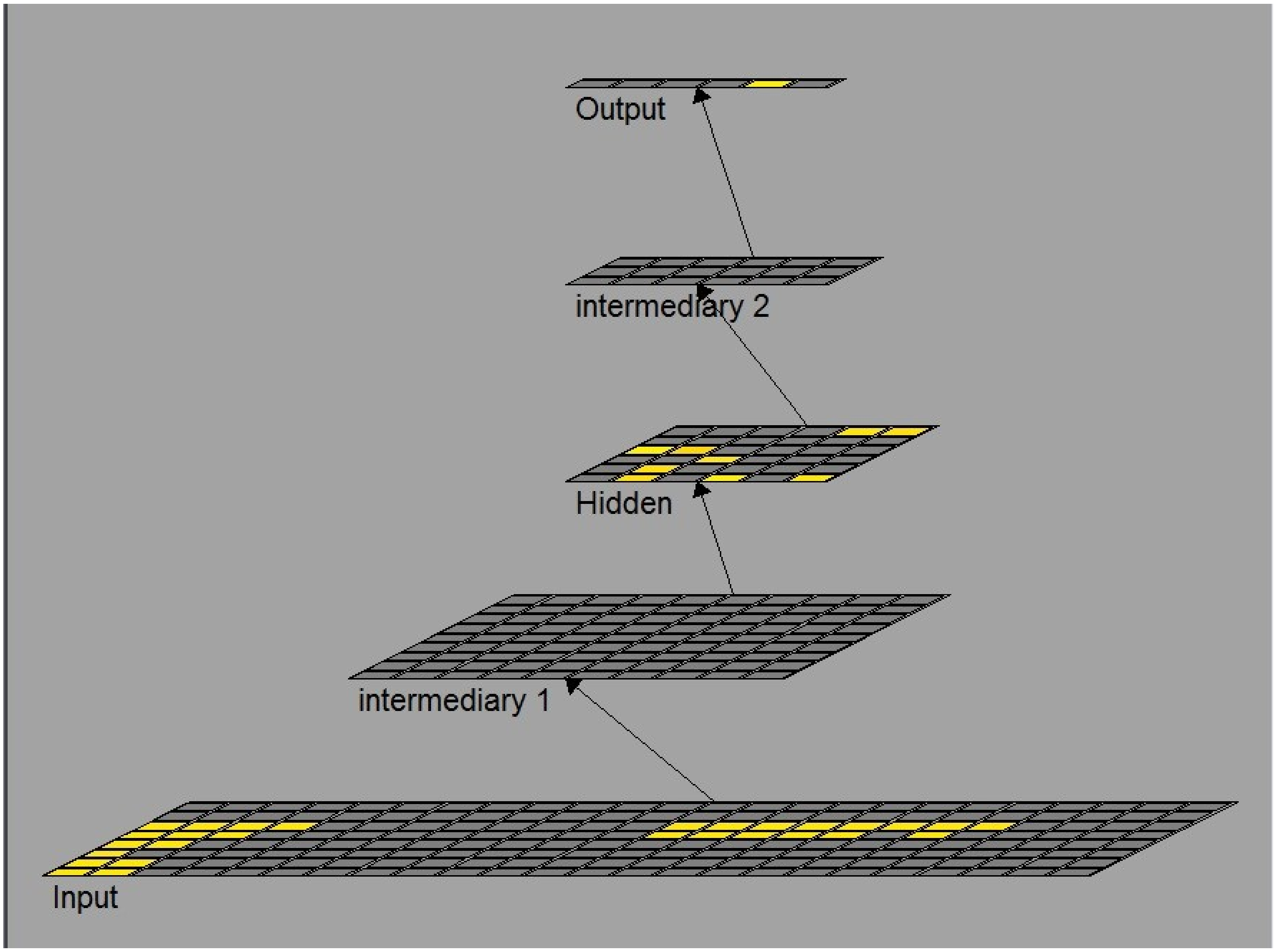
Architecture.

A feedforward network with a single input layer and multiple hidden layers was used in Simulation Five. This architecture was designed to slowly combine representations into single unit outputs without compromising learning by too quick pattern compression. The architecture for Simulation 5 can be seen in Figure 2.

[3] My thanks to David Touretzky for pointing this out to me.

Figure 2

Architecture of Simulation 5



Stimuli.

Training items for Simulation 5 consisted of six items in each of six categories, for a total of 36 items. Inputs were encoded at the featural level, with each feature represented by a group of four input units. This coding was inspired by the feature level coding in the Rumelhart network (Rogers & McClelland, 2004; Rumelhart & Todd, 1993; see pp. 14-17). Items in the same category were more likely to share features among their input representations than items that were not in the same category. Because each feature was encoded using four units, items could have less than a whole feature in common. This means that in addition to sharing whole features, items could therefore share partial features, or be said to partially share features. This gradation of feature sharing was thought to be more naturalistic than the whole feature coding used in the Rumelhart network (Rogers & McClelland, 2004; Rumelhart & Todd, 1993; see pp. 14-17). “Regular” items within a category shared more features, whereas “irregular” items shared fewer features. At output, single units were used to represent category names. Two additional items per category were used for tests of generalization. Of the novel items, one was more regular in terms of the number of units shared with the other members of its category, and one was less regular. A list of words used for training and test items, and images of some of these stimuli, can be found in Appendix B.

It was expected that in a test of this kind, generalization performance would be significantly better in the semi-supervised condition, especially where irregular exemplars were concerned.

Design.

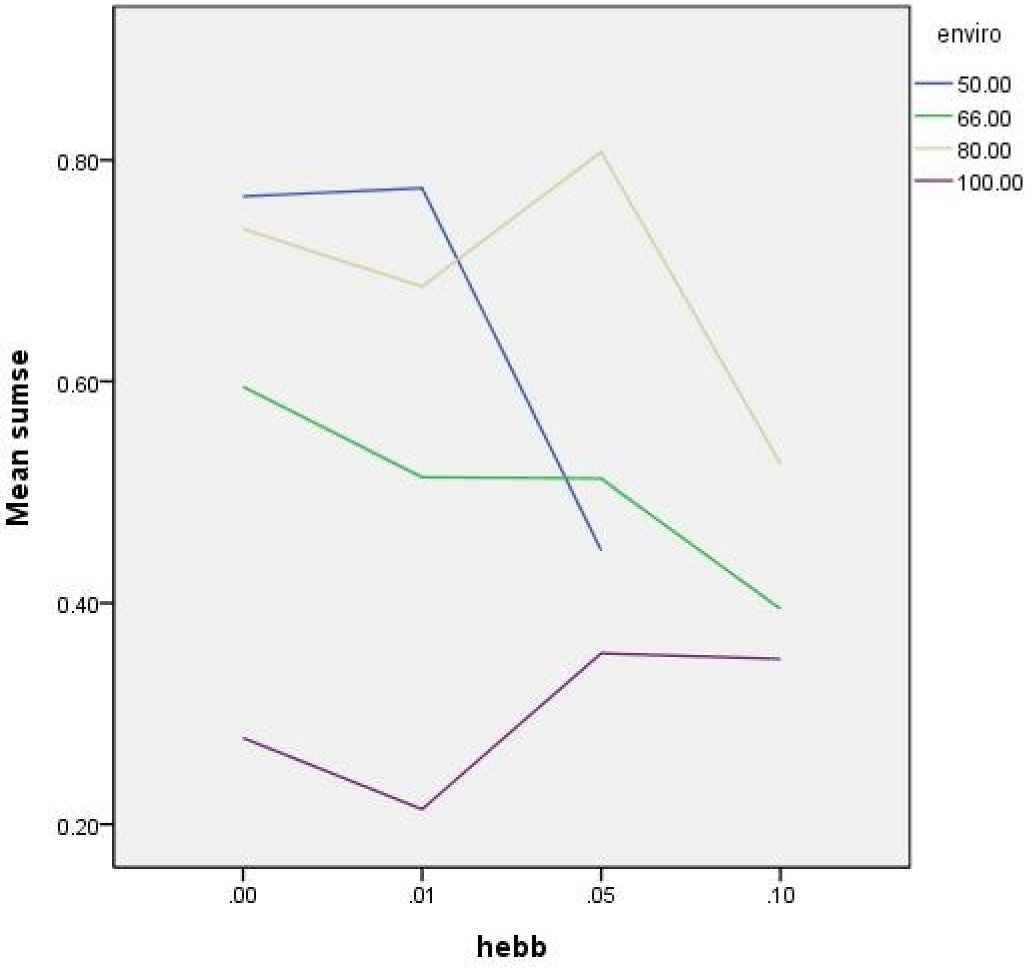
Training duration for this simulation was 20 epochs. Simulation 5 was conducted using a 4 x 4 factorial design with factors being amount of supervision (levels of 50%, 66%, 80%, and 100%) and amount of Hebbian learning ((k_{hebb}) value in the Leabra equation; see Appendix A) (levels of .001, . 01, .05, and .1) and dependent variables being summed squared error on tests of the training and generalization data sets.

Results.

The results of Simulation 5 speak to the interaction of Hebbian learning and amount of supervision in an environment. An ANOVA was used to discover whether the results of this simulation upheld the thesis that networks using semi-supervised learning show similar performance at training and improved performance at generalization compared to fully supervised networks. After training, the network performed correctly on the training set more than 90% of the time in both supervised and unsupervised conditions. No significant differences were detected in the errors of these conditions, *F*(3,1) = 1.371, *p* = .251. Fix degrees of freedom for F to have both numerator and denom values (F is a ratio). Means of summed squared error on the generalization test showed significant differences in performance depending on environment, *F*(3,1) = 30.711, *p* < .01, with significantly better performance at 100% supervision than at the other three environmental conditions (see Figure 3). This means that while supervised and semi-supervised learning showed the same performance on the training set, supervised learning showed better performance on the generalization set. Although the hypothesis for this simulation is thus partly correct, the more important generalization improvement was not apparent in this simulation. Simulation 6 (below) attempted to determine whether the results of Simulation 5 were specific to Simulation 5 or common to networks using semi-supervised learning.

Figure 3

Simulation 5, means of summed squared error on the generalization test



Within the 20-epoch training duration, almost all iterations of the network in all conditions trained to perfect performance on the training data. At this training duration, the interaction of amount of supervision and Hebbian learning value (k_{hebb}) was marginally significant, *F* (9,1) = 1.915, *p* = .051. However, neither the main effects of supervision amount or Hebbian learning were significant (*F*(3,1) = 2.317, *p* = .076 and *F*(3,1) = 2.502, *p* = .06, respectively). This interaction demonstrates that the amount of supervision in the network and the value of the (k_{hebb}) jointly affect network performance during training, where neither may do so alone.

A similar situation obtained in the test of generalization. An interaction between supervision amount and Hebbian learning value (k_{hebb}) was found, F(8,1) = 4.758, p < .01, but the results of Hebbian learning by condition were not consistent. At 50% and 80% supervision, there was a significant effect of Hebbian learning (F(2,1) = 13.215, p < .01 and F(3,1) = 3.334, p = .019, respectively), but at 66% and 100% supervision, there was not. In addition, Tukey's HSD tests on Hebbian learning at the 50% and 80% conditions did not find any patterns in the effect of Hebbian learning on generalization (Table 1). The small differences between means suggest that Simulation 5 was not sensitive enough to detect differences among supervision amounts and Hebbian learning if they do exist. This interaction was investigated more rigorously in Simulation 6 by refining the data set to more precisely control overlap within and between categories.

Table 1

Simulation 5, effect of Hebbian learning on generalization at 50% and 80% supervision

50% supervision		
Hebbian learning value	*q*	*p*
.00-.01	0.01	>.9
.00-.05	0.32	<.01
.01-.05	0.33	<.01

80% supervision		
Hebbian learning value	*q*	*p*
.00-.01	0.05	>.9
.00-.05	0.07	>.8
.00-.1	0.21	>.1
.01-.05	0.12	>.5
.01-.1	0.16	>.3
.05-.1	0.28	>.05

Simulation Six

Rationale.

Given the ambiguous results of some previous researchers (Cozman & Cohen 2006; Nigam & Ghani, 2000; Nigam et al., 2000), and our own ambiguous results (see Simulation 5, pp. 40-42) we must ask: Is it necessarily the case that inclusion of unlabeled data during training improves generalization performance? Do the results of earlier research hold true for connectionist networks?

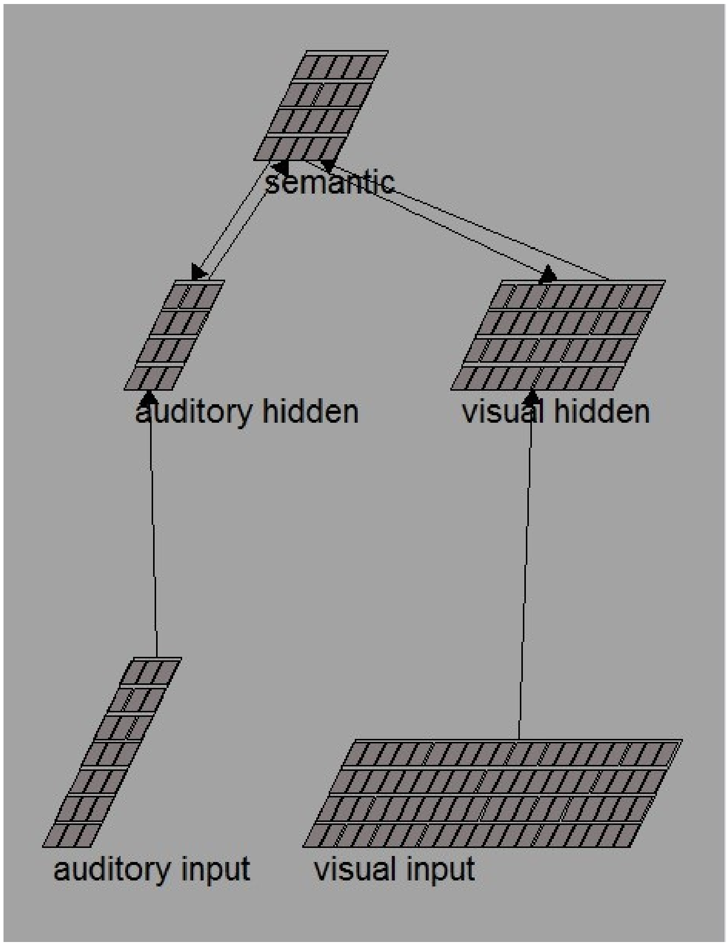
This question was answered with another simulation of category learning. Category learning is an efficient domain for this problem, as it both relates to earlier semi-supervised models, all of which are primarily for classification of data, and ties into a larger theme of the current work, the potential role of semi-supervised type learning in human language acquisition.

