Overgeneralization and recovery:

Learning the negative prefixes of English verbs

Ping Li

July 1992
CRL Technical Report 9203

Center for Research in Language

University of California, San Diego
La Jolla, CA 92093-0526

ABSTRACT

Overgeneralization and recovery have received increasing attention in recent years in the study of language acquisition. This study attempts to use neural network modeling to tap into the processes underlying overgeneralization and recovery by examining the acquisition of the English negative prefixes *un-* and *dis-*. Unlike previous network models designed to map forms to forms, our network learns to map meanings of English verbs onto the negative prefixation forms. Our simulations reveal a number of interesting results. First, the mapping problem in question could be solved by the network in which the learning mechanism has very little predisposed (or innate) structure; moreover, during the course of learning the network develops structured internal semantic representations. Second, overgeneralization is a function of how well a verb fits the inherent meaning categories and its prefixation pattern; in particular, if the novel input is consistent with the prefixation patterns of the learned structure, no overgeneralization occurs. Overgeneralization occurs when the new input is inconsistent with members of the meaning category with respect to prefixation. Third, recovery is hard when similar inputs are mapped to dissimilar outputs; synonyms that undergo different prefixation patterns experience particular difficulties in this regard. Finally, the introduction of new verbs to the network may disturb already learned patterns of old verbs (resulting in U-shaped error function), suggesting a highly dynamic and interactive system in neural network learning. These results indicate that neural networks, because of the properties of distributed representation, are ideal for dealing with form-meaning mappings in language acquisition.

INTRODUCTION

One crucial aspect of human language learning is the learner's ability to generalize existing patterns to novel instances. This ability often leads to "overgeneralization" in learning, a phenomenon characterized by the learner's application of a general pattern to inappropriate cases. Overgeneralization has received increasing attention in recent years in the study of language acquisition (Bowerman, 1987; Clark, 1987; Pinker, 1989). One of the important findings associated with overgeneralization is the U-shaped learning pattern: initially children produce few errors, but later on they overgeneralize a pattern across situations and errors proliferate, and only gradually do they recover from all overgeneralizations and errors cease. What is unclear, however, is the specific kind of mechanisms involved in the over-generalization and recovery process. For example, what are the conditions under which overgeneralization takes place? What drives the learner to eventually give up errors and recover from overgeneralization? How does learning and recovery differ as a function of the interaction between linguistic forms and semantic structures? Researchers from different schools have provided us with different answers to these questions, but so far no general consensus has been reached (Bates & MacWhinney, 1987, 1989; Clark, 1987; Pinker, 1989). This study is an attempt to use neural network modeling to examine overgeneralization and recovery, in order to uncover some psycholinguistic mechanisms underlying the learning process, and to address some important theoretical issues currently under debate.

A pioneering work using neural network modeling to study overgeneralization is Rumelhart & McClelland's (1986) simulation of the acquisition of English past-tense forms. It is well-known in child language studies that at one stage of learning, young children tend to apply the regular past-tense suffix '-ed' to irregular verbs, producing errors such as 'comed' and 'goed' (Kuczaj, 1977; Bybee & Slobin, 1982). Rumelhart & McClelland's modeling results indicate that neural networks are able to capture over-generalization errors displayed in child language, and in particular, the characteristic U-shaped development inherent in children's acquisition of the English past tense. More importantly, these learning patterns, traditionally taken as strong evidence for the application of explicit linguistic "rules", are clearly simulated by the network using a single learning mechanism that does not resort to procedural rules. Rumelhart and McClelland's study has given rise to much debate about the representation of linguistic knowledge and its acquisition. In particular, Pinker & Prince (1988) point out that Rumelhart & McClelland's model suffers from its input encoding scheme, from the way that input was

presented to the system, and from its failure to distinguish between regular and irregular verbs. They conclude that these shortcomings are due to the central tenets of neural networks, and that linguistic knowledge ultimately involves the acquisition and use of symbolic rules. So far, the issue remains highly controversial (Elman, 1990; MacWhinney & Leinbach, 1991; Plunkett & Marchman, 1991).

One of the central issues within the current debate is the network's ability to deal with semantics, i.e., the meaning of linguistic forms. Advocates of symbolic approaches argue that connectionist networks are incapable of handling the complex relations between meaning and form (Pinker & Prince, 1988; Pinker, 1991). Although there have been some models using meaning components, the majority of previous networks in language acquisition have been developed to map forms to forms. For example, given a base form of the verb, the network is required to output its inflected past-tense form, where researchers have been concerned with the specific phonological constraints which govern the verb's suffixation pattern. While it is important to look at the horizontal relations between forms and forms, it is more important to study the vertical relations between meanings and forms, since the form-meaning mapping is an essential aspect of language learning. In order to meet the "meaning challenge" for neural network modeling of language acquisition, we therefore need to examine the meaning-form mappings more closely. This study is an attempt to use a 'semantic neural network' (i.e., a neural network with semantic features) to tap into the process of form-meaning mapping in language acquisition.

Another important aspect of the debate about overgeneralization is the learner's ability to recover from overgeneralization errors. Particularly relevant to the recovery process is the so-called "no negative evidence" problem. In fact, the current study is largely inspired by the various hypotheses around these discussions. Many theorists have proposed that children receive only "positive evidence" about what is correct in their language (i.e., exposure to actual sentences), but no "negative evidence" (i.e., feedback about what is not a possible sentence) (see Braine, 1971; Bowerman, 1987, 1988; Pinker, 1989; Bohannon, MacWhinney, and Snow, 1990, for discussion). If there is no negative evidence, how do children recover from over-generalization errors? Investigators have suggested several solutions, most of which are concerned with syntactic overgeneralization (e.g., Baker, 1979; Chomsky, 1981; Mazurkewich & White, 1984), but none of these seems to be entirely satisfactory (see Bowerman, 1987, for discussion). Many of these solutions appeal to innate linguistic knowledge or predispositions of various kinds, according to which recovery is a function of innate mechanisms or principles that cannot be learned. The current study, though not a direct investigation of the "no negative evidence" problem, will address the issue of how recovery may be governed by a single learning mechanism such as back-propagation. Using the conceptual framework of neural networks as the basis for learning, I hope to provide insights into functional constraints on the processes of over-generalization and recovery. Specifically, modeling the effect of semantic structure on form-meaning mapping in the early and later stages of language acquisition may shed light on the psycholinguistic mechanisms in overgeneralization and recovery.

Much of the current debate centers around the English past tense. While the acquisition of the English past tense forms does involve a meaning component, the problem is mainly (as noted above) a matter of form-form mapping. In this study, I have chosen to study a new problem domain by looking at the acquisition of the English negative prefixes, in particular, *un-* and *dis-*. In contrast to the past tense problem, the problem of the negative prefixation in English is essentially a semantic one.[1] In the next section, I will provide some background about this problem and illustrate its theoretical significance.

[1] This is not deny the phonological, morphological, and historical complexities involved in the problem. But these aspects are of less importance to learning as they are in the case of the English past tense. As we shall see below, semantics plays a major role here.

NEGATIVE PREFIXATION AND ITS ACQUISITION IN ENGLISH

There are a number of negative prefixes in English such as '*de-*', '*dis-*', '*in-*', '*mis-*', and '*un-*'. These prefixes mark negativity in a variety of ways that are not of our direct interest here (see Marchand, 1969, for discussion). The two negative prefixes I am concerned with in this study are '*un-*' and '*dis-*'. Both of them are used to designate the reversal of the action specified by the base verb, as in *untie* and *disconnect*. Note that these verbal prefixes should not be confused with '*un-*' or '*dis-*' attached to adjectives, such as *unkind* or *discontent*. Although clearly related, they mark different meanings: the verbal prefixes mark reversativity rather than simple negativity, while the adjectival prefixes roughly means "not", a simple negation of the quality or state denoted by the adjective. *De-* is not chosen for study here even though it also conveys a reversative meaning, because the verbs associated with *de-* are mostly abstract or mental verbs (e.g., *deactivate*, *decentralize*, *defame,* etc.) and are probably not what children learn at an early stage of lexical acquisition. Furthermore, *de-* is not as productive as *un-* and *dis-* in marking reversativity.

