Cryptotypes, Meaning-Form Mappings, and Overgeneralizations*

Ping Li
Center for Research in Language
University of California, San Diego

1. Introduction

OVERGENERALIZATION, or a learner's application of a general pattern to inappropriate cases, has received increasing attention in language acquisition studies in recent years. Investigators have examined this phenomenon from a number of linguistic, psychological, and cognitive perspectives (Bowerman, 1982, 1987; Clark, 1987; Pinker, 1989) and have found many interesting results. One very important finding concerns the child's acquisition of the English past tense. It has been observed 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). Moreover, there is a clear U-shaped learning pattern associated with the acquisition process: initially children produce irregular forms correctly, but later on they overgeneralize regular forms to irregular verbs, and only gradually do they recover from all overgeneralization errors.

The acquisition of the English past tense has become a issue under hot debate in recent years. One of the areas in which researchers have had intensive discussion is the use of NEURAL NETWORKS (or PDP networks, connectionist networks) in simulating human performance (see Pinker & Mehler, 1988). A pioneering work using neural network modeling to study the acquisition of the English past tense is Rumelhart & McClelland (1986). Rumelhart & McClelland's modeling results indicate that neural networks

* This research was supported by a fellowship from the Human Frontier Science Program Organization. It 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 for their insightful discussions, and to Catherine Harris for her constructive ideas in building semantic representations. 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.

are able to capture overgeneralization errors displayed in child language, and in particular, the characteristic U-shaped development inherent in children's acquisition of past tense forms. 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. So far, the issue remains highly controversial (Elman, 1990; MacWhinney & Leinbach, 1991; Pinker & Prince, 1988; 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. 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. This study is an attempt to use a 'semantic neural network' (i.e., a network with semantic features) to tap into the process of form-meaning mapping in language acquisition.

Relevant to the overgeneralization problem is the child's ability to recover from overgeneralization errors. Although previous research has uncovered some mechanisms underlying the overgeneralization process, it has not provided us with a clear picture of this equally important, related process in language acquisition. The current study will address the issue of how recovery may be governed by a single learning mechanism such as BACK-PROPAGATION. Using the PDP conceptual framework as the basis for learning, I hope to provide insights into functional constraints on the processes of both overgeneralization and recovery.

Much of the current debate centers around the English past tense. In this study, I have chosen to study a new problem domain by looking at the acquisition of 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 following 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.

2. Cryptotypes and the acquisition of negative prefixes

English has a number of negative prefixes to mark reversativity, i.e., the reversal of the action specified by the base verb, such as *un-* in *unfasten* and *dis-* in *disentangle*. 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. (This prefix should not be confused with *un-* prefixed to adjectives to negate a quality or state, such as in *unkind* and *unbroken*). According to Whorf, the set of verbs that can be prefixed with *un-* seem to fall into a cryptotype that has a "covering, enclosing, and surface-attaching meaning". Cryptotypes 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 the meaning for the *un-* prefixable verbs is elusive: "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-*.

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 semantic clusters associated with the *un-* prefixable verbs. These semantic 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, which 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", including *unbind, unbolt, unbuckle, unclasp, unfasten, unlease, unlock, untangle, untie, unzip*, etc. It is important to note that each of the sub-classes is not a unique set, and the features associated with the class are not exclusive, due to the properties of cryptotypes. Furthermore, a verb may also have half of a feature, or .7 of a feature, as part of its inherent meaning, since a particular feature may be important but not characteristic of, or possible but not necessary in a verb's meaning. 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 because of the distributed nature of knowledge representation in neural networks.

Unlike in the case of the *un-* verbs, there seem to be no clear semantic clusters associated with the verbs that can be prefixed with *dis-*. However, *dis-* is also interesting because it is closely related to *un-* in marking reversativity. According to Marchand (1969), *dis-* and *un-* may compete as alternative devices for marking negativity or reversativity. 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-*. Similarly, we have *disentangle* vs. *untangle*, *discharge* vs. *unload.* Furthermore, some *dis-* verbs have counterpart *un-* verbs in their past participle forms, such as *disconnected* vs. *unconnected*, *disconfirmed* vs. *unconfirmed*, and *disarmed* vs. *unarmed*. 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 brought into attention by Bowerman (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 narrow down the range of errors by identifying shared aspects of meaning among the verbs that can take *un-* (corresponding to Whorf's cryptotype of the "covering, enclosing, and self-attaching meaning").

There has been no report on the acquisition of the prefix *dis-*. One reason for this lack of study 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, e.g., *discuss, disturb,* and *distort*. 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

---

[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.

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-*. 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. Moreover, the properties of neural networks (e.g., distributed representations, nonlinearity in information processing) are ideal for dealing with our problem, i.e., the elusiveness of cryptotypic semantic structures.

3. Method

3.1 *The input data*

Since the current study is concerned with the role of semantics in the acquisition of the meaning-form mapping, we used semantic representations as input to the network. 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 'zero' verbs which take neither *un-* nor *dis-*. The final selection of the 47 *un-* verbs and 18 *dis-* verbs was based on consultation with native speakers about their intuition on the acceptability of all the *un-* verbs and *dis-* verbs that appeared in the Webster's dictionary. 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. We chose twenty semantic features 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. These features are supposed to capture the semantic range of the verbs that can be prefixed with *un-* and *dis-*. 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.

The specific values of the semantic features for each verb came from a semantic judgment experiment. In the experiment, fifteen native speakers of English were given a list of 105 verbs and a list of semantic features (see Table 1), and were asked to rate how relevant each feature is to each verb on

a scale of 1 to 7. The averaged rating scores from the fifteen subjects 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). A HIERARCHICAL CLUSTER ANALYSIS conducted on the results indicates that the relative meaning distances between different verbs as revealed by the analysis are consistent with native intuitions about how similar the words are (i.e., synonymous words tend to group together as clusters), providing support for the validity of our experiment (see Li, 1992).