Architecture.

Separate input layers for auditory-like and visual-like stimuli were used, along with separate hidden layers. Recurrent connections were used between the hidden layers and the single output layer. The architecture for Simulation 6 can be seen in Figure 4

Figure 4

Architecture of Simulation 6



.

Stimuli.

Input to the network involved feature-level representation of objects from seven categories, such that features were local to groups of four nodes. The objects so symbolized were therefore overlapping and distributed. Hidden layers combined similar patterns and output to a simplified phonological coding of object name, such that there was some overlap in output representation among categories.

These training sets allowed us to explore learning with differing amounts of Hebbian learning in environments with differing amounts of labeled targets. However, direct comparisons could not be made among the training environments because their error terms are on different scales. Therefore, generalization performance, in addition to being an important component of any discussion of semi-supervised learning, is also the best method for directly comparing the effects of differing amounts of labeled data in learning environments. A new environment that is the same for all networks can be constructed and used in generalization testing.

A separate environment consisting of two exemplars from each category that did not appear in any training set was used to test generalization. The same environment was used for all tests. The relevant dependent variables were summed squared error and settling speed, a measure of how long the network needs to decide how to classify an item.

Design.

In Simulation 6, a 6 x 5 factorial design was implemented, crossing number of labeled items (by percentage, 25, 33, 50, 66, 75, and 100) of the 12 items from each category used for training, and the amount of Hebbian learning included in the implementation of the Leabra algorithm ((k_{hebb}), with values of .000, .025, .050, .075, and .100) (for an explanation of Leabra, see Appendix A or O'Reilly, 2001; O'Reilly & Munakata, 2000).

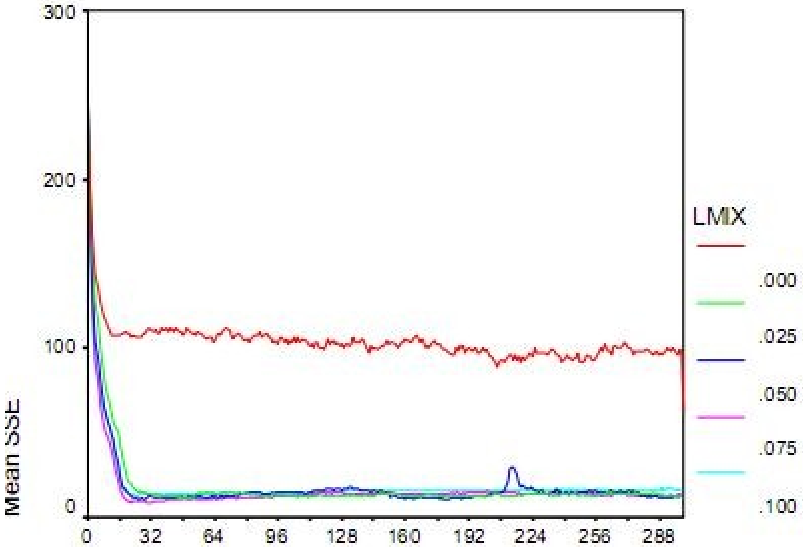
Results.

We can begin by examination of learning curves for these combinations of labeling amount and Hebbian mixture, shown in Figures 5-10. Figure 5 shows the base example, learning from 100% labeled data – the traditional modeling paradigm for a Leabra network, and a replication of O'Reilly's

(2001) key result. With .000 Hebbian learning, the network settles at a mean summed squared error of approximately 100, with much of the decrease from a starting value (at epoch 0) of close to 300 occurring within the first 20 epochs of training. In contrast, networks with Hebbian learning of .025 or greater settle at or close to a summed squared error of 0 within 30 epochs.

Figure 5

Simulation 6, 100% supervision

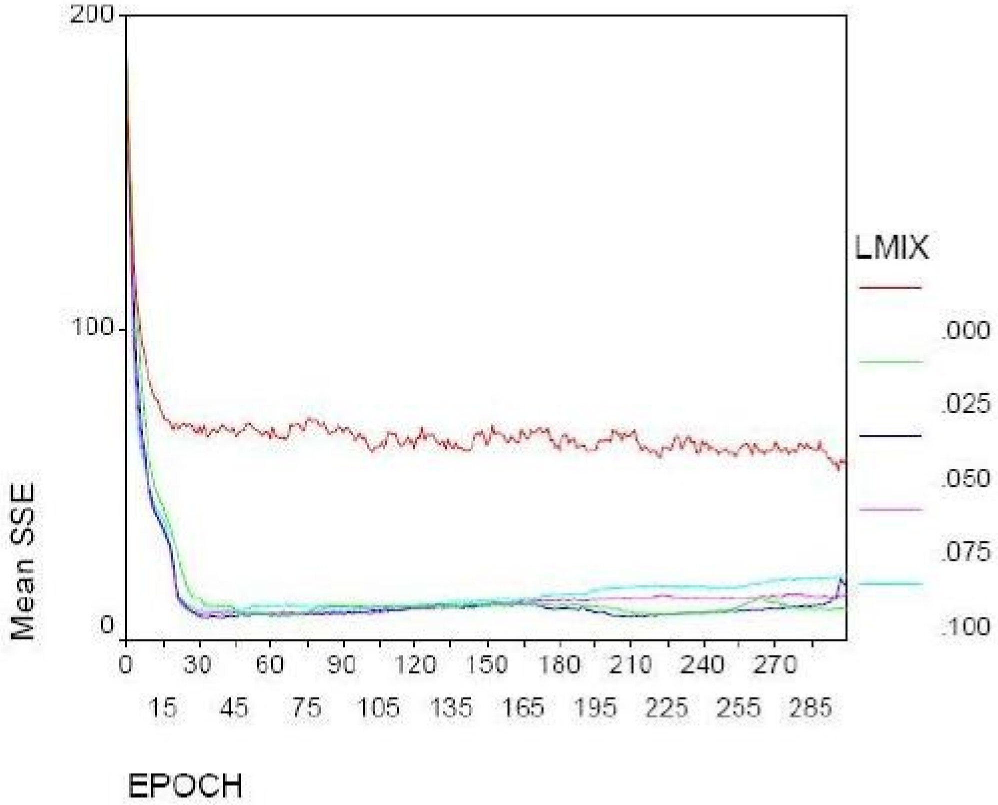


These results change as we remove labeling from training items. Two things must be kept in mind while considering these curves. First, this procedure is quite different from that of the algorithms applied to earlier semi-supervised data classification problems. Those problems start with large amounts of unlabeled data to which labels are assigned laboriously by hand. In contrast, the current network is being compared to those connectionist networks that traditionally use only labeled data. Here, our ultimate goal is to learn something about how human cognition copes with partially labeled environments. Second, performance here is measured by the summed squared error statistic. Error terms can be calculated only for those items that are labeled, because without a labeled target the backpropagation algorithm cannot compute the difference between the generated output and the target. Therefore, the different training environments cannot be compared directly because their error, in effect, is on different scales. This problem will be addressed in Simulation 7.

The curve for learning from 75% labeled data, shown in Figure 6, is similar to the 100% curve in that with Hebbian learning (k_{hebb}) equal to 0 the training set is never learned, whereas for all the other values of Hebbian learning, it is learned in relatively few epochs. However, as training progresses, the error values increase slightly and finally separate at approximately 135 epochs. The lowest error is seen with Hebb values of .025 and .050, with .075 being higher and .100 having the most error of this group.

Figure 6

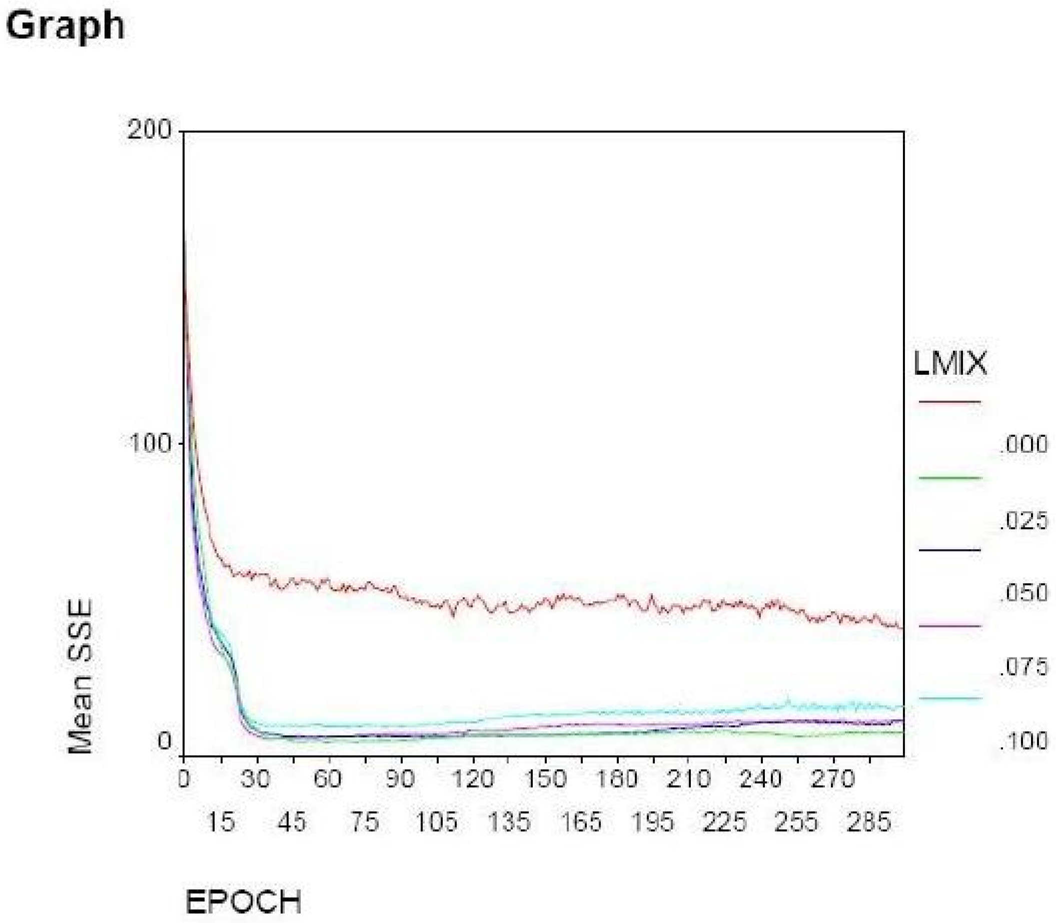
Simulation 6, 75% supervision



When 66% of the training examples are labeled, the results are very similar to those of the 75% curves, with the exception of the curve for a Hebbian learning value (k_{hebb}) of .100. From the time when the learning curves asymptote, the .100 curve shows a higher summed squared error, with progressively more “jitter” as learning continues. At this point we begin to see the ways in which the Hebbian learning value influences learning (Figure 7). Whereas too little Hebbian learning in the algorithm leads to failure to learn the environment, too much can result in a decline in performance at higher training durations. In human learning, this might result in the inability to make generalizations based on prior learning. All changes in knowledge would have to be memorized. Too much Hebbian learning would result in too much generalization, such that the same behavior would be applied to all situations. These might be analogous to the behaviors expressed in some psychopathologies, though it is currently unknown whether any such behaviors result from failures in learning at the neuronal level.

Figure 7

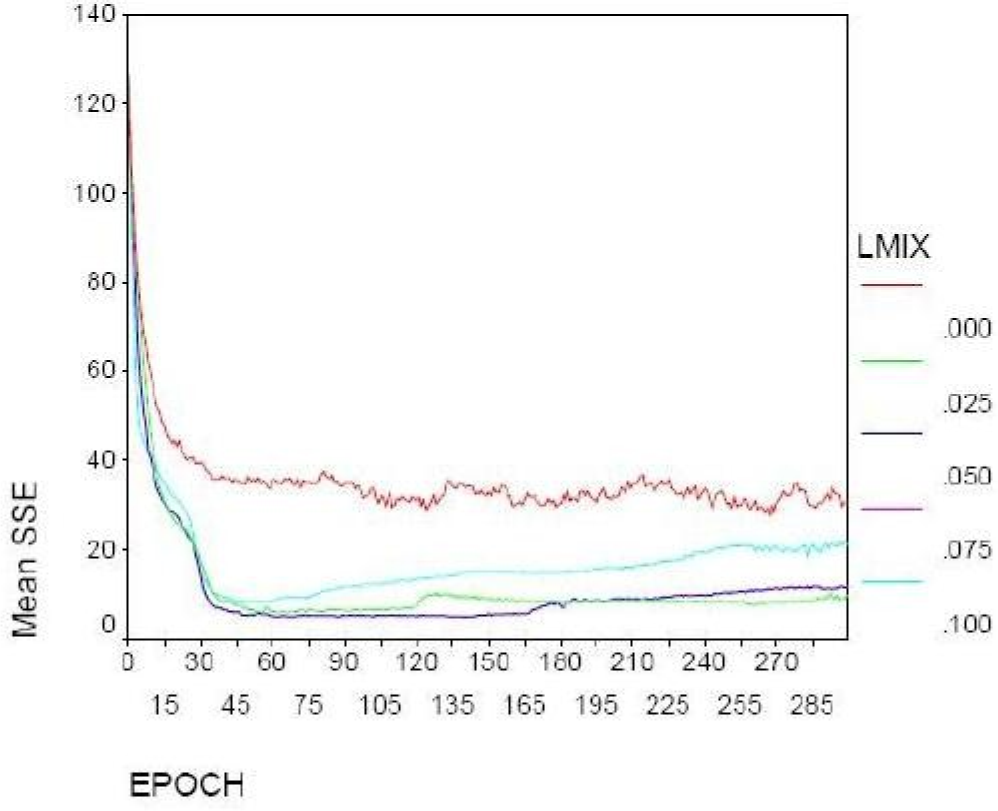
Simulation 6, 66% supervision



At 50% labeling, the differences among the learning curves become greater. Learning is most successful at Hebbian learning (k_{hebb}) values of .025 and .050 and less so for other values (Figure 8). The lower starting error in these and subsequent curves is related to the loss of labeled targets by which error can be computed. (The response to an item without a labeled target is never counted as "incorrect" by the algorithm.)