In a monograph on semantic categories, Whorf (1956) used the verbal prefix *un-* to illustrate the notion of 'cryptotype'. In English, one can say *uncoil, uncover, undress, unfasten, unfold, unlock, untie, untangle*, etc., but not *unbury, unget, unhang, unhate, unpress, unspill, unsqueeze,* etc. According to Whorf, the set of verbs that can be prefixed with *un-* seem to share a "covering, enclosing, and surface-attaching meaning". Whorf noted that this meaning falls into a cryptotype, a covert semantic category. Cryptotypes, according to Whorf, are semantic categories that are not marked overtly (as past time, for example, is marked by *-ed* in English), but are only definable negatively in terms of the restrictions they place on how morphemes can be combined. Whorf noted that for the *un*-prefixable verbs, "we have no single word in the language which can give us a proper clue to this meaning ...; hence the meaning is subtle, intangible, as is typical of cryptotypic meanings". However, Whorf also pointed out that despite the difficulty in characterizing the *un-* verbs with an overt semantic label, native speakers of English do have an intuitive feel for what can be prefixed with *un-*. If a new verb is coined with a specific meaning, one can readily accept it or reject it for *un-* prefixation.

Tapping into this native intuition about meaning may seem to be a daunting task, but a detailed analysis of the existent *un-* verbs indicates that there are clear clusters of meanings associated with the *un-* prefixable verbs. These meaning clusters may not be, as Whorf would like, a uniform set that can be described with a single semantic label. In contrast, they may be distinct clusters that are related to one another. In other words, they are different "mini-cryptotypes". For example, many of the *un-* verbs seem to share a meaning of "circular movement" in denoting the activity. They belong to what we may call "rotating or turning verbs", including *uncoil, uncurl, unfold, unreel, unroll, unscrew, untwist, unwind*, etc. Another sub-class of *un-* verbs may be the "binding or locking verbs", which include *unbind, unbolt, unbuckle, unclasp, unfasten, unlease, unlock, untangle, untie*, *unzip*, etc. Yet a third class may share the feature of "covering", which include *uncover, undress, unmask, unveil, unwrap,* etc.

It is important to note that each of these sub-classes is not a unique set, and the meaning associated with the class is not exclusive, due to the properties of cryptotypes. For example, the verb *screw* in *unscrew* may be viewed as having both a meaning of circular movement and a meaning of binding or locking, and the verb *zip* in *unzip* may be viewed as sharing both the binding/locking meaning and the covering meaning. In addition to such crossing-overs or overlaps of meaning features, a verb may also have half of a feature, or .7 of a feature, as part of its inherent meaning. For example, the verb *wrap* may be typically viewed as having the covering meaning. But in some cases the action of wrapping may also involve circular movements. Since the feature of circular movement is a possible but not a necessary component of (or is important but not characteristic of) the verb's meaning, we can propose that *wrap* has half of the meaning feature (in fact, this turns out to be exactly the case in results from our semantic judgment experiment with human subjects, see below). As we will see, these properties, i.e., crossing-over of features and degraded feature composition, make neural network models extremely suitable for studying the mapping problem in question, due to the distributed

nature of knowledge representation in neural networks.

Unlike *un-*, the negative prefix *dis-* in English has received little discussion in the literature. But *dis-* is equally interesting and important in our view. Although the verbs prefixed with *dis-* do not seem to form overt or covert semantic clusters (in contrast to the verbs prefixed with *un-*), they interact with *un-* verbs in significant ways. First, it is clear that *dis-* has its own territories for marking negativity, which is not covered by *un-*. Verbs like *disbelieve, dislike*, and *distrust* are not marked for reversative action, but for negation of a mental state, in which the base verbs *believe, like,* and *trust* denote mental activities. These mental verbs are rarely used with *un-* to marking reversativity. Second and more importantly, *dis-* and *un-* do compete as alternative devices for marking negativity. According to Marchand (1969), these two prefixes (together with *de-*) are rivals in expressing reversative meaning. For example, in some cases, it is arbitrary whether the verb in question takes *dis-* or *un-*. *Connect* and *link* are synonymous words, but *connect* is prefixed with *dis-*, while *link* is prefixed with *un-*. *Entangle* and *tangle* are also synonyms (although they differ in morphological structure), but the former takes *dis-* and the latter *un-*. Similarly, *charge* and *load* also have different reversative forms: *discharge* vs. *unload*. In some other cases, the verb may be prefixed with both *dis-* and *un-* , but resulting in different meanings. For instance, there are two negative forms related to the verb *cover*, *discover* and *uncover*, which have different meanings. Although *uncover* seems to be more like the direct counterpart of *cover* (the meaning of *discover* is more of a metaphorical sense, compared to the positive form *cover*, and *discover* is more like one unit rather than a prefix plus a stem), there is no reason to presuppose that children would not treat *uncover* and *discover* as competing forms. Finally, some *dis-* verbs have counterpart *un-* verbs in their past participle forms, such as *disconnected* vs. *unconnected*, *disconfirmed* vs. *unconfirmed*, and *disarmed* vs. *unarmed*. Again, although the meanings of these pairs are not the same, they indicate the nature of competition between the two reversative prefixes.[2]

Few empirical studies have examined the acquisition of negative prefixes in English. The issue was first raised and studied by Bowerman (1981, 1982). Bowerman pointed out that most previous studies of the acquisition of morphology have been exclusively concerned with children's understanding of the morphological markers themselves, regardless of the meaning of the words with which the markers co-occur. She used *un-* as an example to illustrate the importance of semantics, in particular, cryptotypes in children's acquisition of prefixation patterns. Consistent with other cases of overgeneralization, Bowerman's data indicate that children also display a U-shaped pattern in learning *un-*. They initially treat *un-* and its base verb as an unanalyzed whole and produce the *un-* verbs in appropriate contexts (e.g., *unbuckle, unfasten, untangle*, etc., analogous to the child's saying *went* without realizing it as the past tense form of *go*). At a later stage, in which the reversative meaning of the prefix is extracted, they start to use *un-* with a variety of verbs, producing errors such as *unbury, unget, unhang, unhate, unpress, unspill, unsqueeze,* etc. The role that semantics play in the acquisition of *un-* is clearly shown in how children recognize shared aspects of meaning among the verbs that can take *un-* and how they eventually recover from errors on the basis of their semantic analysis. As Bowerman shows, the speed with which children recover from errors may be different: in some cases, children may produce incorrect combinations of *un-* with verbs that fall outside Whorf's cryptotype of "covering, enclosing, and self-attaching", and later on narrow down the application of the prefix to only the verbs that fall into the cryptotype; in other cases, they may identify the semantic characteristics of *un-* verbs simultaneously as they work out the reversative meaning of the prefix, and thus produce no novel forms that violate the cryptotypic meaning.

There has been no report on the acquisition of the prefix *dis-*, although *dis-* is equally interesting and important as *un-* (because of its competition and interaction with *un-* as a negative prefix, as discussed earlier). One reason for the lack of study of

[2] Although we should note here that *dis-* and *un-* have different historical backgrounds (*dis-* is a Latinate form and *un-* is of native origin dating back to Old English), it is hard to conceive that the etymological difference would affect children's acquisition of these prefixes. See Pinker (1989) for a different kind of arguments with respect to how etymology may influence children's acquisition of syntactic structures.

the acquisition of *dis-* may be due to the limited productivity of this prefix in everyday speech. Moreover, many *dis-* verbs are considered as a whole rather than a base verb plus a negative prefix. In some cases, there is no positive form of the *dis-* verb, e.g., *discuss, disturb,* and *distort.* In other cases, there is no apparent meaning relationship between the positive verb form and the negative *dis-* form, e.g., *dismiss, dispose*, and *display.* Furthermore, *dis-* is used for some abstract mental verbs (see earlier discussion) that children are unlikely to use until much older. These facts suggest that the child may have to learn many of the *dis-* verbs by rote. Generalization of *dis-* to novel forms may be suppressed by this kind of rote learning. However, the limited productivity of *dis-* in the adult language does not necessarily suppress the child's early ability to generalize, especially if the child has encountered both *un-* and *dis-* in the same kind of negative context. Therefore, by examining the performance of *dis-* (together with *un-*) in our network we will be better able to understand the processes involved in the acquisition of negative prefixation.