Figure 8

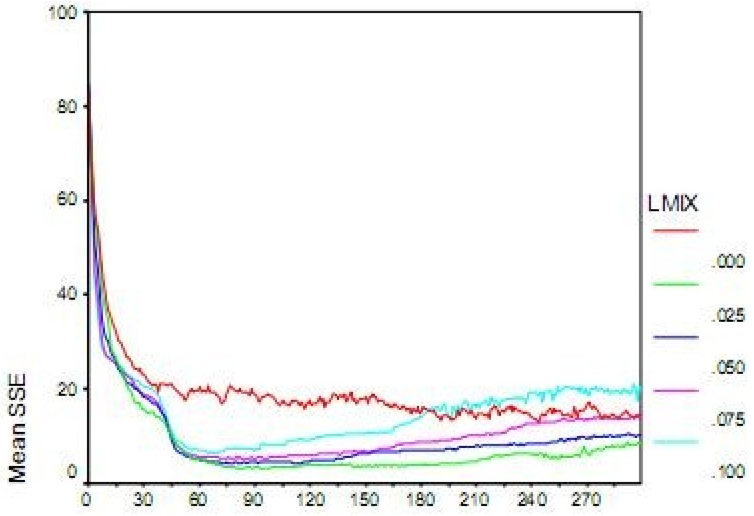
Simulation 6, 50% supervision



The relationships among the curves become more complicated when labeling is as low as 33% (Figure 9). By 195 epochs, the curve for the Hebbian learning (k_{hebb}) value of .000 is no longer the worst performer. Instead, summed squared error is greatest for a Hebbian learning value of .100. It seems that it is possible, with few enough labeled training exemplars, to have too much Hebbian learning in the algorithm. It is not the case, however, that omitting Hebbian learning can compensate for the absences of larger amounts of labeled data. A Hebb value of .025 still gives the best performance. Unlabeled data can be used for learning only when some form of unsupervised learning is applied. Therefore, in this instance, we are seeing that learning only from labeled exemplars – merely a third of the training data – provides a better estimate of the data distribution than learning from labeled and unlabeled exemplars when reliance on Hebbian learning is too great.

Figure 9

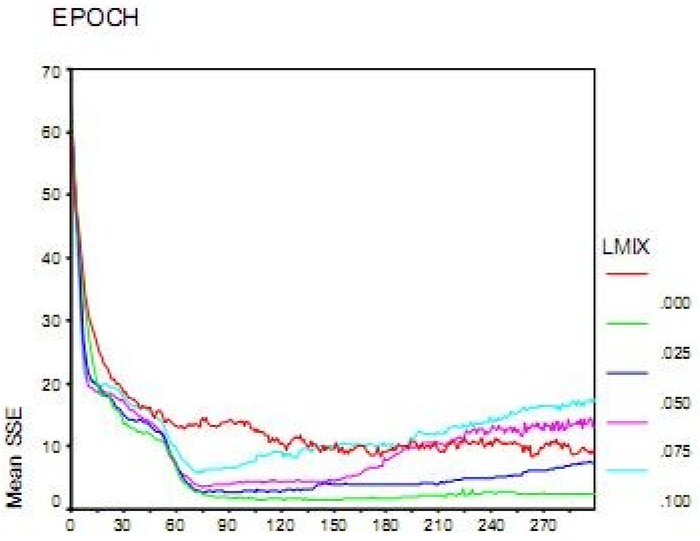
Simulation 6, 33% supervision



Finally, at 25% labeling (Figure 10), networks with Hebbian learning values of both .100 and . 075 finish training with summed squared errors higher than that of the network with .000, showing a stronger degree of the same phenomenon observed at 33% labeling. However, the best performance is still shown with a Hebb value of .025. Training was not conducted with 0% labeled exemplars because the type of learning under scrutiny, task learning, requires supervised learning (and hence, labeled data, and hence, targets).

Figure 10

Simulation 6, 25% supervision

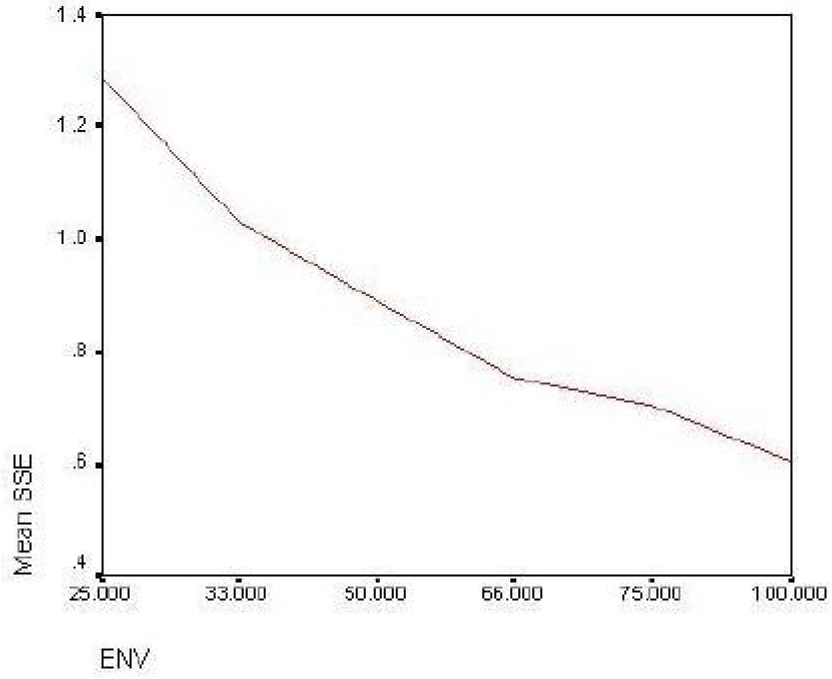


In the generalization task, summed squared error was significant for both environment and Hebbian mixture; cycles to settle was significant only for Hebbian mixture. The interaction between environment and Hebb mix was not significant for either variable, confirming the results outlined above. The best values of Hebbian learning were similar for all environments.

Multiple comparisons using Tukey's HSD showed that the only significant difference in summed squared error among supervision amounts was between 25% and 100%, where q = .683 with $p < .05$. Figure 11 shows summed squared error at test by supervision percentage collapsed across Hebbian learning values (k_{hebb}). The Hebbian value of .000 was significantly different from all other values for both dependent variables.

Figure 11

Simulation 6, summed squared error at test by supervision percentage collapsed across Hebbian learning values



Results consistently showed that for any environment, the best generalization was found with a Hebbian learning (k_{hebb}) value of .025. It is also the case that the best generalization overall was found with the fully labeled training environment, which contradicts the results of Simulations 1-4. However, for Hebbian values greater than 0, differences in generalization are not significant from environment to environment. This result is important because it shows roughly equivalent learning among environments with large amounts of labeled training data and those with low amounts. It might be tempting to dismiss this result as a curiosity. If the best learning – the best generalization – is achieved from fully supervised environments, why pursue semi-supervised learning at all?

Conclusions.

The training set results show that the amount of Hebbian learning in the algorithm (k_{hebb}) affects observed learning over time, though in a more complex manner than originally predicted. For these simulations, it was hypothesized that better performance on less supervised data sets would be seen with greater values of Hebbian learning. Instead, a relatively small value of Hebbian learning, .025 or .05, is best for all percentages of labeling in the data set. This is addressed by Cozman and Cohen (2006) in their discussion of classifier degradation by inclusion of unlabeled data. In expectation-maximization classifiers (see p. 25), the asymptotic bias of the estimator trained on labeled data may be different from that of the estimator trained with unlabeled data. In other words, networks trained on these different exemplars perform as if they were learning different distributions of data. Inclusion of too much unlabeled data produces an incorrect estimate of the actual distribution. A similar phenomenon may be occurring in the connectionist network under discussion. Inclusion of too much Hebbian learning in the Leabra algorithm may give too much weight to the constant amount of unlabeled data. It is possible that such incorrect weighting gives rise to the temporary overgeneralization of word endings seen in children at a certain stage of language development (saying “gived” instead of “gave”, for example) or the undergeneralization of morphology seen in specific language disorder.

The Goal of Semi-Supervised Learning

If we are seeking to understand human cognition, we must be able to suit our models to the environmental conditions of human learning. Unless we want to argue that humans learn from environments that are universally fully supervised, we must make a better attempt to capture realistic learning environments in our models. This will help us bring to light the flaws in our current conceptions of human learning. It is important do this, as at least some human learning may be semi-supervised.

Simulation 7

Rationale.

Simulation 7 marks an attempt to optimize – or at least to find the relevant patterns regarding – class and feature learning in a semi-supervised network. This represents an attempt to understand the role of multidimensionality of stimuli in a learning environment.

In the results of Simulation 2, the performance of the semi-supervised network was superior to that of the fully-supervised network. The Simulation 2 network called for a horizontal-vertical line discrimination, but one in which each unit was equally likely to participate in either kind of line. Only the dependencies among the units active in an item contained the information necessary to make the discrimination. This is suggestive because it points to the power of Hebbian learning. It is also true that in many learning environments it is not necessary to force this reliance. Even a fully supervised network can make use of Hebbian learning when it updates weights with the Leabra algorithm. This increases a network's power in learning and particularly in generalization. My results have followed those of O'Reilly (2001; O'Reilly & Munakata, 2000) on this point. However, in an environment like that of Simulation 2, in which enacting supervised learning on the individual features of an item does not lead to correct learning, forcing the network to rely on Hebbian learning significantly improves performance. “Forcing” reliance on Hebbian learning simply means that when an input item is untargeted, weight updates made in response to that item are made solely on the basis of Hebbian learning; the error-driven term in the algorithm goes to zero. The co-active units are “tied together” more strongly than they would be if the error-driven learning were present to work against these co-occurrences.

Were the weight updates of the entire sequence of training described in a single equation, the co-active units would be weighted more heavily than singly active units. To speak of it in psychological terms, this forcing highlights, or makes more salient, co-occurring sets of features. It can be predicted, therefore, that when co-occurrences of features provide information that single features do not, we will see the benefits of semi-supervised learning. These benefits will be seen more strongly in those data sets in which the co-occurrence of features is more important for correct learning than when it is not. In Simulation 2, what the network learned from error driven learning contradicted what it learned from Hebbian learning. Semi-supervision provides a mechanism by which error driven learning can be given less weight.

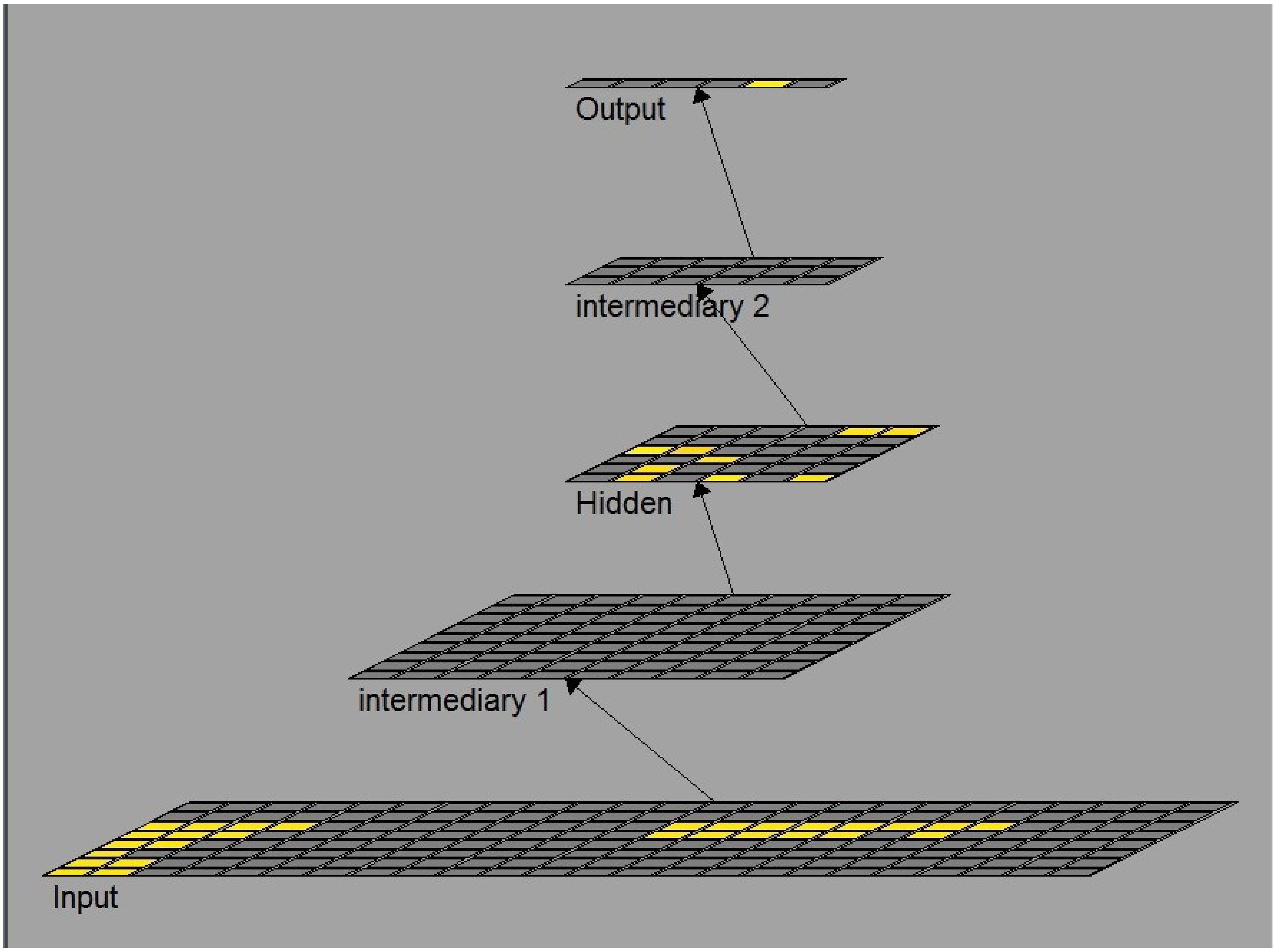
Specifically, in the simulation outlined below, the largest improvement of semi-supervised over fully supervised learning should be seen in the data set with the greatest amount of interclass overlap in the greatest number of classes. In these sets, feature co-occurrence will be the single most important factor in obtaining correct network performance. In the remaining data sets, except for the control, semi-supervision and full supervision are likely to be functionally equivalent; in the control data set semi-supervision is likely to hurt performance.