Using neural network modeling, this study will investigate the process of overgeneralization and recovery with both the prefixes *un-* and *dis-*. It looks into the psycholinguistic mechanisms underlying these processes, i.e., the conditions under which overgeneralization and recovery take place and the factors that influence and determine these processes. Because empirical studies are generally constrained in their flexibility of systematically manipulating candidate determinants, neural network modeling, by varying relevant factors at a micro level, can provide important information about language acquisition that is not readily available in natural empirical studies. Furthermore, the properties of neural networks are ideal for dealing with the kind of problems we have here, i.e., learning the negative prefixes in connection with learning cryptotypes. Recall that cryptotypes are covert semantic categories whose meaning is elusive and the structure of which is hard to capture with single labels. While it seems hard for symbolic accounts to come up with satisfactory explanation of Whorf's problem (i.e., by using crisp and principled rule-like descriptions for cryptotype and its acquisition), the properties of neural networks (i.e., distributed representations, nonlinearity in information processing) seem to be ideal for handling the elusiveness of cryptotypic semantic structures. The point will become clearer as we discuss our results.

The organization of the paper is as follows. First, input representation, network architecture, and training and testing schedule are presented in the method section. Next, major simulation results are presented and analyzed in the results section. Finally, in the general discussion, the simulation results are evaluated with respect to their theoretical significance for understanding the psycholinguistic mechanisms of over-generalization and recovery.

METHOD

The input data

In this study, the network was trained to map meanings of English verbs onto the different prefixation patterns. Thus, the input data to the network were different verbs encoded as semantic feature vectors. 105 verbs were selected for this study, on the basis of two sources: The *Webster's Ninth New Collegiate Dictionary* and the Kucera & Francis (1967) corpus. Our data set consisted of 47 *un-* verbs, 18 *dis-* verbs, and 40 *ø* verbs which take neither *un-* nor *dis-*. The Webster's dictionary contains more verbs prefixed with *un-* or *dis-*, but many of these are unacceptable to modern-day English speakers, such as *unwish*, *unlive*, and *disannul.* The final selection of the 47 *un-* verbs and 18 *dis-* verbs was based on native speakers' judgments of the acceptability of all the *un-* verbs and *dis-* verbs that appeared in the Webster's dictionary (except those marked 'archaic'). In the case of *un-*, fourteen subjects were asked to rate how good each verb sounded to them on a scale of 1 to 7 (1 means that it sounds completely weird, 7 means perfectly natural). Only those verbs which on the average scored 5.0 or more were selected. In the case of *dis-*, we did not run such a group experiment and simply rely on the scores for verb selection, because the *dis-* verbs that appeared in the Webster's constitute a very heterogeneous set with respect to their status of parts of speech (for example, both *discontent* and *disadvantage* were marked as transitive verbs in the Webster's dictionary, but modern-day English speakers would use *discontent* as an adjective and *disadvantage* as a noun). Moreover, many of the *dis-* verbs do not have

their positive counterpart forms, e.g., *dispel, distort,* and *disturb* (see earlier discussion), and therefore cannot be used in our network in which only positive verb forms without prefixes are used as input for the mapping task. The final 18 *dis-* verbs were selected on the basis of consultation with a few native speakers and on removing some highly abstract items such as *disentail, disfranchise* and *disinherit* (children will probably not learn these words until a later stage, after they have grasped the general use of *dis-*).[3] Finally, there were also 40 plain verbs that cannot take either *dis-* or *un-* in this study. These were all high frequency verbs selected from Kucera & Francis (1969). These 40 verbs were included in order to have in the input data some verbs that cannot be prefixed in contrast to the ones that can. A complete list of the input verbs can be seen in Figure 1.

Each of the 105 verbs was encoded as a set of semantic features with different values. Because there has been no detailed linguistic analysis of the *un-* and *dis-* verbs (except Whorf, 1956), the selection of appropriate semantic features was a difficult task. Twenty semantic features were chosen for the current study (see Table 1 for a complete list of these features), partly on the basis of reviewing relevant literature (Whorf, 1956; Marchand, 1969; Pinker, 1989), and partly on the basis of our own linguistic analysis.[4] These features are supposed to capture the semantic range of the verbs that can be prefixed with *un-* and *dis-*. Of course, the features are also relevant to many other verbs that do not take *un-* or *dis-,* but it is the combination of the various features that may distinguish the verbs that can be prefixed from the ones that cannot.

Instead of 'hand-crafting' the specific values of each semantic feature into the network (which is easily subject to criticism, see Lachter & Bever, 1988), this study used a more natural featural representation, derived from a semantic judgment experiment with human subjects. In the experiment, fifteen native speakers of English were given a list of 105 verbs and a list of semantic features (the twenty features we have chosen for study), and were asked to evaluate each verb with respect to each feature and determine whether the particular feature applies to that particular verb. Subjects marked the verb with a cross if they thought that the feature was characteristic of, or typically involved in the situation denoted by the verb. Thus, for each subject, we had a feature-by-verb matrix of 0s and 1s (0 means that the feature is irrelevant to the verb, and 1 means that the feature is relevant). The averaged rating scores from the fifteen subjects were therefore graded representations that were used as input to the network. In this case, each verb was encoded as a vector of the twenty features with values between 0 and 1. Some examples are given in Table 1.

Insert Table 1 about here

To evaluate the validity of the results from our semantic judgment experiment, a hierarchical cluster analysis was conducted on the verbs encoded as vectors of features. Figure 1 shows that the relative meaning distances between different verbs as revealed by this analysis are consistent with native intuitions about how similar the words are (i.e., synonymous words tend to group together as clusters).[5] Verbs that are similar to each other are grouped together lower in the tree, while clusters of words that are similar to other clusters are connected higher in the tree. For example, at the bottom of the graph, the words *connect* and *link* form the smallest cluster. If we examine Table 1, we can see that these two words share the majority of features in common. Some of the features have exactly the same values, others differ by small degrees. In contrast, the verb *turn* is very different from either *connect* or *link*. Features 4, 8, and 16 characterize both *connect* and *link*, but are almost completely irrelevant to *turn*.

In Figure 1, we can easily discern a few big clusters: toward the top there is a cluster which indicates 'turning or rotating verbs', most of which take the prefix *un-* (*roll,*

---

[3] I owe my special thanks to Cathy Harris and Edward Yang for their help on the selection of the *dis-* verbs.

[4] I should especially like to acknowledge Cathy Harris's constructive discussions on the relevant semantic features selected for this study.

[5] The capitalized marker after each verb in Figure 1 is a mnemonic for the prefixation pattern of each verb. The actual input consists of the base verbs without any prefixes. This also holds for other figures of this paper.

reel, turn, wind, twist, coil, curl, etc.) Further down we can discover the cluster of 'visual or auditory activities', none of which takes negative prefixes (*talk, ask, say, hear, see, look, speak,* and *tell*). Somewhere to the bottom, we can find a cluster of 'binding or locking verbs', most of which can be prefixed with *un-* (*bind, buckle, lock, plug, clasp, fasten*, etc.) These results indicate that there are meaning structures inherent in the input, and the network (the child too, we believe) is learning the mapping between these structures and negative prefixes, rather than learning a messy, unorganized set of verbs for prefixation patterns.

Insert Figure 1 about here

In addition to the semantic clusters, we can also see that the tree splits into two big branches according to whether or not the verb can take prefixes. Figure 1 is clearly divided into a cluster which takes no prefix (the upper part) and a cluster which takes the prefix *un-* (the lower part), although there are exceptions (the "rotating verbs" are grouped with the ø verbs). The verbs that take *dis-* are scattered in both clusters. These results provide further evidence for a naturally structured learning target. It also provides support for the validity of our rating experiment.

Network architecture

A standard three-layer back-propagation network is used for all simulations in the current study (Rumelhart, Hinton, & Williams, 1986). In this type of network learning, information is first encoded at the *input layer,* then it funnels through the *hidden layer*, where internal representation is formed, and finally results are given at the *output layer*. Each layer consists of different units, representing different states/processes of information processing (from input to output). Learning in this case is a function of adjusting the weights and connections between different units. The adjustment is done according to the back-propagation algorithm, with which the network discovers a discrepancy between its actual output and the desired output in the teaching signals, then an error term is propagated back through the system, and accordingly weights and connections are changed.

Figure 2 illustrates the architecture in which the network was configured in this study. There are 20 input units encoding different semantic features of verbs, 6 hidden units encoding the internal semantic representations during learning, and 3 output units representing either *un-*, *dis-*, or no prefixation (*ø*). All the simulation results to be reported below were based on the same architecture. All simulations were conducted using the TLEARN program configured at the Center for Research in Language, UCSD.

Insert Figure 2 about here

Task and procedure

The task for the network was to learn to map verbs (encoded as semantic feature vectors) onto one of the three prefixation patterns: *un-*, *dis-*, or *ø*. That is, each time the network was given an input pattern,[6] it was required to output an *un-*, *dis-*, or *ø* form of the verb. There are two types of simulations in our study. The first is that the network is presented with the complete corpus right from the beginning, as in Simulation 1 below, in which case all instances in the data set are seen by the network simultaneously (though in a random order). The second, as in Simulations 2 and 3, is that the network is first trained on a subset of the whole corpus (i.e., some randomly selected instances of the 105 verbs) until it achieves asymptotic performance on correct prefixation for all the items in the subset, and then some novel instances (i.e., instances that the network has not seen during initial training) are introduced to the system to test how well the network generalizes on the basis of what it has been trained on.

---

[6] For lack of better terminology, the word 'pattern' is used in this paper both in context of *learning patterns, overgeneralization patterns,* and in context of *input/output patterns, hidden unit activation patterns*. The latter use is more of a technical sense, meaning 'vectors' in network modeling. Although in context it is clear which sense is intended, the reader may need be reminded of this difference here.

To evaluate network performances during training and testing, we use the root mean squared (RMS) difference scores for report of all simulation results, i.e., the squared differences between the actual values generated by the network and the desired values in the teaching signals. As will be seen (e.g., in Figure 3), learning in this case is a function of the decrease of error rates over time, with time being represented as "learning cycles" in our simulations.[7] For global learning performances of the network, RMS errors will be used to show the time course of learning, summed over at a given interval specified by the experimenter (every 1000 instances in this study, see below). For individual patterns (i.e., different verbs), RMS errors indicate how distant the actual output deviates from the desired output. In this study, our network is considered to have mastered the relevant mappings if the global learning error drops to below 0.1 and the error for individual patterns drops to below 0.01.

The network's performance is also assessed with hierarchical clustering analysis (see Elman, 1990 for application of this method in network analysis). This technique allows us to discover the relative meaning distances between different verbs that are represented at the hidden unit level, at each point in time during learning. In the discussion of input patterns, we used this method to analyze the input patterns, and the results there revealed some clear meaning clusters. In the reports below, we will use this method to analyze the hidden unit activation patterns in order to determine if the network has discovered meaningful structures inherent in the meaning-form mappings during the course of learning.

RESULTS AND ANALYSIS

In this section, I will present a number of simulation results that are relevant to different aspects of the theoretical issues of overgeneralization and recovery. First, I will analyze the performance of the network trained on all of the 105 instances of verbs, showing that the network is capable of handling the mapping problem at issue; more importantly, in learning the correct prefixation mappings, it discovers semantic clusters of verbs that correspond to what Whorf called *cryptotypes*. Second, I will report results from testing the network's ability to generalize and patterns of overgeneralization that may occur in the generalization process. Third, I will narrow the focus down to some individual pairs of verbs that share semantic features in common and compare their performances at different phases of learning. These results are later integrated in a general discussion.

Simulation 1: Discovering inherent semantic structures

In this simulation, the network was trained on the complete data set (105 verbs). The order in which each verb was presented to the network was randomized. Figure 3 shows the global error decrease within a time frame of 80000 learning cycles, averaged over 1000 randomly sampled patterns at each point.

Insert Figure 3 about here

As seen in this graph, error rates started high but gradually decreased to a level of near error-free performance. A sharp drop occurred at around 20000 cycles, and at around 60000 cycles, errors dropped drastically further down (RMS below 0.02) and all verbs were learned for their correct prefixation patterns at around 80000 (RMS below 0.01). These results are straightforward in showing that the negative prefixation problem could be solved by a network like ours in which there is very little predisposed structure in the learning mechanism. Learning takes place in a simple task of mapping meanings onto forms, in which all the inputs were seen by the network in a random fashion.

Inherent in this straightforward learning curve are the interesting individual patterns for different verbs. Figure 4 plots a few examples which differ dramatically in how fast they were learned with respect to the prefixation patterns. The verb *fold* was

[7] A *cycle* consists of a single forward activation and a backward weight adjustment pass, i.e., a random presentation of a verb to the network and the network's processing of it. For readers familiar with *epochs*: one epoch is the number of cycles equivalent to the number of instances in our training data.

learned correctly almost right from the beginning, while the verb *do* had not been learned within the given time frame (30000 cycles). The verb *appear* is somewhere in between, in which its error rates started high at the beginning and dropped down after a period of vacillation.

Insert Figure 4 about here

It is not surprising that different verbs would behave differently in how fast they are learned. However, it is interesting that the learning speed reveals a function of how well a particular verb fits the inherent semantic categories (what Whorf may call 'cryptotypes') with respect to the prefixation patterns. Notice that the verb *fold* falls nicely into a cluster of "rotating verbs", or verbs that indicate a circular movement, most of which take the prefix *un-* (see Figure 1). Learning to map *fold* onto *un-* in this case should be easy, once the network discovers the correlation between the rotating verbs and the *un-* prefixation. In contrast, the verb *do* does not fall neatly into any clear clusters, and it belongs to the big ø branch in which most verbs are not prefixed with *un-* (see earlier discussion of Figure 1), while *do* itself does take the prefix *un-*. The verb *appear* is in a similar situation as *do*, with most of its close neighbors taking no prefixes (e.g., *come, run, go, walk*, etc.) while the verb itself can be prefixed with *dis-*. However, *appear* managed to get out of trouble earlier than *do*, even though they are in a similar situation (Figure 5 shows that by about 30000 cycles of learning, *appear* has been grouped together with other *dis-* verbs, e.g., *believe, like*, and *embarrass,* see below). This faster learning rate with *appear* could be due to the fact that the set of verbs that can be prefixed with *dis-* is much smaller than the set of *un-* verbs, so that it is easier for *dis-* verbs to get out of attractions and competitions from neighborhood items.

In the next several graphs, I will present some cluster analyses of the hidden unit activation patterns of the network that has just been described; these hidden unit activation patterns are the network's internal representations of the verbs. The analyses illustrate how the network discovers inherent semantic categories or cryptotypes during the course of learning.

Figure 5 shows the tree structure of all the trained verbs that are represented at the hidden unit level at 30000 cycles, at which time both *fold* and *appear* have been learned, but *do* has not. Although the network has not learned all verbs in question at this point (Figure 3 indicates that that is accomplished around 60000 cycles), there is evidence that the network has already discovered a lot of structure in the input, with respect to both the prefixation patterns and the semantic categories.