Architecture.

Simulation 7 used the simplest architecture of the language-like simulations (Figure 12). A single input layer was connected in a feedforward manner to a single hidden layer, which was connected in a feedforward manner to the single output layer. The simplicity of the architecture, however, belied the complexity of Simulation 7's stimuli.

Figure 12

Architecture of Simulation 7



Stimuli.

In the previous simulations, the dimensionality of the environments varied greatly, from the two dimensions of Simulation 1 to the 48 dimensions of Simulation 6. Number of dimensions in and of itself does not determine learnability, as we can easily see by comparing the results of two-dimensional Simulation 1 to those of the also two-dimensional Simulation 2. The great difficulty of the Simulation 2 task involves the relationship of the horizontal and vertical dimensions, and the low predictive value of any active unit in the simulation. In this very difficult task, the potential of semi-supervised learning first became apparent. These dimensions were completely orthogonal in a literal sense, with no *conceptual* overlap between them. However, in their instantiation in the network, they were completely overlapped, existing in the same input space, with every node equally likely to participate in the representation of both horizontal and vertical lines. The horizontal-vertical distinction, for any input item, could only be captured in in the dependencies among multiple units – which units were jointly active in a given input item.[4] In this simulation, neither fully supervised nor semi-supervised learning completely learned the task. However, semi-supervised learning performed significantly better on a generalization test involving the correct classification of novel patterns. The inference is that semi-supervised learning is important not only where stimuli are multidimensional, but where those dimensions exist in certain relationships.

Multidimensionality is a common feature of stimulus sets, and can be pre-established or discovered after the fact, depending on the way one chooses to define the dimensionality of the given network. Unless otherwise stated, the dimensionality referred to here will be the dimensionality of the input or environment, as it is these environmental characteristics that are manipulated when semi-supervised learning is introduced. Dimensionality is enumerated here as the number of unique input characteristics regardless of the nodal arrangement. Therefore Simulations 1 and 2 have two dimensions, though both have 16 input nodes; Simulation 6 has 48 dimensions, but 192 nodes in the input layer. All dimensions here are "multi" – all the simulations take inputs of two or more dimensions. The number and relationships of these dimensions may be under scrutiny; their existence is not. Hebbian learning, crucial to the semi-supervised learning paradigm, has no meaning for a one-dimensional input, just as error-driven learning has no meaning without target outputs. Not every input stimulus instantiates each dimension. If we were to describe the layer in vector notation, a 0 would be

[4] My thanks to Dave Touretzky for pointing this out to me.

entered for the dimensions not instantiated in each input's vector.

Data sets for this simulation had differing numbers of classes. The numbers of classes used were three, as a very small number that is larger than two (as Simulations 1 and 2 were two-class networks); seven, as this matches the CELL model (Roy & Pentland, 2002) that provided the original starting point for Simulation 6, of which this is a modification; 14, as being twice as large as seven; and 28, as being twice as large again. These numbers provide a good range without reaching an unwieldy number of classes, and their selection is designed to allow patterns to be detected. An "optimum" number of classes may not be detected. However, looking for an optimum number of classes may not be a reasonable goal. Presumably, the brain parses the world into useful and fairly accurate numbers of classes; no upper limit on this number has been determined.

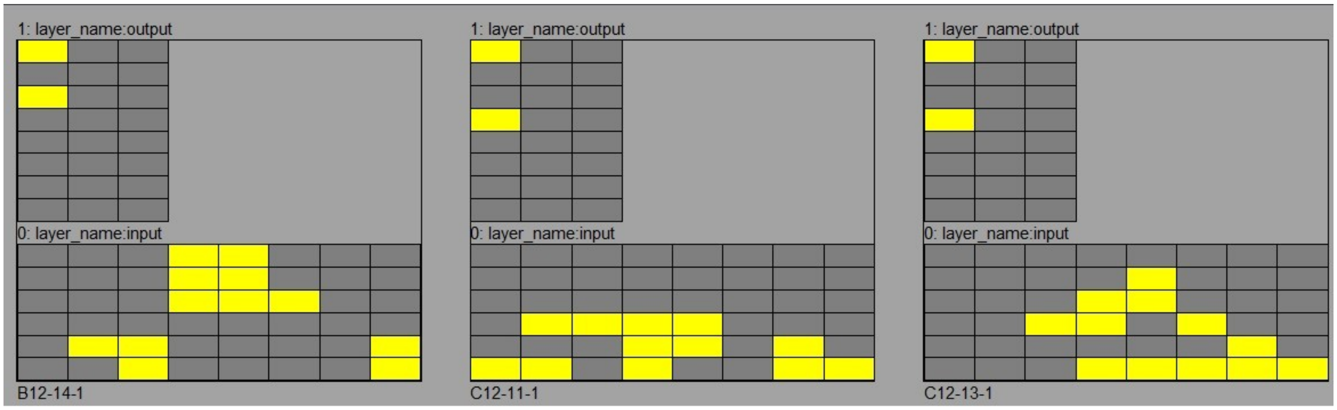
Within these numbers of classes, features were be manipulated in a variety of ways. First, a control data set was generated, having one sufficient feature per class (and therefore no inter-class overlap). This set provided a basis for comparison with other data sets, though no conclusions were drawn from the control itself.

For the other data sets, a "prototype" (in quotation marks to differentiate the use of a prototypical example here from any endorsement of prototype theory) for each class was created, consisting of three, six, or 12 features. Each exemplar in the class differed from the prototype by one, three or six features. The prototypes in the different classes overlapped by one, three, or six features. In the data sets, the features deviating from the class prototypes were rotated to create varying amounts of inter-class overlap. All of this, simulated in both fully supervised and semi-supervised paradigms, summed to a large number of data sets, but it offered a sound space in which to explore issues of feature, class number, and overlap. The question was therefore be formulated as follows: Given a large number of items that belong to *k* categories, when the items are defined in *d* dimensions, and the items within each category share *s* dimensions, where $s < d$, what are the advantages or disadvantages of semi-supervised learning in learning of categories; and more specifically, how do these advantages and disadvantages vary with the magnitude of correlation among item features between categories?

One fully labeled and two partially labeled data sets were constructed for simulation 7. It is possible to match a partially labeled data set to a fully labeled one on the basis of number of labeled items (in which case the partially labeled data have more total items) and on the basis of number of total items (in which case the partially labeled data have fewer labeled items). For simulation 7, both kinds of matches were constructed. These data sets will be referred to below as ssl-a and ssl-b, respectively. The semi-supervised data sets were equated with each other on the number of unlabeled items. These data sets were intended to represent linguistic stimuli, in that they were multidimensional, distributed, and varied around particular features (as, for example, a phoneme may vary based on the phonemes preceding or following it). Examples of the Simulation 7 stimuli can be seen in Figure 13.

Figure 13

Examples of stimuli from Simulation 7



Design.

For all data sets, the network was trained with .025 Hebbian learning incorporated in the Leabra algorithm, and a learning rate of .01. A 3 x 3 x 3 x 3 factorial design was used, with factors of number of features, within-category difference, between-category difference, and data set as outline above.

Results.

In performance on the training set, measured by summed squared error after 3000 epochs of training, significant differences were found among all 3 data sets, $F(2,1) = 51.129$, $p < .01$. Significance values for all three pairwise comparisons were less than .01 (see Table 2). The best performance was seen with the ssl-b data set, with fewer labeled items than the ssl-a and fully labeled data sets. Here there is not only a beneficial effect of semi-supervised learning, but of a greater percentage of unlabeled items in the data – which was not always the case in Simulation 6. Differences among number of categories were also significant, $F(7,1) = 918.144$, $p < .01$, as was the interaction term, $F(14,1) = 31.409$, $p < .01$. In most cases, summed squared error increased as number of categories increased, but this was not the case for all conditions. Just as unlabeled data made a difference in Simulation 6, so it makes a difference in Simulation 7, a difference that may be dependent on the dimensionality and overlap of the information in the training data.

In a reversal from some earlier simulations, semi-supervised learning was not beneficial to generalization performance, measured by summed-squared error on a set of novel exemplars from the same categories as the training data. Analysis of variance over the data sets was not significant, $F(2,1) = .397$, $p > .5$. The conditions for number of categories was significant, however, $F(7,1) = 529.220$, $p < .01$, with summed squared error increasing as number of categories increased. The interaction term showed a trend toward significance, $F(14,1) = 1.679$, $p = .053$ (see Table 2)

Table 2

Simulation 7, pairwise comparisons for training performance by data set.

contrast	*q*	*p* value
fully supervised - ssl-b	.136	< .01
fully supervised - ssl	.084	< .01
ssl-b - ssl	.052	< .01

.

Conclusions.

From this analysis of Simulation 7, we can see that environmental dimensionality is a consideration in evaluating a semi-supervised paradigm. It is interesting that thus far the benefits of semi-supervised data sets are apparent in training set performance rather than generalization. Even here, however, better performance (lower summed squared error) is seen in conditions with fewer categories. More testing is necessary to determine the exact cause of this result; a recurring necessity when so many variables are involved.

Chapter Six

Integration and Future Directions

The most important point of correspondence between the semi-supervised model and empirical studies of lexical acquisition is in environmental input conditions. Both the Rumelhart (Rogers & McClelland, 2004) and SUSTAIN (Love et al., 2004) models meet some conditions of ecological validity; the semi-supervised model meets a criterion that the others have yet to demonstrate, that it learns from both labeled and unlabeled data.

Behl-Chadha (1996) has shown that infants as young as 4 months of age can form categories of visual representations of artifacts in the absence of naming. This is a parallel to unsupervised learning in both the semi-supervised and the SUSTAIN (Love et al., 2004) model. However, once a child has learned about 50 words, they will form such categories only during naming conditions (Smith et al., 2002, cited by Landau, 2004). The SUSTAIN model does not account for this type of change, as it does either supervised or unsupervised learning, but never both. The semi-supervised model, on the other hand, could make this change smoothly. The findings suggest a state change in the nature of human learning, from unsupervised to supervised learning, a change that has not been explored in the semi-supervised model, but easily could be.

It is suspected, however, that rather than a change in the way learning occurs at the neuronal level, what changes is the way that infants use the cues in their environments. The change seen by comparing Behl-Chadha's (1996) results to Smith et al.'s (2002) results is closely paralleled by the infant's switch from reliance on perceptual cues to social ones between 10 and 19 months of age, and that this change is gradual (Hollich, Hirsh-Pasek, & Golinkoff, 2000, cited by Hirsh-Pasek et al., 2004). This could be captured in the semi-supervised model by a gradual decrease in the proportion of Hebbian learning used in the weight change equation. Because reliance only on supervised learning is unlikely in terms of biological plausibility (see O'Reilly & Munakata, 2000), the mixed-model learning of an equation like Leabra is the best way to capture the neural-network level processing that underlies word learning.

It could be argued instead that feedback is always available to learners, and sometimes they use this feedback and sometimes do not. For example, it is highly unlikely that when infants are four months old, their environments undergo a universal radical shift to include object naming all the time whereas prior to four months object naming is never present. Rather, it is probable that objects are named in child-directed speech some of the time at all ages, though the amount of object naming may increase as infants become more able to direct or to follow caregiver attention. This is what we see in studies of object naming and shared attention (Carpenter, Nagell, & Tomasello, 1998; Goodman, Dale, & Li, 2008; Hoff & Naigles, 2002, cited by Woodward, 2004). In fact, different kinds or levels of feedback can be characterized on the basis of these studies: no feedback, naming feedback, naming and joint attention feedback. From this point of view, although the semi-supervised model uses the most complete representation of the environment to date, it too underestimates the varieties of input available to the infant word learner.