The first clear cluster that the network has discovered is the *dis-* verbs. With only a few exceptions, all the *dis-* verbs have been grouped together. Even the exceptions show some internal consistency in that they are also grouped together (i.e., *entangle, believe, like, appear,* and *embarrass*). This shows that the network has singled out all the verbs that can be prefixed with *dis-*, probably not on the basis of meaning, but simply on the basis of the mapping pattern. Because the set of *dis-* verbs is relatively small, discovering the cluster of *dis-* verbs is not a very difficult task for the network.

Insert Figure 5 about here

A second cluster that emerged from the analysis in Figure 5 is the set of verbs that all share the semantic feature of 'circular movement'. It can be seen that, except one verb, all of what we called "rotating or turning verbs" have been grouped together at this point of learning. Interestingly, the only exception is the verb *turn*, which differs from all other rotating verbs in that it does not take the prefix *un-*. This shows that the network has not only discovered the category of rotating verbs on the basis of meaning, but also discriminated words on the basis of their prefixation patterns. It should be clear that these clusters or structures have emerged from the internal representations of the network as a function of learning, not as a property that is given at the beginning. In Figures 6 and 7, I analyze the network's representations at two initial stages and show how the network discovers the semantic category of rotating verbs.

Insert Figure 6 about here

Figure 6 represents the hidden unit activation patterns for each verb at a very early stage, 200 cycles. In contrast to what was found in Figure 4, the network has formed no apparent structure for the category of rotating verbs, as seen in that the different rotating verbs (indicated with arrows) are scattered all over without being grouped at all. Turning to Figure 7, which is how the verbs are represented by the network at 1000 cycles, we found that the network has developed some structures for the rotating verbs at this stage. These verbs now tend to come together as groups, forming small clusters at several levels. Although they have not been completely grouped together, and occasionally there may be one verb left alone in no groups (e.g., *fold*), the results present us with a snapshot in which the network is making its effort in discovering the semantic categories.

Insert Figure 7 about here

Comparing Figure 7 with Figure 4 we find that the representation of some individual verbs also differs as a function of learning the mapping between semantic categories and prefixation patterns. In Figure 4, it was seen that the network singled out the verb *turn* from the other rotating verbs since *turn* does not take the prefix *un-* while all others do, even though semantically it belongs to the rotating category. In Figure 7, *turn* apparently has not been isolated from other rotating verbs; instead, it is treated by the network similarly to *reel* and *roll* in the midst of forming the semantic cluster of rotating verbs. At this stage, there is no apparent structure in the network's internal representation with respect to the prefixation patterns. Only at a later stage of learning (e.g., at 30000 cycles, as represented in Figure 4) does the network distinguish *turn* from its semantic associates by finding out their mapping differences in prefixation. Although in a different way, these analyses for *turn* also applies to *appear*, a verb that we examined earlier. In Figure 7, *appear* is with other motion verbs (i.e., *walk, come,* and *run*) in a small cluster, while in Figure 4 the small cluster has broken down and *appear* has been grouped together with other *dis-* verbs.

To summarize the results in Simulation 1, first, our network is able to solve the prefixation problem in a simple task in which all verbs were learned at once and in which the learning mechanism has very little built-in (i.e., innate) structure. Second, the speed with which a particular verb is learned seems to be a function of how well the verb fits the cryptotype and its prefixation pattern. Third and most importantly, the network has developed different internal representations for the verbs at different stages of learning on the basis of prefixation patterns and meaning patterns. Our cluster analyses show clearly the dynamic process in which the network discovers covert semantic categories or cryptotypes inherent in the mappings between verb meanings and prefixation patterns.

Simulation 2: Learning novel input

In this simulation, a new network was trained on a subset of the whole corpus, and then presented with some novel instances to test how well it can generalize on the basis of what it has learned earlier in the training session. This training and testing regimen is comparable to a learning situation in which the child has extracted some basic patterns from the data and is faced with learning new lexical items for the same mapping problem. Typically, the child may produce two kinds of performance results in such a situation: (1) generalize correctly on the new items on the basis of the learned pattern, because the new items fit in with the learned pattern; (2) overgeneralize on the new items, because the new items do not fit in with the learned pattern but the child thinks they do and applies the learned pattern to the new items. Neural networks are ideal for simulating this kind of performance because of its flexibility in controlling relevant factors that may cause overgeneralization and influence the timing of recovery.

Figure 8 presents the results when the network was trained on 85 of the 105 verbs and tested on the remaining 20 verbs. Items were randomly selected for training and testing. Up to 20000 cycles, the network was trained on only the 85 verbs. At 20000 cycles

when the network learned all the 85 verbs to near error-free, the 20 novel verbs were introduced into the system with continuing learning. It can be seen that the error rates went up radically as the new items came in, indicating that the network made overgeneralization errors at the beginning when encountering new items. Errors decreased gradually as learning continues, showing the network's eventual recovery from all overgeneralizations. Notice that in this particular training and test schedule, it actually took the network longer to learn the 20 verbs in the test set than to learn the 80 verbs in the training set, which could be due to some particularly difficult items in the test set.

Insert Figure 8 about here

Although the error rates were initially very high when the network was exposed to new test verbs, not all verbs were incorrectly overgeneralized with respect to their prefixation patterns. In Figure 9, we chose three verbs of different nature to illustrate that not only do verbs differ in how fast they are learned, as discussed in Simulation 1, but they also differ in whether they are overgeneralized and how fast they can recover from overgeneralizations.

Insert Figure 9 about here

Figure 9 shows the RMS scores of three verbs in the testing phase (after 20000 cycles). It happens that the verb *appear* was in the test set, a verb that we discussed in Simulation 1. Similarly to what was found in that simulation, *appear* displayed high initial learning errors. However, unlike in Simulation 1, *appear* was not seen by the network until a later stage, when the network had learned many other verbs. The other two verbs, *twist* and *clasp*, were introduced to the system at the same time as *appear*, but their results were drastically different from that of *appear*. There were almost no errors for these two verbs. The contrast between *appear* and *twist* here reminds us of the contrastive pattern found in Simulation 1, in which the learning speed of *appear* and *fold* were dependent on how well the verb in question fits the cryptotype and its prefixation pattern. Note that *twist* is a verb like *fold* that falls into the rotating verb category. These results suggest that the fit to cryptotypes influences not only the speed of learning, but also the network's generalization ability. *Appear* was overgeneralized by the network initially because it involved learning a motion verb that differs from other motion verbs in prefixation. In contrast, there were no overgeneralizations for *twist* and *fold* because these verbs fell into a cluster of rotating verbs that all take the prefix *un-* (except *turn*). It was shown in Simulation 1 that the network quickly discovered the cryptotype of rotating verbs, so the learning errors of *fold* decreased rapidly. In this simulation, the network had already formed some structure for rotating verbs before learning the verb *twist* (see Figure 10), so the new verb was incorporated into the right category and generalized correctly from the beginning. The same holds true for the verb *clasp*, which belongs to a large cluster of verbs that share a locking or binding meaning, all of which take the prefix *un-* (also see Figures 10 and 11).

Figure 10 shows the network's hidden unit activation patterns at 20000 cycles, when all verbs in the training set were learned correctly. It is clear in this graph that the network has developed structured internal representations for at least three semantic clusters: motion verbs, locking or binding verbs, and rotating verbs, even though these clusters differ in size. Figure 11 represents the very beginning of the testing phase, when the network was just being exposed to the new test items at 20000 cycles. This graph indicates how the network squares the relevant verbs into the corresponding category. Apparently *twist* and *clasp* fit right into the prefixation pattern of their neighbors in the cluster, while *appear* is surrounded by competitors that take no prefixes.