The most direct way to simulate non-use of feedback by a learner is to exclude labeling on some items in the model. This was done in the semi-supervised model under discussion. However, this method may mistakenly attribute to the environment a condition endogenous to the learner. In other words, there may be times at which feedback is absent from the environment, and others at which feedback is available to the learner but is not used. A minimum of two approaches is necessary to explore this question. First, the attentional conditions of word learning must be studied in more depth. Habituation and joint-attention paradigms are appropriate for this type of study. Event-related potential methods would be even better, if infant attention can be measured adequately in this fashion. The goal of such studies would be to determine some proportion of naming events that are not attended to by infants. Some such details can be inferred from Carpenter's (et al., 1998) study of caregiver-infant joint attention, in that in dyads with higher frequencies of joint attention, infants displayed larger vocabularies and larger gains in vocabulary than in those with lower frequencies of joint attention. This is not enough detail, however, to fit a model, so a more precise enumeration of instances of caregiver-infant joint attention and its effects on lexical acquisition is required. Secondary to such findings, a more elaborate connectionist model that explicitly incorporates attentional conditions would be of benefit, as such a model would allow parsing of the effects of attention, feedback, and synaptic weight change in word learning.

Whether variables are explored mathematically as in a structural equation model, or implemented in a connectionist network, modeling is becoming a more important endeavor as the environmental and cognitive conditions for lexical acquisition are becoming better understood. A multiplicity of factors makes experimental control more difficult, especially in this domain, which is already hard to study in a non-natural setting. The careful and creative use of models will help make explicit sense of the many variables involved, and allow to simulate holding variable constant that may be opaque to manipulation in the natural environment and able to be manipulated only over short durations in the laboratory. In these circumstances, models can increase our understanding of a cognitive domain and help us make better predictions about the interacting effects of multiple variables.

The first point is that, when implemented with the Leabra algorithm, semi-supervised learning meets O'Reilly's (1998) principles for biologically based modeling and is biologically plausible to the same degree as fully supervised models in Leabra. This may seem intuitive and, hence, insignificant, but its importance should not be discounted. The central goal of this particular modeling endeavor is to develop a more ecologically valid technique for modeling the environmental conditions under which language is learned. If this goal were achieved, but biological plausibility discounted, superiority of the semi-supervised over the fully supervised model could not be claimed, as it too would capture only half of the conditions of lexical acquisition. If we ignore what we know when developing a new model, the science has stood still even though the information has changed. It is therefore important to demonstrate that the most biologically plausible of learning algorithms is compatible with an alternative paradigm for the environmental context of learning.

Second, we see that in some circumstances, semi-supervised learning gives better generalization to novel exemplars than supervised learning. The human capacity for generalization is not only remarkable, but often poorly captured by supervised networks. An improvement in a model's ability to generalize is therefore a key factor in the preference to model lexical acquisition with semi-supervised learning. It also shows that the findings of other researchers on semi-supervised learning hold true when the paradigm is implemented with a biologically plausible algorithm for a human-like cognitive task. Even the course of generalization ability over time can be interpreted in a developmental framework. Generalization ability gradually improves over the course of training, and then decreases, indicating a narrowing or refining of the items identified by a lexical item. Of course this interpretation is, for now, only hypothetical, as a more refined model would be needed for a precise attempt to predict the time course of generalization ability over development in humans. However, it is provocative, as such a success would indicate a more accurate model of the influence of the environment on the learner than has previously been seen.

Third, in addition to the predicted better generalization performance, better training set performance has also been seen with semi-supervised learning. This result should be interpreted cautiously for Simulation 4, as the semi-supervised and fully supervised training sets were essentially on different scales. However, this problem was addressed in Simulation 7, and it was found that training performance was better with the semi-supervised data set than with the fully supervised set, with the best performance given with the semi-supervised set that was matched to the fully-supervised set on total number of items.

Simulations Five and Six

Taken together, Simulations Five and Six show that the learning paradigm of a network, and specifically the amount of labeling, interacts with other parameters, including the regularity of the data set, the duration of training, and the proportion of Hebbian learning in the weight-change algorithm. Apart from any other analysis of the results of these simulations, two cautions can be taken. The first is that modeling with the semi-supervised paradigm expands the number of degrees of freedom the researcher must handle. This can be of benefit, as the real-world learning environment may more precisely be matched, but it also increases the chance of any single result being found in error, and opens the researcher to charges of finding effects that are unimportant or nonexistent. Second, expanding this line of thought beyond computational modeling and into considerations of human learning, it may be the case that in order to fully understand lexical acquisition, it may be necessary to consider more variables than are usually studied.

Simulation Seven

The interesting thing about Simulation Seven is that the best performance is seen when the network has the least amount of information available. It is possible to speculate that our language-learning abilities specifically take advantage of a limited amount of information in the environment, and would in fact show a decrement in learning if more information were available. Furthermore, it gives rise to the speculation that some kinds of cognitive impairments, such as specific language impairment, are the result of a brain that relies too heavily on supervised learning.

Generalization in CELL and Simulation 7

A direct metric for comparing the Simulation 7 model to the CELL model is impossible. Simulation 7 was designed to identify differences in performance among the three data sets. No differences between supervised and semi-supervised generalization ability were found. On the other hand, the CELL model compared unsupervised cross-channel learning to unsupervised single-channel learning. Although the CELL model speaks to the utility of unsupervised learning in language learning domains, its function is essentially different from that of the Simulation 7 model. Therefore, any attempt to make a direct comparison between the two is futile.

If we operate from the premise that neither supervised nor unsupervised learning alone is sufficient for lexical acquisition, then semi-supervised learning seems a not unreasonable compromise between the two. However, only Simulation 7 and the other simulations discussed above can make any assessment of this compromise. The possibility of a comparative model, with a semi-supervised robotic system learning words from real-world data, is intriguing.

Parameters for Semi-Supervised Learning

This exploration of semi-supervised learning has served to show the promise of semi-supervised learning as a paradigm for connectionist modeling. We have seen how the learning environment is inadequately captured by the more typical learning paradigms of supervised and unsupervised learning, and how in environments with little labeled data, incorporating the ability to learn from unlabeled data can improve learning and generalization. We have seen that semi-supervised learning can be applied to biologically plausible connectionist learning algorithms (Leabra), and that in such cases we find improvements in learning and generalization, dependent upon the dimensionality of the feature space, the overlap among items, the balance between Hebbian and error-driven learning in the algorithm, and the ratio of labeled to unlabeled data.

These parameters should not be looked upon as qualifications upon semi-supervised learning's success, but rather as a description of the conditions under which this paradigm will be particularly useful. It may not be appropriate to apply semi-supervised learning indiscriminately to every learning situation. However, neither fully-supervised nor unsupervised learning are universally appropriate either. The correct pursuit of modeling should involve the careful selection of a learning paradigm based upon the learning task, the cognitive process, and the characteristics of the environment when the task under study is learned by humans. It is the latter parameter that has so often been neglected by cognitive scientists, but we have now seen how environmental characteristics can indeed make a difference to the performance of a model.

Final Thoughts

We speculate that semi-supervised learning is an important component of the infant's language learning ability: Joint-attentional situations constitute labeling in child-directed speech. Because joint attention occurs relatively infrequently, language learners must also use unlabeled data in their learning process. Semi-supervised learning provides a substitute for this learning from both labeled and unlabeled linguistic data.

This speculation is supported by the finding that training with unlabeled data improves human classification performance (Zhu et al., 2007) and the results of our Simulation 7. Though Simulation 7 did not show the predicted improvement in generalization performance compared to the fully-supervised model, it did show an unexpected improvement in training set performance. This suggests that the semi-supervised model was better able to learn the observed data distribution than the fully-supervised model. This simulation was designed to be languagelike in some specific ways. The relative success of the Simulation 7 model allows us to strengthen our hypothesis that language is a domain that makes use of semi-supervised learning. Based on the work of Singh et al., (2008) who have characterized the domains for which semi-supervised learning confers an advantage over supervised learning (p. 29), we may speculate that language fits these characteristics.

Proposal for Empirical Studies

This paper has addressed the theory of applying semi-supervised learning to connectionist cognitive models. While we have discussed the biological plausibility of semi-supervised learning, at least when applied with biologically plausible learning algorithms, it remains true that there has not yet been a study to establish definitively whether humans learn in a semi-supervised fashion. It has been implied by a number of researchers, and would go some way toward solving some of the complex learning problems with which humans are faced, those in which the degree of supervision in the environment seems insufficient for correct learning. However, implication and logic are not a substitute for empirical research. Therefore, I close with proposals for two experiments to address semi-supervised learning in humans.

Learning multidimensional environments.

First, we have seen the degree to which the success of semi-supervised learning depends on a multidimensional environment in which features correlate in complex ways. We saw this in a preliminary way in Simulation 2, and to an even greater degree in Simulation 7. Therefore the first experiment will look at learning the characteristics of a complex multidimensional environment, one in which the featural dimensionality can be precisely described. This experiment will use as stimuli a large number of pictures of flowers. This is ideal for a number of reasons. First, flowers are a natural kind. There is no need to create, perhaps badly, artificial stimuli that may not vary along a sufficient number of dimensions. Second, with the exception of those persons who are expert in botany or gardening, most kinds of flowers are, simply, flowers – an average person will not be able to recognize and name the majority of flower species available to choose for stimuli. If asked to name the stimuli, people will tend to call all of them “flowers” as opposed to the pseudo-novel stimuli sometimes used in similar learning experiments, which often resemble particular animals and could be grouped on the basis of those familiar kinds. Third, it is easy to manipulate photographic images of flowers to create stimuli with specific characteristics – photographs can be cropped, enlarged, or rotated, the colors changed, etc. – so categories can be created on the basis of any number of specifically chosen characteristics. Finally, as long as the subjects are not flower experts, the labeled/unlabeled status of each stimulus can be manipulated, which is critical for any study of semi-supervised learning.

This experiment will follow Simulation 7 closely. One of the values of cognitive modeling is its predictive ability; therefore, we will use the simulation already completed to predict the results of the experiment. Quite simply, semi-supervised learning in the sense of learning from labeled and unlabeled exemplars will produce in this experiment's human subjects better generalization, that is, better naming of examples not presented during training; and these results will vary according to the dimensionality and overlap of categories and category exemplars.

Semi-supervision in lexical acquisition.

The second experiment is more to the point of language learning, language being a cognitive process that has been cited as requiring both supervised and unsupervised learning (Bloom, 2000). In language learning, we know that when an object name is used, caregiver and child place joint attention on the object 70% of the time. This gives us an environmental substrate for supervised learning and an obvious labeled-to-unlabeled ratio. An excellent study already exists exploring the relationship of joint attention and language learning (Carpenter, Nagell, & Tomasello, 1998; see Chapter 1 for a description of this experiment).

The four levels of joint attention found by Carpenter et al. (1998) can be translated into four levels of supervision in a model, and the dependent variable will be number of words learned at different points over the course of training. Because the Carpenter et al. (1998) study was observational, it may also be possible to introduce an experimental manipulation of joint attention in a laboratory situation, perhaps over the course of a sequence of days rather than months, to look at the learning of specific words. If it could be established that supervised and unsupervised learning, together, produce the best word acquisition in this scenario, it would be solid evidence that language learning is semi-supervised, and validate the use of semi-supervised paradigms in connectionist models.

Appendix A: Technical Appendix[5]

Early cognitive modelers proposed that complex cognitive processes arise out of the interactions of many simple processing units (Minsky & Papert, 1969). At that time, Minsky and Papert argued that the usefulness of neural networks was limited by their inability to learn categories based on the nonlinear combination of multiple input features, such as the exclusive-OR (XOR) problem. In the XOR problem, a neural network must learn that when either of two input units is active, the output unit will be active, but when both input units are inactive (Boolean NOT) or both input units are active (Boolean AND), the output will be inactive. This problem requires the nonlinear combination of inputs because simply summing the inputs results in an incorrect solution. However, the parallel distributed processing (PDP) approach, introduced in the 1980s (Hinton, 1981; McClelland, Rumelhart, & Hinton, 1986; Rumelhart, Hinton, & McClelland, 1986), demonstrated that problems requiring the nonlinear combination of inputs can be solved with the addition of hidden layers to the neural networks (McClelland, Rumelhart, & Hinton, 1986; Rumelhart, Hinton, & McClelland, 1986). Hidden layers are those that mediate between inputs and outputs. They are “hidden” because in the observation of an organism the stimulus and the organism's response can be seen, but the neural processes driving the response cannot; such processes are “hidden” in the “black box” that strict behaviorism disregards and cognitive psychology seeks to illuminate. The innovation of hidden-layer models led to the explosion of cognitive research using neural networks that continues to this day. Over time, the PDP approach and others like it, have collectively come to be called connectionist models. These approaches are distinguished by the emergence of complex behaviors from the interactions of simple units and storage

[5] Because the current work is so dependent on the principles of connectionist modeling, an explanatory appendix was deemed more appropriate than integrating such important information with the substance of the literature reviews and novel simulations contained herein.

of information in the weights associated with the connections between units.

One key argument for connectionism is that it is meant to emulate the brain. The processing units of a connectionist network are like neurons in the brain. The weighted connections between units are analogous to synapses in that they may be strengthened or weakened by experience. Because of this correspondence, connectionist modeling allows us to simulate not only cognitive processes, but also the neural substrates of those processes. Further, what works in a connectionist network may also be shown to work in the brain (Hinton & Shallice, 1991). It is also useful in discussions of innateness; if a connectionist model can learn a task (say, a semantic task) without preprogrammed structures, rules, or guiding principles, then this supports the possibility that in humans, such things are not innate and can be learned (Rogers & McClelland, 2004). (Such a discussion is beyond the scope of the present work.) This mechanistic viewpoint of connectionism bridges the disciplines of psychology and neuroscience.

This provides a crucial bridge as the disciplines of cognitive psychology and neuroscience grow closer together. Even though the coding of information in a simulated model cannot necessarily be taken as a theory of how information is coded in the brain (Cree et al., 1999), it still provides a good link between abstract cognitive theories and neuroscience. There is no obvious connection between neural functioning and the activation and spread of information; but there is between neural functioning and connectionist models, and between connectionist models and the spread of information.

Connectionist theories propose that all information comes in patterns (Cree et al., 1999; Delgado et al., 2000; Hinton & Shallice, 1991; Masson, 1995; Newell, 1986; Plaut & Booth, 2000; Rogers & McClelland, 2004) (Newell (1986) calls these patterns symbols). These patterns of representation ("symbols") are controlled by the weights ("synaptic strengths") that connect the processing units ("neurons"), which are slowly adjusted over learning (Rogers & McClelland, 2004). Current theories establish distributed representations of information as a basic premise (Hadley & Cardei, 1999; Hinton & Shallice, 1991; Masson, 1995; O'Reilly, 1998; Rogers & McClelland, 2004).

Although some models may use localist coding schemes, using a single unit in the network to represent a single unit of information, this is done not to suggest that the brain uses localist coding, but rather for simplicity.

This use of distributed representations - the sharing of information among connected nodes - is one of the hallmarks of the connectionist approach (Hinton & Anderson, 1981; Hinton, McClelland, & Rumelhart, 1986). The pattern of activation, or state of the entire network, corresponds to the active item of information, which may be a concept, percept, or memory, for example (Rumelhart & Todd, 1993). The state space (Churchland, 1986) of such a network is represented as vector (Landauer & Dumais, 1997) or attractor (Cree, McRae, & McNorgan, 1999) space, and the relatedness of two items of information is indicated by a similar pattern of activation (in the neural network's hidden layer), or the angle between the vectors, or the distance between the attractors.

The discussion here will make reference to both learning *algorithms* and learning *paradigms*. Learning algorithms are the mathematical rules that govern the adjustments made to the connection weights during learning, in both computers and in organisms. Contemporary learning algorithms are more realistic than their predecessors, and some, like the Leabra algorithm (O’Reilly, 2001; O’Reilly & Munakata, 2000) have a high degree of biological plausibility.

In contrast, a learning paradigm refers to the way a learning problem is posed; it describes the way data are structured for presentation to a network (Jordan & Rumelhart, 1992; Rumelhart, 1986). The common learning paradigms used in computational modeling are supervised and unsupervised learning, as even a casual investigation of the literature will show. Both of these paradigms can be implemented with a variety of learning algorithms, and in some cases it is difficult to determine which paradigm is being implemented.

Unsupervised Learning

Unsupervised learning is based on the extraction of correlated information from an environment. The classic algorithm for unsupervised learning is the Hebbian learning rule, named for neuroscientist D. O. Hebb, who first formulated the idea that the synapse between two concurrently active neurons will become stronger (Hebb, 1949). This is the instantiation, in neural terms, of association learning as discussed by behaviorist researchers such as Pavlov or Skinner. Here, the network or organism learns the structure of an environment through the co-occurrences of that environment's features.

The phrases “unsupervised learning,” “self-organizing learning,” and “Hebbian learning” are often used interchangeably (see, i.e., O’Reilly, 1998), and this kind of learning can also be referred to as model learning (O’Reilly, 1998; 2001). Unsupervised learning is used in situations in which the data are not *labeled*, where a label is related information that is not concurrently available from the environment. Essentially, unsupervised leaning is driven by the environment. Examples are clustered based on the similarity of the patterns generated by the input to the network; natural differences in input patterns can be preserved by this clustering, but the clusters so formed may not be accurate, especially in the case where data are not linearly separable into clusters. This clustering happens through the strengthening of inter-nodal connections among simultaneously active nodes. Delgado and colleagues (2000) describe this process as “the emergence of global order from local interactions” (p. 502); unsupervised learning also has the advantage of being computed solely through information locally available to the units, such as would happen in the brain (Delgado et al., 2000; O’Reilly, 2001; O’Reilly & Munakata, 2000). This sort of learning is crucial for constructing task-independent representations of the environment in the neural network, and does much to foster the ability of a network to generalize (O’Reilly, 1998, 2001).

Classic Hebbian learning, which is known to exist in the brain and is often the primary example writers have in mind when using the phrase “unsupervised learning,” describes a synapse between two neurons that becomes stronger when the firing of those neurons is correlated (Hebb, 1949). In a computational model, Hebbian learning is the name given to learning that occurs on the basis of such correlations (O’Reilly, 1998). In terms of the environment, these correlations can be interpreted as redundancy in the data (Collins & Singer, 1999), or as correlations between inputs to multiple sensory systems (Roy, 2000). The strength of a connection between two units, the connection's *weight*, becomes stronger when the units are active at the same time, and decreases when the units are not active at the same time. The conditional probability of the units being simultaneously active is reflected in the weight strengths that are learned (O’Reilly, 1998). The biological basis of Hebbian learning is well-established (Delgado et al., 2000).

In mathematical terms, a Hebbian learning algorithm takes the form of.

$$wt = \eta x(t)y(t) \qquad \text{(Eq. 1)}$$

(Montague & Sejnowski, 1994). The change in the connection weight between two neurons at time t is a function of a fixed learning rate η times the presynaptic activity $x(t)$ times the postsynaptic activity $y(t)$ (Montague & Sejnowski, 1994). The learning rate η is the amount of change made to the weight at one iteration of the algorithm. This algorithm strengthens the synapse between two neurons when their firing is correlated in a precise way, such that presynaptic firing precedes postsynaptic firing by some small amount of time (Gerstner & Kistler, 2002; Scarpetta et al., 2002). Montague and Sejnowski (1994) suggested that the Hebbian rule is symmetric, that is, learning is not dependent on which neuron fires first. However, more recent research has established that if the presynaptic neuron fires first, the synapse is potentiated, or made stronger, but if the postsynaptic neuron fires first, the synapse is depressed, or weakened (Gerstner & Kistler, 2002; Scarpetta et al., 2002).

Unlabeled data are sufficient in and of themselves for dividing a data set into classes or clusters

(O'Reilly, 2001; Roy, 2000). However, because all the information used to make weight changes in Hebbian learning is at the level of the synapse, there is no way to drive the solution of tasks requiring more remote information (O'Reilly, 2001). Supervised learning is necessary for assigning labels to the newly classified sets of data (Nigam et al., 2000).

Supervised Learning

Supervised learning (also called "error-driven", or "task" learning) is a learning paradigm that is used on labeled data (Gharamani & Jordan, 1994; Hanson, 1995; Jordan & Rumelhart, 1992; O'Reilly, 1996, 1998, 2001; Rumelhart, Hinton, & McClelland, 1986; Rumelhart, Hinton, & Williams, 1986; Stone, 1986). In most instances of supervised learning, the network is fed an input and computes an output representation. For example, an animal sees its preferred kind of food (input) and moves toward it in order to eat it (output). The difference between the output that the network produces and the desired target, or label, is used to derive an error signal. This is fed back into the network for the purpose of altering the connection weights such that on successive iterations, the network will generate an output closer to the desired target (Hanson, 1995; Jordan & Rumelhart, 1992; O'Reilly, 1996, 1998; O'Reilly & Munakata, 2000; Rumelhart, Hinton, & McClelland, 1986; Rumelhart, Hinton, & Williams, 1986; Rumelhart & Todd, 1993; Sarrukai, 1997; Stone, 1986). For example, if the animal does not reach the food after taking one step, it will take another step in order to get closer to the food. It is also possible to use supervision only to cluster the input data (group based on similarity), not to apply output labels to it (Sarrukai, 1997). The exception to this formulation is the Hopfield network, in which the desired output is *clamped* (desired activity levels in the output units are set by hand and not permitted to vary) and the network settles into a state that will generate that outcome (Hopfield, 1982). In a fully trained network, it is possible to use the input to predict what the output will be (Gat, 2001; Gharamani & Jordan, 1994).

Supervision in the form of feedback is necessary for development of some kinds of behavior, such as those that emerge in operant conditioning (Sutton, 1984), and song learning in some birds. Bottjer and Arnold (1997) have found that experience is necessary for enervation of major song-controlling areas in the male zebra finch brain, and without auditory feedback (i.e., if the birds cannot hear themselves sing) this circuit does not develop. More importantly, supervised learning is necessary for learning how to categorize inputs that cannot be categorized using only concurrent environmental information. For example, an ape may need to use its previous experience to determine that a bad-smelling fruit is still good to eat. Also, recall that in the XOR problem, a correct solution could not be reached simply by summing the activations of the two input units. In both these cases, it is necessary to use information that cannot be directly obtained from the environment.

Supervised learning is often formulated in such a way as to make necessary a "*teacher signal*," that is, a mechanism by which information about a network's output error is passed back to earlier stages of the network (Rumelhart, Hinton, & McClelland, 1986; Rumelhart, Hinton, & Williams, 1986; Stone, 1986). This signal can be thought of as explicit correction from an external source, but there are other ways of interpreting the teaching signal that do not require that it be explicit, an approach that is probably more plausible from a biological standpoint (O'Reilly, 1998). For example, Jordan and Rumelhart (1992) interpret the target in their model of distal, or temporally distant, learning as the outcome of some action taken as a result of a network state (i.e., the brain sends particular signals to hands and arms, you throw a ball, and either make the basket or fail to; pp. 310-311). This are very similar to uses of reinforcement learning algorithms, (Barto, 1989; Jordan & Rumelhart, 1992; Sutton, 1984). In all these cases, the error signal is the difference between the desired and the obtained outcome (O'Reilly, 1998).

One algorithm for supervised learning that makes use of a teacher signal is the backpropagation algorithm, or delta rule:

$$\Delta_p w_{ij} = \eta (t_{pj} - o_{pj}) i_{pi} \qquad \text{(Eq. 2)}$$

(Rumelhart, Hinton, & Williams, 1986). In this rule,

$$\Delta_p w_{ij} \qquad \text{(Eq. 3)}$$

refers to the change in weight *w* between units *i* and *j* following presentation of pattern *p*. The learning rate, η gives the size of the weight change at each iteration. The term

$$(t_{pj} - o_{pj}) \qquad \text{(Eq. 4)}$$

gives the difference between the desired and actual output of the *i*th unit, and

$$i_{pi} \qquad \text{(Eq. 5)}$$

gives the value of the *i*th unit (Rumelhart, Hinton, & Williams, 1986). While both the Hopfield and the delta rule learning mechanisms were motivated by a desire to understand how a collection of relatively uniform interconnected units could produce a complex behavior, including pattern storage (memory) and cognition (Hopfield, 1982; Rumelhart, Hinton, & McClelland, 1986), neither is particularly plausible from a biological standpoint.

The introduction of parallel distributed processing models into psychology allowed cognitive models to have a degree of biological realism, and biological models a degree of cognitive realism (Delgado et al., 2000; Rumelhart & McClelland, 1986). But the delta rule's lack of more specific biological constraints have become a liability in the face of greater knowledge about the workings of the brain. Therefore, more plausible algorithms for supervised learning are necessary. For example, as originally formulated, the delta rule requires that error information be transmitted backward along the same pathways through which information was propagated in a forward direction; that connection weights be both positive and negative; and that the same synapses be both excitatory and inhibitory, none of which actually occur in the brain. Because activation and error are propagated via different mechanisms, use of the delta rule gives rise to nonlocal learning (Crick, 1989; O'Reilly, 1996, 2001;

Zipser & Andersen, 1988), which also violates the contemporary principle of biological plausibility. All variables used in learning in the brain must be available to the individual neurons at their locations.

To sum up, supervised learning is the only way to associate an exemplar with a specific category or external data label, and it is essential for situations of predictive learning. However, it does not handle generalization well, in that exceptional exemplars often perturb the network's performance on the entire data set. Unsupervised learning, on the other hand, is much more effective at generalization and at handling exceptions, but much poorer at task learning, and it cannot associate exemplars with categories or external labels, but merely cluster them on the basis of input similarity. The different strengths of supervised and unsupervised learning suggest that a combination of the two paradigms may be involved in the most effective computational system for general learning.

Mixed-Model Learning

Supervised and unsupervised learning have been combined to a greater or lesser extent in a number of frameworks. In the distal learning paradigm, Jordan and Rumelhart (1992) envision the teacher as the desired outcome of an action. O'Reilly (1998) posits something similar, suggesting that the teacher "takes the form of actual environmental outcomes that can be compared with internal expectations" (p. 457). Jordan and Rumelhart (1992) provide the example of learning to throw a basketball. Feedback in this case is how close the throw comes to the basket, and subsequent throws can be adjusted on the basis of the degree of success achieved in a previous throw. In distal learning, the teacher is necessarily present, but in a covert form (Jordan & Rumelhart, 1992).

Sakaguchi (1990) rejected the notion that neurons represent only environmental stimuli, arguing instead that other forms of representation are needed to adequately carry out information processing. Sakaguchi explored this concept by introducing a teacher signal into a model that organized a region of neurons topographically to represent the environment, using a Hebbian learning mechanism. The mapping is stable, and, like the somatosensory cortex in mammals (Gardner & Kandel, 2000), large representational areas are devoted to frequently stimulated input areas (Sakaguchi, 1990). With self-organization, a columnar microstructure is formed, very similar to that of somatosensory cortex. In Sakaguchi's (1990) model with the added teacher signal, connections within each column arise through Hebbian learning, but the global organization of the field is shaped by the teacher signal. Use of the teacher is justified in terms of the integration of multimodal signals and output to a response system (i.e., the motor system), where the environment provides feedback about whether a given action had the desired outcome (not unlike Jordan and Rumelhart's (1992) distal learner). For example, consider seeing a moving target in the visual periphery, and turning one's head to look. If the location of the object is mapped accurately, one is rewarded with a clearer view. If the location mapping is inaccurate, the target may disappear from view.

There is no reason to think that Sakaguchi's mechanism is implausible, but without more specific neural mechanisms being assigned to this process, the possibility (as opposed to the plausibility) of this claim is impossible to evaluate. As an additional critique, the connectionist simulations performed in this research assumed the preformation of columnar microstructure. This assumption is safe if columns are formed without respect to environmental input. However, if the intent is to show that column formation is driven by the environment (where, presumably, the teacher signal comes from), then the network needs to be instructed properly, not gifted with a pre-established organization. Sakaguchi may be agnostic about this point; here, the utility of columns is apparently more important than their origin.