Insert Figures 10 and 11 about here

The above results are also observed in other trainings using a different combination of training vs. test items. Figure 12 shows the individual learning patterns of three verbs after training on another set of verbs (again, 85 randomly selected verbs in the training set and

20 in the test set). Although this time the network did not learn the verbs in the training set until 30000 cycles, these results here are completely consistent with results presented in Figure 9, in which verbs sharing the same meaning and taking the same prefixes as members of the cryptotype were generalized correctly while verbs sharing the same meaning but having different prefixation patterns were overgeneralized. *Buckle* is similar to *clasp* in Figure 9, and *coil* is similar to *twist*. *Entangl*e is more of an extreme case than *appear*: the network could not recover from overgeneralization within 60000 cycles (although eventually with prolonged learning it did fully recover from errors, see Figure 15 below). Looking into the network's internal representations at different learning levels we found that the reason it was hard for *entangle* to recover is due to its close synonymy with the verb *tangle* which belongs to a cluster of binding or locking verbs, all of which take *un-*, while the verb *entangle* itself takes *dis-*. An inspection of the input representations of *tangle* vs. *entangle* shows that they share the majority of features in common and thus constitute close synonyms.[8] Again, these results indicate that verbs that have clear semantic associates but do not undergo the same prefixation pattern of these associates will be most easily susceptible to overgeneralization and more difficult to recover from errors; in such cases, the semantic associates become grammatical competitors of the learning target. We will return to the issue of how semantic similarities affect learning and recovery later in more detail.

Insert Figure 12 about here

Finally, although verbs that fall nicely into cryptotypes like *coil, twist, clasp,* and *buckle* were generalized correctly from the beginning, as we have seen, they are not necessarily impervious to overgeneralization at all times. In the above training schedule, novel verbs were introduced to the system only when all verbs in the training session were learned correctly, at which time the network has developed structured internal representations for the verb meanings. What happens when the network has not formed semantic structures for relevant verbs when exposed to new items? This is a very realistic situation in child language acquisition. In Figure 13, we plotted results of several verbs from testing the same novel items in the same network as reported in Figure 12, except that novel items were introduced to the network at a very early stage (200 cycles), when the network is unlikely to have formed any sensible structured representation for the semantics of the verbs.

Insert Figure 13 about here

This graph shows clearly that *buckle, clasp, coil,* and so on are also subject to errors, if no sensible structures have emerged from the data that the system has learned at the time when it is exposed to new items. Recovery in this case is a function of learning the semantic structures of the relevant verbs. Note that in Figure 13 the recovery process is very fast for both the rotating verbs and the locking/binding verbs, due to the network's quick discovery of relevant meaning clusters (see also earlier discussion in Simulation 1); in real learning situation this need not be the case. The fact that the network displays different learning patterns for the same verbs bears important significance to child language acquisition, because it shows that overgeneralization and recovery have to do with the existing knowledge and its organization of the learning mechanism. However, given the exploratory nature of our simulations in this respect, any conclusions about how children acquire new patterns on the basis of old information must remain speculative.

To summarize results in this set of simulations, verbs that differ in cryptotypic meanings and prefixation patterns are not only different in the overall speed of learning when all words are learned simultaneously, they are also differentially susceptible to overgeneralization and differ in speed of recovery. When novel input is consistent with the prefixation patterns of the learned structure, no overgeneralization occurs. Specifically, if the new item takes the same

---

[8] Although these two verbs differ in their morphological structure, they are nonetheless represented in our input similarly; recall that our semantic encoding did not include phonological, morphological, or historical differences.

prefix as its semantic associates in a cryptotype, the network generalizes correctly from the beginning. Overgeneralization occurs when the new input is inconsistent with members of the cryptotype in prefixation. Recovery is hard when the new item shares a large number of semantic features with, but takes a different prefix as members of its meaning category. Finally, new input is more susceptible to overgeneralization when the learning system has not formed any sensible semantic structures, while the same input may not be overgeneralized if it is presented to the system at a later stage when semantic structures have clearly emerged from learning.

Simulation 3: Synonymy effects in recovery

We have already observed some synonymy effects in Simulation 2, in which the difficulty of recovery with *entangle* was attributed to its close synonymy with the verb *tangle* in the training set, which has a different prefixation pattern. In this simulation, we are particularly concerned with how synonymy, i.e., semantic similarity, may affect the recovery process.[9] There have been a number of hypotheses about how synonymous words may compete with each other to influence the learner's eventual correction of over-generalization errors in the child language literature (Bates & MacWhinney, 1987; Clark, 1987; Bowerman, 1987). Although the current study is not specifically designed to test the alternative hypotheses, the results from our simulations will shed light on the psycholinguistic mechanisms that govern the recovery process.

In this simulation, we systematically studied a number of synonym pairs by training on one member of the pair and testing on another. The specific pairs that were trained and tested on are: *link* vs. *connect, tangle* vs. *entangle,* and *load* vs. *place*. Note that the two members of each pair all take different prefixes in the output: un*link* vs. dis*connect*, un*tangle* vs. dis*entangle*, and un*load* vs. dis*place*. Although these pairs differ in their exact degree of synonymy (i.e., they share a different number of semantic features or share the same number to different degrees in our input representations), the same kind of results were found for all pairs, even though the exact form of recovery may be different for different pairs.

Figures 14a and 14b show the results of two different pairs of synonyms in the test phase, in which *entangle* and *connect* were introduced to the network as novel input after the network had learned the other 104 verbs (including *tangle* and *link*) in the training phase. Error rates were sampled at every 500 cycles for both graphs.

Insert Figures 14a and 14b about here

What is clearly revealed in these graphs is that the network's performance with the novel input (i.e., *entangle* and *connect*) is almost a complete mirror image of the performance with its synonyms (i.e., *tangle* and *link*): as error rates drop down with one verb, they go up correspondingly with the other verb. In Figure 14a, errors with *entangle* were initially high and gradually decreased; errors with *tangle* gradually increased until around 75000 cycles, and then dropped down together with *entangle*. In Figure 14b, the same kind of contrastive pattern for the novel input and its synonym can be observed, but the recovery process took a different form: error rates went up and down radically at the beginning for both *connect* and *link*, as if the system was experiencing radical and unstable changes. This contrastive pattern of performance, i.e., the decrease of errors for one verb mirroring the increase of errors for its synonym, can be observed in all phases of learning, although to a lesser extent in the later stages of the network's recovery.

It appears that the mirror images in our simulation reflect some kind of competition effect between the novel input and its synonyms. In principle, synonyms *attract* each other for the same mapping because of their semantic similarities, but in effect they may *compete* with each other because of their different mapping patterns that is at issue (such as in the above cases). Each time the strength of the mapping of a verb to the same prefix as its synonym increases, the error rates for that verb also increase, and accordingly

---

[9] In talking about synonymy, I am not committed to particular semantic theories on the nature of synonymy. The use of *synonymy* or *synonym* is a convenience for labeling semantic similarity or words that share semantic similarities.

weights need to be adjusted. In this case, learning proceeds as a function of suppressing the strength of mapping synonyms to the same prefix. Only by gradual adjustment of activation strengths (in a contrastive way) could the system learn to map similar input to different output. Therefore, the attraction and competition effects could well result in mirror images of the kind seen in our results. Although Figures 14a and 14b show very different forms of recovery, the fact that both show mirror images indicates that the synonyms were in competition in both cases. Note that this competition is not a competition for the same mapping (as described in other studies, e.g., Bates & MacWhinney, 1987; Bowerman, 1987; Clark, 1987), but competition for different mapping patterns between items in the same category. The different forms with the two pairs in the graphs are the results of two different ways of getting out the problem by competition: in Figure 14a, the system learned *entangle* in a kind of conservative way, so that the learning is a function of smooth transition from high errors to error-free performance (the disturbance to its synonym was also of a gradual nature), whereas in Figure 14b, the system was eager to get out of trouble, resulting in radical swings of learning errors, but also consequently in rapid recovery.

Earlier in Figure 12, we showed that synonyms that undergo different prefixation patterns are susceptible to overgeneralization and difficult to recover from errors. It was seen that the network could not learn the correct prefixation of *entangle* within 60000 cycles. In Figure 15, we plotted the same results of *entangle* from Figure 12 (with prolonged training) against the results of its synonym *tangle*. The form of the recovery here is not exactly the same as in Figure 14a (there was a period of rough zigzags between 50000 and 60000 cycles, indicating the instability of the system during competition). But consistent with the results from the current simulation, the synonyms also displayed the same contrastive pattern, providing further evidence for our analysis of the competition effect.