Contrastive Hebbian learning (CHL) is a supervised implementation of Hebbian learning (Xie & Seung, 2003). CHL, uses two activation phases: a phase with output units clamped to the target value, and a phase where the output units settle freely. The difference between these phases is the error signal (Xie & Seung, 2003). CHL uses Hebbian weight updates and is used in feedback networks.

These feedback connections are symmetric with the feedforward connections except for the inclusion of an additional multiplicative term to make the feedback connections weaker than the feedforward ones (Xie & Seung, 2003), a circumstance which occurs in the brain (O'Reilly, 2001). The feedback connections allow clamping of the output units to affect earlier parts of the network, namely, the hidden layer (Xie & Seung, 2003). CHL in this formulation is equivalent to backpropagation, but is more biologically plausible (Xie & Seung, 2003).

The most sophisticated of the mixed model algorithms, the Leabra learning algorithm (O'Reilly, (2001, O'Reilly & Munakata, 2000) implements both error-driven learning and Hebbian learning over the same data. Leabra is an extension of O'Reilly's (1996) earlier algorithm, the generalized recirculation (GeneRec) algorithm. This is an algorithm for error-driven learning developed specifically for use in interactive networks, those with both feedforward and feedback connections (O'Reilly, 1996, 1998; Pineda, 1995). Networks that use GeneRec have a greater biological plausibility than earlier formulations of error-backpropagation not only due to their interactivity, but also because the error signals are computed locally using activation variables, a type of information that is available to neurons in the brain (O'Reilly, 1996, 1998). Using local information also allows weights to update independently of each other (Istook & Martinez, 2002), which means that information processing is controlled by the neurons, not by an external force (Fletcher, 2000).

However, in other ways the GeneRec algorithm is less plausible (O'Reilly, 1996). As with CHL, which is also used for supervised learning in interactive networks, a network trained with the GeneRec algorithm will generalize poorly. Leabra adds both Hebbian learning and k-winners-take-all (kWTA) inhibitory competition (O'Reilly, 2001; O'Reilly & Munakata, 2000) to GeneRec. These two features add generalization ability to interactive networks, Hebbian learning by detecting such correlations as mentioned above, and inhibitory competition by forcing the development of sparse representations (O'Reilly, 1998, 2001; O'Reilly & Munakata, 2000). Inhibitory competition also reduces settling time, which is correlated with generalization ability. Incidentally, generalization is also fostered by large hidden layers in Leabra, though it is impaired by them in other algorithms (O'Reilly, 2001).

The Leabra algorithm

$$\Delta w_{ij} = \varepsilon[k_{hebb}(\Delta_{hebb}) + (1 - k_{hebb})(\Delta_{sberr})] \quad \text{(Eq. 6)}$$

uses a term for Hebbian learning,

$$k_{hebb}(\Delta_{hebb}) \quad \text{(Eq. 7)}$$

and assigns the rest of the weight change on an iteration to supervised learning,

$$(1 - k_{hebb})(\Delta_{sberr}) \quad \text{(Eq. 8)}$$

(O'Reilly, 2001; O'Reilly & Munkata, 2000). The sum of these terms, times the learning rate ε, gives the value of the weight change. In a network trained using Leabra, the weight change for every input is driven by similar features (clustering) and by an error signal. Inhibitory competition, combined with Hebbian learning, makes such a network effective not only at learning of a training set of inputs, but also at generalizing to new examples (O'Reilly, 2001; O'Reilly & Munkata, 2000). Leabra's power and flexibility in these respects make it an ideal mixed-model algorithm. (However, it is still unable to learn in certain environments, as discussed in Chapter 2).

Appendix B: Glossary

Activation value In a neural network, the numerical representation of the probability with which a unit will fire on a receiving unit.

Attractor space Representing the state of a neural network by distance from the point of lowest enegry or error.

Backpropagation Sending information in a retrograde or "backwards" direction along the connection between two neurons. May be more plausibly implemented as a parallel backwards connection.

Clamping Neural network procedure in which desired output activations are set by hand and not permitted to vary, to determine the network state(s) that make such an output possible

Clustering Grouping of inputs based solely on unlabeled information

Connectionism An approach to computational modeling in which behaviors emerge from the parallel interaction of many simple neuron-like units. See **parallel distributed processing**.

Delta rule A backpropagation algorithm that calculates the difference between an expected and obtained output in order to strengthen or weaken connection weights appropriately.

Distal learning Learning of a response remote in time from an input.

Distributed coding Using multiple units in an artificial neural network to represent an environmental occurrence, thought, or behavior. See **localist coding.**

Feedback Information, usually about error, is sent from the output of a network in a retrograde or backwards direction to affect the processing of units earlier in the network.

Feedforward Information proceeds from presynaptic (sending) units, to postsynaptic (receiving) units.

Global task Task the solution of which requires the combined outputs of a neural network.

Hebbian algorithm Mathematical equation describing the increase of connection weights between two simultaneously active neurons

Hidden layers Groups of neurons in a neural network that are interposed between the network's input and output

Hierarchy A form of semantic network organization in which words are superordinate or subordinate to related words, and associations from one word to another must proceed along these superordinate/subordinate connections

k-Winners-take-all inhibition A variable number of strongly active units inhibit the activation of more weakly active units

Label An item of information not currently available from the environment. For example, one may see an unfamiliar bird and be told its name after it flies away.

Leabra A learning algorithm that implements Hebbian and backpropagation learning over the same connections simultaneously

Learning algorithms Mathematical rules that govern the adjustments made to connection weights between two units during learning

Learning paradigms The way a learning problem is posed; how data are structured for presentation to a neural network

Learning rate The amount by which the connection between two units changes on one iteration of the learning algorithm

Localist coding Using one unit in an artificial neural network to represent an environmental occurrence, thought, or behavior. See **distributed coding.**

Local information Information available to an individual neuron; the neuron's response can be c omputed without respect to the information available to other neurons

Neural network Group of neurons, either biological or artificial, that interact with each other to form a larger processing unit

Node In an artificial neural network, a neuron-like unit that performs computations on its input and sends output to other, similar nodes

Nonlinear combination of multiple input features The inputs to a neuron do not sum linearly. For example, the postsynaptic neuron should fire when one but not both of its inputs fires on it.

Nonlocal learning The connection weight between two units changes in response to the output of the network as a whole rather than the signaling of the individual receiving unit

Parallel distributed processing (PDP) An approach to computational modeling in which behaviors emerge from the parallel interaction of many simple neuron-like units. See **connectionism**.

Patterns of representation The set of active processing units in a neural network that corresponds to an environmental occurrence, thought, or behavior

Self-organizing learning See **Unsupervised learning**.

Settling time In some neural networks, the speed at which each neuron takes on an activation value in response to some input.

Sparse representations Representations in a network are distributed, but use few neurons for each representation to maximize the number of representations that can be implemented by the network.

State space All possible patterns of activity that can be taken on by a neural network

Target Desired outcome of a neural network, the output the network learns to produce in response to a given input

Teacher signal Mechanism for feeding error information back into a neural network

Unsupervised learning Learning from information concurrently available in the environment, for example, the sight of a potential food item and an appetizing smell.

Vector space The representation in linear algebraic terms of the state of a network

Weight In an artificial neural network, the strength of the connection between two processing units; the output of the sending unit is muliplied by this value to solve for the input of the receiving unit

XOR problem A classic nonlinear problem, in which one but not both conditions must be true in order for a consequence to be true. An extensive discussion of the XOR problem can be found in D. E. Rumelhart, J. L. McClelland, and the PDP Research Group (Eds.), (1986). *Parallel distributed processing: Explorations in the microstructure of cognition vol. 1* (45-76). Cambridge, MA: MIT Press.

Appendix C: Stimuli for Simulation 5

Training items for Simulation 5

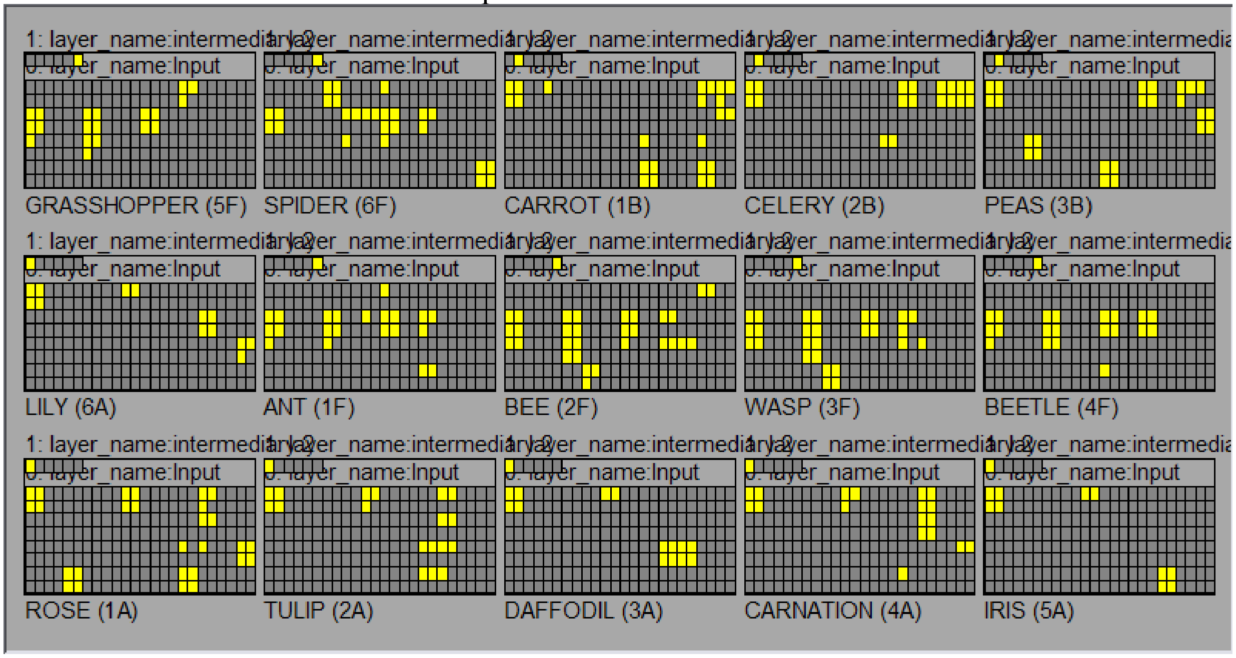
rose
tulip
daffodil
carnation
iris
lily
ant
bee
wasp
beetle
grasshopper
spider[6]
carrot
celery
peas
lettuce
cabbage
corn
apple
orange
banana
fig
grape
peach
dog
cat
horse
pig
lizard
turtle
eagle
robin
sparrow
duck
owl
penguin

Generalization items for Simulation 5

violet
daisy
cockroach
fly
broccoli
potato
pear
apricot
lion
elk
bluejay
ostrich

[6] Taxonomically, "spider" is not a member of the insect category; however, its similarity to the insect members makes it suitable for inclusion in this simulation.

Samples of Stimuli from Simulation 5



References

Anagnostopoulos, G. C., Bharadwaj, M., Georgiopoulos, M., Verzi, S. J., & Heileman, G. L. (2003). Exemplar-based pattern recognition via semi-supervised learning. *Proceedings of the IEEEINNS-ENNS International Joint Conference on Neural Networks (ISCNN '03).* Portland, Oregon. Retrieved 28 March 2003 from pegasus.cc.ucf.edu/~mbharadw/thesis/gca-ssl-ijcnn-2003.pdf.

Anagnostopoulos, G. C., & Georgiopoulos, M. (2001). Ellipsoid ART and ARTMAP for incremental unsupervised and supervised learning. *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (ICJNN '01)*, vol. 2 (1221-1226). Washington, DC: IEEE, INNS, ENNS. Retrieved 1 July 2004 from http://my.fit.edu/~georgio/research/publications/gca-eam-icjnn-2001.pdf.

Anagnostopoulos, G. C., Georgiopoulos, M., Verzi, S. J., & Heileman, G. L. (2002). Reducing generalization error and category proliferation in ellipsoid ARTMAP via tunable misclassification error tolerance: Boosted ellipsoid ARTMAP. To appear in *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (ICJNN '02)*, Vol. 3 (2650-2655), Honolulu, Hawai'i: IEEE, INNS, ENNS. Retrieved 1 July 2004 from http://my.fit.edu/~georgio/research/publications/gca-beam-icjnn-2002.pdf.

Balota, D. A., & Lorch, R. F. (1986). Depth of automatic spreading activation: Mediated priming effects in pronunciation but not in lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition 12*, 336-345.

Barto, A. (1989). From chemotaxis to cooperativity: Abstract exercises in neuronal learning strategies. In R. Durbin, C. Miall & G. Mitchison, (Eds.), *The Computing Neuron* (pp. 73-98). Wokingham, England: Addison-Wesley.

Bennet, K., & Demirez, A. (1998). Semi-supervised support vector machines. In M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in Neural Information Processing Systems vol. 12* (368-374). Cambridge, MA: MIT Press.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Bottjer, S. W., & Arnold, A. P. (1997). Developmental plasticity in neural circuits for a learned behavior. *Annual Review of Neuroscience 20*, 459-481.

Carnegie Mellon University. (1995). *PDP++*.

Carpenter, G. A., Grossman, S., Markuzon, N., Reynolds, J. H., & Rosen, D. B. (1992). Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks 3*, 698-713.

Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development 63*, v-143.

Chapelle, O., Schölkopf, B., & Zien, A. (2006). Introduction to semi-supervised learning. In O. Chapelle, B. Schölkopf, & A. Zien (Eds.), *Semi-supervised learning* (pp. 1-12). Cambridge, MA: MIT Press.

Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive Science 26*, 609-651.

Chen, Y., Wang, G., & Dong, S. (2003). Learning with progressive transductive support vector machine. *Pattern Recognition Letters 24*, 1845-1855.

Churchland, P. M. (1989). *A neurocomputational perspective: The nature of mind and the structure of science.* Cambridge, MA: MIT Press.

Collins, A.M., & Loftus, E.F. (1975). A spreading-activation theory of semantic processing.

Psychological Review 82, 407-428.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning & Verbal Behavior 8*, 240-247.

Collins, M., & Singer, Y. (1999). Unsupervised models for named entitiy classification. In *Proceedings of EMNLP/VLC-99*.

Cozman, F., & Cohen, I. (2006). Risks of semi-supervised learning. In Chapelle, O., Scholkopf, B., & Zien, A. (Eds.) *Semi-supervised learning* (pp. 57-72). Cambridge, MA: MIT Press.

Cree, G. S., McRae, K., & McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science 23*, 371-414.

Crick, F. H. C. (1989). The recent excitement about neural networks. *Nature 337*, 129-132.

Dale, P.S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers 28*, 125-127.

Delgado, J. F. R., Dalenoort, G. J., & Garcia, A. P. (2000). Biological and theoretical relevance of some connectionist assumptions. The development of conceptual networks. *Psicothema 12*, 500-505.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*, 1-38.

Echols, C.H., & Marti, C.N. (2004). The identification of words and their meanings: From perceptual biases to language-specific cues. In Hall, D.G., & Waxman, S.R. (Eds.), *Weaving a lexicon* (41-78). Cambridge, MA: MIT Press.

Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E., Hartung, J. P., Pethick, S., & Reilly, J. S. (1993). *The MacArthur communicative development inventories: User's guide and technical manual*. San Diego: Singular.

Fletcher, P. (2000). The foundations of connectionist computation. *Connection Science 12*, 163-196.

Gardner, E. P., & Kandel, E. R. (2000). Touch. In E. R. Kandel, J. H. Schwartz, & T. M. Jessell (Eds.), *Principles of neural science* (4th ed.) (pp. 451-470). New York: McGraw-Hill.

Garzon, F. C. (2003). Connectionist semantics and the collateral information challenge. *Mind & Language 18*, 77-94.

Gat, Y. (2001). A learning generalization bound with an application to sparse-representation classifiers. *Machine Learning 42*, 233-239.

Gerstner, W., & Kistler, W. M. (2002). Mathematical formulations of Hebbian learning. *Biological Cybernetics, 87*, 404-415.

Gharamani, Z., & Jordan, M. I. (1994). Supervised learning from incomplete data via an EM approach. In J.D. Cowan, G. Tesauro, & J. Alspector (Eds.), *Advances in neural information processing systems vol. 6* (120-127). Morgan Kaufman.

Gogate, L.J., & Bahrick, L.E. (2001). Intersensory redundancy and 7-month-old infants' memory for arbitrary sellable-object relations. *Infancy 2*, 219-231.

Goodman, J.C., Dale, P.S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language 35*, 515-531.

Guo, Z, Zhang, Z.(M.), Xing, E.P., & Faloutsos, C. (2008). Semi-supervised learning based on semiparametric regularization. Society for Industrial and Applied Mathematics Data Mining Conference, April 24-26, Atlanta, Georgia, USA. Accessed June 15, 2010 from http://www.siam.org/proceedings/datamining/2008/dm_08_12_Guo.pdf.

Gurney, K. (1997). Adaptive resonance theory: ART. In *An introduction to neural networks.* (147-166). London: UCL Press.

Hadley, R. F., & Cardei, V. C. (1999). Language acquisition from sparse input without error feedback. *Neural Networks 12*, 217-235.

Hanson, S. J. (1995). Some comments and variations on back-propagation. In Y. Chauvin and D. Rumelhart (Eds.), *Backpropagation: Theories, architectures, and applications. Developments in connectionist theory* (292-323). Hillsdale, NJ: Erlbaum.

Hebb, D. O. (1949). *The organization of behavior.* New York: Wiley.

Hinton, G. E., (1981). Implementing semantic networks in parallel hardware. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel models of associative memory.* (pp. 161-188). Hillsdale, NJ: Erlbaum.

Hinton, G.E., & Anderson, J.A. (1981). *Parallel models of associative memory.* Hillsdale, NJ: Erlbaum.

Hinton, G.E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition vol. 1* (77-109). Cambridge, MA: MIT Press.

Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review 98*, 74-95.

Hirsh-Pasek, K., Golinkoff, R.M., Hennon, E.A., & Maguire, M.J. (2004). Hybrid theories at the frontier of developmental psychology: The emergentist coalition model of word learning as a case in point.   In Hall, D.G., & Waxman, S.R. (Eds.), *Weaving a lexicon* (173-204). Cambridge, MA: MIT Press.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational capabilities. *Proceedings of the National Academy of Sciences 79*, 2554-2558.

Istook, E., & Martinez, T. (2002). Improved backpropagation learning in neural networks with windowed momentum. *International Journal of Neural Systems 12*, 303-318

Jordan, M. I., & Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science 16*, 307-354.

Keil, F.C. (1989). Concepts, kinds, and cognitive development. Cambridge, MA: MIT Press.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English.* Providence: Brown University Press.

Landau, B. (2004). Perceptual units and their mapping with language: How children can (or can't?) use perception to learn words. In Hall, D.G., & Waxman, S.R. (Eds.), *Weaving a lexicon* (111-148). Cambridge, MA: MIT Press.

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

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review 104*, 211-240.

Li, T., Zhu, S., Li, Q., & Ogihara, M. (2003). Gene functional classification by semi-supervised learning from heterogeneous data. ACM SAC Bioinformatics Track, March 9 to 12, 2003, Melbourne, Florida, USA. Retrieved 28 March 2003 from www.cs.rochester.edu/~zsh/pub/bio-132.pdf.

Love, B.C., Medin, D.L., & Gureckis, T.M. (2004). SUSTAIN: A network model of category learning. *Psychological Review 111*, 309-332.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk, vol. 2: The database (3rd ed.)*. Mahwah, NJ: Lawrence Erlbaum.

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language 12*, 271-295.

Mandler, J.M. (1996). Preverbal representation and language. In Bloom, P., Peterson, M.A., Nadel, L., & Garret, M.F. (Eds.), *Language and space: Language, speech, and communication* (365-384). Cambridge, MA: MIT Press.

Masson, M.E.J. (1991). A distributed memory model of context effects in word identification. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 233-263). Hillsdale, NJ: Erlbaum.

McClelland, J. L. (1994). The interaction of nature and nurture in development: A parallel distributed processing perspective. In P. Bertelson, P. Eelen, & G. d'Ydewalle (Eds.), *International perspectives on psychological science, volume 1: Leading themes*. (57-88). Hillsdale, NJ: Erlbaum.

McClelland, J.L., McNaughton, B.L., & O'Reilly, R.C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review 102*, 419-457.

Minsky, M., & Papert, S. (1969). *Perceptrons.* Cambridge, MA: MIT Press.

Montague, P. R., & Sejnowski, T. J. (1994). The predictive brain: Temporal coincidence and temporal order in synaptic learning mechanisms. *Learning & Memory 1*, 1-33.

Murphy, G.L., & Medin, D.L. (1985). The role of theories in conceptual coherence. *Psychological Review 92*, 289-316.

Muslea, I., Minton, S., & Knoblock, C. A. (2002). Active + semi-supervised learning = robust multiview learning. *Proceedings of the 19th International Conference on Machine Learning* (IMCL 2002), 435-442. Retrieved 28 March 2003 from www.isi.edu/info-agents/papers/muslea02-icml-robust.pdf.

Nelson, K. (1988). Acquisition of words by first language learners. In Franklin, M.B., & Barten, S.S. (Eds.), *Child language: A reader* (50-59). New York: Oxford University Press.

Newell, A. (1986). The symbol level and the knowledge level. In Z. W. Pylyshyn & W. Demopoulos (Eds.), *Meaning and cognitive structure: Issues in the computational theory of mind* (pp. 31-

39). Westport, CT: Ablex.

Nigam, K., & Ghani, R. (2000). Analyzing the effectiveness and applicability of co-training. In *Ninth International Conference on Information and Knowledge Management (CIKM-2000),* 86-93. Retrieved 1 July 2004 from http://www.kamalnigam.com/papers/cotrain-CIKM00.pdf.

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. M. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning 39,* 103-134.

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation 8*, 895-938.

O'Reilly, R. C. (1998). Six principles for biologically-based computational models of cortical cognition. *Trends in Cognitive Sciences 2*, 455-462.

O'Reilly, R. C. (2001). Generalization in interactive networks: The benefits of inhibitory competition and Hebbian learning. *Neural Computation, 13*, 1199-1241.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience.* Cambridge, MA: MIT Press.

Pathak-Pal, A., & Pal, S. K. (1987). Learning with mislabeled training samples using stochastic approximation. *IEEE Transactions on Systems, Man, and Cybernetics 17*, 1072-1077.

Pineda, F. J. (1995). Recurrent backpropagation networks. In Y. Chauvin & D. E. Rumelhart (Eds.), *Backpropagation: Theories, architectures, and applications. Developments in connectionist theory* (99-135). Hillsdale, NJ: Erlbaum.

Pinker, S. (1997). Words and rules in the human brain. *Nature 387*, 547-548.

Plaut, D. C., & Booth, J. R. (2000). Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review 107*, 786-823.

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology 10*, 377-500.

Robare, R. J. (2004). Generalization and discrimination in a semantic network trained with semi-supervised learning. *Proceedings of the Sixth International Conference on Cognitive Modeling* (pp. 412-413). Mahwah, NJ: Lawrence Erlbaum.

Robare, R. J. (2005, November). Parameters for semi-supervised learning in neural-network models of cognition. Poster session presented at Dynamical Neuroscience XIII: Computational Cognitive Neuroscience, Washington, DC.

Rogers, T.T., & McClelland, J.L. (2004). *Semantic cognition: A parallel distributed processing approach.* Cambridge, MA: MIT Press.

Rosch, E., & Mervis, C.B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology 7*, 573-605.

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

Roy, D. (2000). Learning from multimodal observations. *IEEE International Conference on Multimedia and Expo (I)*, 579-582. Retrieved 7 May 2003 from dkroy.www.media.mit.edu/people/dkroy.

Roy, D. (2006). Human speechome project press event. Accessed June 15, 2010 from http://www.media.mit.edu/events/movies/video.php?id=zetera-2006-05-15.

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

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group

(Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition vol. 1* (45-76). Cambridge, MA: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition vol. 1* (318-362). Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1986). PDP models and general issues in cognitive science. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group, *Parallel distributed processing, Vol. 1: Foundations*. (pp. 110-146). Cambridge, MA: MIT Press.

Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. In D.E. Meyer & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 3-30). Cambridge, MA: MIT Press.

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science 274*, 1926-1928.

Sakaguchi, Y. (1990). Topographic organization of nerve field with teacher signal. *Neural Networks 3*, 411-421.

Sarrukai, R. R. (1997). Supervised networks that self-organize class outputs. *Neural Computation 9*, 637-648.

Scarpetta, S., Zhaoping, L., & Hertz, J. (2002). Hebbian imprinting and retrieval in oscillatory neural networks. *Neural Computation, 14*, 2371-2396.

Singh, A., Nowak, R.D., & Zhu, X. (2008). Unlabeled data: Now it helps, now it doesn't. NIPS 08. http://pages.cs.wisc.edu/~jerryzhu/pub/NIPS08_SSL_v6.pdf.

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

Smith, L.B., Jones, S.S., Landau, B., Gershkoff-Stowe, L., & Samuelson L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science 13*, 13-19.

Snedeker, J., & Gleitman, L.R. (2004). Why it is hard to label our concepts. In Hall, D.G., & Waxman, S.R. (Eds.), *Weaving a lexicon* (257-293). Cambridge, MA: MIT Press.

Spelke, E.S. (1994). Initial knowledge: Six suggestions. *Cognition 50*, 431-445.

Spelke, E.S., Phillips, A., & Woodward, A.L. (1995).. Infants' knowledge of object motion and human action. In Sperber, D., Premack, D., & Premack, A.J. (Eds.), *Causal cognition: A multidisciplinary debate* (44-78). New York: Clarendon Press/Oxford University Press.

Stone, G. O. (1986). An analysis of the delta rule and the learning of statistical associations. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition vol. 1* (444-459). Cambridge, MA: MIT Press.

Sutton, R.S. (1984). Temporal credit assignment in reinforcement learning. Ph.D. dissertation, Department of Computer Science, University of Massachusetts, Amherst, MA. Published as *COINS Technical Report 84-2*.

Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research 2*, 45-66.

Vapnik, V. N. (1998). *Statistical learning theory.* New York: Wiley.

Waxman, S.R. (2004). Everything had a name, and each name gave birth to a new thought: Links between early word learning and conceptual organization. In Hall, D.G., & Waxman, S.R. (Eds.), *Weaving a lexicon* (295-335). Cambridge, MA: MIT Press.

Weimer-Hastings, K. (1998). Abstract noun classification using a neural network to match word context and word meaning. *Behavior Research Methods, Instruments, & Computers 30*, 264-271.

Weisstein, Eric W. (1999). Log likelihood procedure. From *MathWorld--A Wolfram Web Resource.* http://mathworld.wolfram.com/LogLikelihoodProcedure.html. Retreived December 2006.

Woodward, A.L (2004). Infants' use of action knowledge to get a grasp on words. In Hall, D.G., & Waxman, S.R. (Eds.), *Weaving a lexicon* (149-171). Cambridge, MA: MIT Press.

Xie, X., & Seung, H. S. (2003). Equivalence of backpropagation and contrastive Hebbian learning in a layered network. *Neural Computation, 15*, 441-454.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 189-196.

Zhu, X., Rogers, T., Qian, R., & Kalish, C. (2007). Humans perform semi-supervised classification too. AAAI 07. http://pages.cs.wisc.edu/~jerryzhu/pub/humanSSL.pdf

Zipser, D., & Andersen, R. A. (1988). A backpropagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature 331*, 679-684.