Insert Figure 15 about here

Although the above analyses showed results only for pairs of synonyms, the competition effect between synonyms is not limited only to pairs. As shown in Figure 16, learning the new verb *place* (prefixed with *dis-*) caused trouble for the already learned verbs *load* (prefixed with *un-*) and *put* (prefixed with none), both of which are synonyms of the novel verb. The troubled verbs displayed very similar patterns (both were mirror images of *place*), except that at the beginning *load* had higher error rates than *put*.

Insert Figure 16 about here

It is important to note that prior to learning the novel verb, the network has already learned the correct prefixations of all the verbs in the training stage, including the synonyms of the novel verbs such as *link, tangle, load,* and *put*. Apparently, the recurrent errors with the old learned verbs are due to the disturbance or interference from the network's learning of novel verbs in the new test phase. This disturbance may take different forms. In some cases, the introduction of the novel verb seem to only gradually cause trouble for its synonym (as in Figure 14a), whereas in other cases, learning the new verb may cause dramatic changes for its synonym right away (as in Figure 14b). But both indicate that learning to map similar input to dissimilar output can disturb the learned patterns in the old data. We expect that this kind of disturbance or interference will not occur when the novel verb and its synonym undergo the same prefixation pattern, i.e., when the system learns to map similar input to similar output. In fact, we have conducted some simulations in which the novel verb (e.g., *entangle* or *connect*) was artificially set to take the same prefix as its synonym (i.e., *un-* as in *untangle* or *unlink*), and found that there were no recurrent errors with the synonym at all.

In this connection, we wish to note that the recurrent errors constitute part of an interesting U-shaped pattern (see Figures 14a and 15, less obviously in Figures 14b and 16), in which a verb has been learned at an earlier stage, errors reoccur with it after new items come in and only later drop together with errors of the new items. Also note that this U-shaped pattern is a little different from

traditional accounts of U-shaped function, because it follows a stage in which the learning mechanism has already experienced a period of erroneous to correct performance, rather than starting with an initial correct performance as described in traditional accounts.

To summarize results in Simulation 3, we have found that the learning of new verbs may disturb already learned patterns of old verbs, in particular, synonyms that undergo different patterns of prefixation as the new verbs. The mirror images found with various pairs of synonyms show clear contrastive patterns between words that have similar input but different output, reflecting effects of attraction and competition between the old and the new items. Overgeneralization in this case is the result of synonymy and recovery is the result of competition.

GENERAL DISCUSSION

In this study, we have used neural network modeling techniques to examine a classical phenomenon in language acquisition, i.e., overgeneralization and recovery. The goal of the study is to uncover the psycholinguistic mechanisms underlying the learning processes, in hope of providing insights into functional constraints on the processes of overgeneralization and recovery. The focus of the study has been on the role of semantics in the acquisition of English negative prefixes. Although the results described here are exploratory in nature, they suggest that neural networks are not only sensitive to aspects of acquisition in the form-form mappings, but also sensitive to the complex interactions between meanings and forms. Neural networks, because of the properties of distributed representations, are ideal for dealing with the kind of semantic problem that Whorf (1956) worried about, i.e., the mapping between cryptotypes and prefixation patterns.

In the connectionist perspective, Whorf's cryptotype problem can be solved relatively easily by the use of distributed representations. Whorf described the meaning of cryptotypes as 'subtle', 'intangible', and not characterizable by a single label, but he also noted that native speakers do have an intuitive feel for it. The reason that cryptotypes cannot be described with a single label, in our view, may be that the meaning features that unite different members of the cryptotype are represented in different items in a complex distributed fashion, so that they are not easily accessible to explicit symbolic analysis, but are accessible to native intuition. It is therefore not the case that the meaning of the cryptotype itself is intangible, but that the semantic relationship between different items is not easily subject to symbolic analysis. In a sense the meaning of a cryptotype constitutes a semantic network, in which verbs differ from one another with respect to (a) how many features each verb contains, (b) how strongly each feature is represented in the verb (e.g., whether the feature is a necessary component of the meaning, whether it is distinctive and critical, etc.), and (c) how strongly verbs are connected to one another and to what extent a verb qualifies as a member of the category. Learning to map prefixation patterns to verb meanings in this case is no longer learning to apply a rule (in the traditional sense) to a class of verbs, but learning the activation strengths of the connection between a particular prefix and some set of semantic features inherent in particular verbs. In doing this, the learning mechanism groups together those verbs that share a large number of semantic features and that take the same prefix. Thus, the verbs gradually form clusters with respect to both their meaning patterns and their prefixation patterns. Rather than being a uniform set, the verbs are connected to each other with different strengths even though they may be maximally different from other verbs; together they constitute what Whorf might call *cryptotypes*.

Results from Simulations 1 and 2 clearly indicate the plausibility that cryptotypes are emergent properties of distributed semantic representations rather than pre-defined meaning categories. Our input representation consists of only equally distinctive semantic features of verbs, and there is no categorical feature or features that tell whether a verb belongs to a certain semantic category. Our features are represented in a highly distributed fashion, in that the same feature has different strengths in different verbs (i.e., a verb might have .5 of a feature) and that the same verb has different features with different strengths. However, with an appropriate amount of learning, the network develops categorical representations for verbs of different nature that correspond to cryptotypes or mini-cryptotypes. More importantly, with the forming of cryptotypes, overgeneralization and recovery become a

function of how well the verb fits in with the cryptotype and its prefixation pattern. Specifically, if the new item takes the same prefix as members of its cryptotype, the network generalizes correctly. Over-generalization occurs when the novel input is inconsistent with members of the cryptotype with respect to prefixation. Recovery is hard when the novel verb takes a different prefix as its synonym in the same category and is likely to disturb the learned pattern.

One important aspect of over-generalization and recovery concerns the relation between the learning of new verbs and the learned patterns of old verbs. Our study reveals interesting results in this regard. First, Simulation 2 shows that new input is more susceptible to overgeneralization when the learning system has not formed any sensible semantic structures, while the same input may not be overgeneralized at a later stage when semantic structures have clearly emerged from learning. It suggests that depending on the structure of old information, the learning of new verbs may be different. Second, Simulation 3 shows that the learning of new verbs may disturb already learned patterns of old verbs, in particular, synonyms that undergo different prefixations (which results in mirror images between synonyms). The recurrent errors observed with synonyms reflect the attraction and competition between old and new items in the mapping from similar input to dissimilar output. The fact that learning errors can reoccur with verbs that have been learned at an earlier stage, together with the fact that error rates for the same verb can go up and down at different stages of learning, suggest to us a highly dynamic and interactive system in neural network learning. It may be argued, however, that the disturbance of new verbs to old patterns is attributable to the effect of "catastrophic interference" (McCloskey & Cohen, 1989; Ratcliff, 1990), a phenomenon due to the linear sigmoid function of the back-propagation network. But we think that the result captures the dynamics of human language learning in significant ways, whether or not it may be ascribed to the inherent functions of the network. In fact, the results of these simulations suggest a number of explicit new directions for research on overgeneralization and recovery in human children, with special attention to the short- and long-term fate of synonyms.

Although our current study is not a direct comparison between network performance and child performance in the acquisition of the English negative prefixes, we have seen some clear similarities between the two. Bowerman (1982) has pointed out that one important aspect in children's acquisition of the *un-* prefixation is that children recognize shared aspects of meaning among the verbs that can take *un-* and they eventually recover from errors on the basis of this semantic analysis, i.e., they tend to narrow down the range of errors by identifying the cryptotypes. Our study shows clearly how the network makes its effort in discovering relevant cryptotypes distributed among the meaning patterns of verbs (see Figures 5 to 7), and how learning differs as a function of whether the network has formed clear semantic structures (see Figures 12 and 13). Bowerman's data also indicate that overgeneralizations that fall within Whorf's cryptotype, such as *unsqueeze*, are much harder to correct than overgeneralizations that fall outside it, such as *unhate*. Although the exact patterns are not the same, our results similarly display different speed of recovery with different verbs. It was seen in both Simulations 2 and 3 that the items that fall into the cryptotype but have a different prefixation pattern as members of the cryptotype will be most difficult to get away from over-generalizations.

Unfortunately, because there are so few empirical studies of the acquisition of negative prefixes, it is not possible for us to make detailed comparisons between network's and children's patterns of acquisition. The learning patterns that are observed with various simulations in this study (e.g., differential learning speed, disturbance of new verbs to old patterns), we believe, are suggestive of a more detailed picture of overgeneralization and recovery, and are generalizable to human language learning. In this regard, we hope that our results will inspire future empirical research.

Although we have not explicitly addressed the "no negative evidence" problem in this study, we have indicated that our network is able to solve the prefixation problem and recover from all over-generalization errors with a simple learning mechanism such as back-propagation, without resort to innate, language-specific constraints. Our results may be ruled irrelevant to the negative evidence problem, since one could argue that after all a back-propagation network does receive some sort of negative feedback (i.e., the error term in proportion to the

discrepancy between the actual output and the desired output) during the process of learning. However, it is not clear whether the kind of error adjustment in a back-propagation network is equivalent to explicit negative evidence to the child. There has been much debate in the child language literature about the nature of negative evidence and about whether children receive other indirect forms of negative feedback even if there is no explicit negative evidence (e.g., Bohannon & Stanowicz, 1988; Gordon, 1990; Bohannon, MacWhinney, & Snow, 1990; Li, 1991). In our view, the kind of negative evidence provided to our network should be regarded as indirect and is in a form which is very similar to what children might also receive. Back-propagation is a fine-tuning process of weights and connections adjustment rather than an explicit error correction. Learning with this kind of algorithm is a process of *constraint satisfaction*, a process analogous to the child's efforts in matching his language to the adult language spoken around him.

In conclusion, we would like to point out that neural networks are very useful and important tools for tapping into language acquisition processes, not only in the domains of form-form mappings, but also in form-meaning mappings. There have already been several studies that demonstrate the usefulness of neural networks in language acquisition since Rumelhart & McClelland (1986) on the acquisition of the English past tense (MacWhinney, 1990, MacWhinney, Leinbach, Taraban, & McDonald, 1989; MacWhinney & Leinbach, 1991; Plunkett & Marchman, 1991). The current study examines a new domain and shows that a single learning mechanism of the sort embodied in PDP networks can display overgeneralization and recovery in principled ways that are similar to child language acquisition. However, although the various patterns found with our simulations may provide insights for empirical research, they are exploratory and preliminary. In future research, we need to address more systematically how overgeneralization and recovery differ as a function of the degree of semantic similarity (i.e., different levels of synonymy), the time of learning (i.e., the time at which new items are introduced and generalized), and the relationship between the stability of the learned pattern and its vulnerability to disturbance (i.e., more stable structures may be less likely to be disturbed). Only with investigation into these questions (together with future empirical studies) will we be able to provide a more coherent and comprehensive picture of overgeneralization and recovery in language acquisition.

ACKNOWLEDGMENTS

This research was supported by a fellowship from the Human Frontier Science Program Organization. Partial support for the preparation of the paper was also provided by the McDonnell-Pew Center for Cognitive Neuroscience, San Diego. The study was initially inspired by a discussion with Melissa Bowerman and Jeffrey Elman on the "no negative evidence" problem. I am very grateful to Elizabeth Bates, Melissa Bowerman, Jeffrey Elman and members of the UCSD PDPNLP group for their insightful discussions during various phases of this study, and to Catherine Harris for her initial contribution to the construction of semantic representations. I would also like to thank Edward Yang for his assistance in running the experiments with human subjects. Address for correspondence: Ping Li, Center for Research in Language, University of California, San Diego, La Jolla, CA 92093-0526. E-mail: liping@crl.ucsd.edu. Please send correspondence to the following address after August 15, 1992: Ping Li, Department of Psychology, the Chinese University of Hong Kong, Shatin, Hong Kong.

REFERENCES

Bates, E., & MacWhinney, B. (1987). Competition, variation, and language learning. In B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Erlbaum.

Bohannon, J., MacWhinney, B., & Snow, C. (1990). No negative evidence revisited: Beyond learnability or who has to prove what to whom. *Developmental Psychology,* **26**, 221-226.

Bohannon, J., & Stanowicz, L. (1988). The issue of negative evidence: Adult responses to children's language errors. *Developmental Psychology,* **24**, 684-689.

Bowerman, M. (1981). The child's expression of meaning: expanding relationships

among lexicon, syntax, and morphology. In H. Winitz (Ed.), *Native language and foreign language acquisition. Annals of the New York Academy of Sciences*, **379**, 172-189.

Bowerman, M. (1982). Reorganizational processes in lexical and syntactic development. In E. Wanner & L. Gleitman (Eds.), *Language acquisition: The state of the art*. Cambridge University Press.

Bowerman, M. (1987). Commentary: mechanisms of language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Erlbaum.

Bowerman, M. (1988). The 'no negative evidence' problem: How do children avoid constructing an overly general grammar? In J. Hawkins (Ed.), *Explaining language universals*. New York: Basil Blackwell, Inc.

Braine, M. (1971). On two types of models of the internalization of grammars. In D. Slobin (Ed.), *The ontogenesis of grammar*. New York: Academic Press.

Bybee, J., & Slobin, D. (1982). Rules and schemes in the development and use of the English past tense. *Language*, **58**, 265-289.

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht: Foris.

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

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

Elman, J. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning,* **7,** 195-225.

Gordon, P. (1990). Learnability and feedback. *Developmental Psychology,* **26**, 215-218.

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

Kuczaj, S. (1977). The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, **16**, 589-600.

Lachter, J., & Bever, T. (1988). The relation between linguistic structure and associative theories of language learning -- A constructive critique of some connectionist learning models. In S. Pinker & J. Mehler (Eds.), *Connections and symbols*. The MIT Press.

Li, P. (1991). Overgeneralization, recovery, and the 'no negative evidence' problem. Paper presented at the *Workshop on Connectionism and Developmental Theory,* Center for Research in Language, University of California, San Diego, July.

MacWhinney, B. (Ed.) (1987). *Mechanisms of language acquisition*. Hillsdale, NJ: Erlbaum.

MacWhinney, B. (1990). Connectionism as a framework for language acquisition theory. Manuscript. Carnegie Mellon University.

MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition*, **40**, 121-157.

MacWhinney, B., Leinbach, J., Taraban, R., & McDonald, J. (1989). Language learning: Cues or rules? *Journal of Memory and Language*, **28**, 255-277.

Marchand, H. (1969). *The categories and types of present-day English word-formation: a synchronic-diachronic approach.* Munich: C.H. Beck'sche Verlagsbuchhandlung.

Mazurkewich, I. & White, L. (1984). The acquisition of the dative alternation: Unlearning overgeneralizations. *Cognition*, **16**, 261-283.

McCloskey, M., & Cohen, N. (1989). Catastrophic interference in connectionist networks: the sequential learning problem. In Bower, G. (Ed.), *The psychology of learning and motivation*. New York: Academic Press.

Pinker, S. & Prince, A. (1988). On language and connectionism. In S. Pinker & J. Mehler (Eds.), *Connections and symbols*. The MIT Press.

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

Pinker, S. (1991). Rules of language. *Science*, **253,** 530-535.

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in

a multi-layered perceptron: Implications for child language acquisition. *Cognition*, **38,** 43-102.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, **97**, 285-308.

Rumelhart, D. & McClelland, J. (1986). On learning the past tenses of English verbs. In J. McClelland, D. Rumelhart, and the PDP research group (Eds.), *Parallel distributed processing*. Vol. 2. The MIT Press.

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning internal representations by error propagation. In J. McClelland, D. Rumelhart, and the PDP research group (Eds.), *Parallel distributed processing*. Vol. 1. The MIT Press.

Whorf, B. (1956). *Language, thought, and reality*. The MIT Press.