Semantic Features of Verbs	*connect*	*link* (dis-)	*turn* (un-)	(zero)
1. the action is a physical manipulation of an object		.7	.9	.6
2. entities form a collection or a group		.6	.5	.0
3. two entities have complex interrelations		.5	.6	.0
4. one entity tightly fits into another		.6	.7	.0
5. there is a physical distortion of the object		.1	.0	.3
6. the action has an effect on the entity		.2	.3	.5
7. there is a change from one location to another		.1	.1	.5
8. one entity touches another		.9	.9	.1
9. there is a qualitative change at the end of the action		.3	.4	.3
10. one entity is a salient part of another		.5	.7	.1
11. one entity is surrounded by another		.2	.3	.1
12. the action is a circular movement		.0	.0	.6
13. entities are placed in an orderly structure		.3	.5	.0
14. there is a change from one state to another	.3	.3	.1	
15. two entities can be separated		.6	.4	.1
16. two entities can be connected		.7	.8	.0
17. there is a container involved		.0	.1	.1
18. one entity hinders another in some way		.1	.2	.1
19. one entity obscures another		.0	.0	.0
20. the action is mainly a mental activity		.0	.1	.0

Table 1: Examples of verbs encoded as semantic feature vectors

3.2 *Network architecture and task*

A standard three-layer back-propagation network is used for all simulations in this study (Rumelhart, Hinton, & Williams, 1986). There are 20 input units encoding semantic features, 6 hidden units encoding internal semantic representations, and 3 output units representing either *un-*, *dis-*, or no prefixation (*zero*). All simulations were conducted using the TLEARN program configured at the Center for Research in Language, UCSD.

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 *zero*. That is, each time the network is given an input pattern, it is required to output an *un-*, *dis-*, or *zero* form of the verb.

The network's performance is assessed with RMS scores, i.e., the root mean squared differences between the actual values generated by the network and the desired values in the teaching signals. A hierarchical cluster

analysis is also used in the analysis (cf. Elman, 1990). This technique allows us to discover the relative meaning distances between different verbs that are represented at the hidden unit level, so that we may determine if the network has found meaningful structures in the meaning-form mappings during the course of learning.

4. Results and analysis

4.1 *Simulation 1: Discovering inherent semantic structures*

In this simulation, the network was trained on the complete data set (105 verbs). Figure 1 shows the global error decrease within a time frame of 80000 learning cycles, averaged over 1000 randomly sampled patterns at each point.

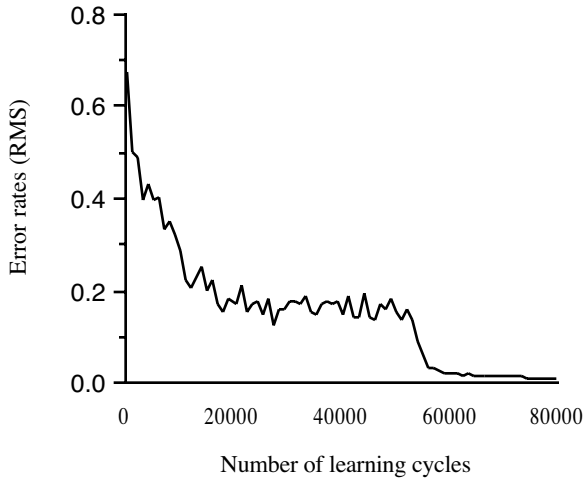

Fig. 1 Global errors in the network's learning of the complete data

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 2 plots a few examples which differ dramatically in how fast they were learned with respect to the prefixation patterns.

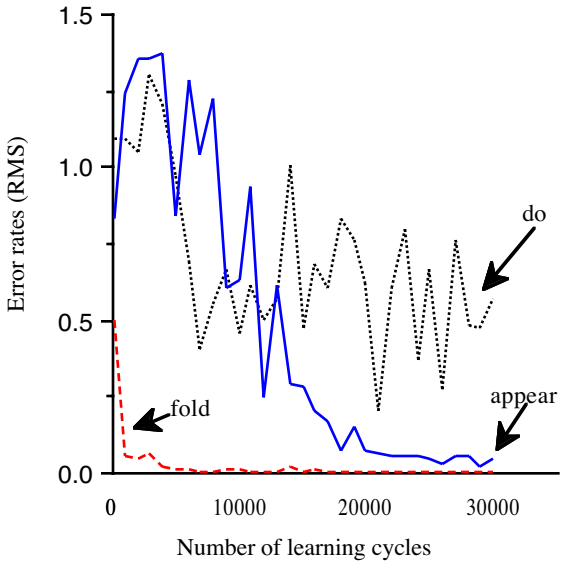

Fig. 2 Learning as a function of the verb's fit to cryptotypes and prefixation patterns

Interestingly, the learning speed with different verbs 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-*. 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. The verb *appear* is harder, because it belongs to a set of motion verbs (e.g., *come, run, go, walk*, etc.), which do not take any prefix, but it takes *dis-* itself. In contrast to *fold* and *appear*, the verb *do* does not fall neatly into any clear semantic clusters and thus experiences difficulty in learning (the network could not learn it within 30000 cycles).

In a hierarchical cluster analysis of the internal activation patterns of the network described above, we found some more interesting results on the network's performance in discovering cryptotypes during learning.[3] There

[3] For reasons of space the graphs of the cluster analysis are not presented here. See Li (1992) for details.

is clear evidence that the network has formed a lot of structure at the later stages of learning, with respect to both the prefixation patterns and the semantic categories. It is also clear from the analysis that the 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. This is shown in that at initial stages when the network has not formed any sensible structure, verbs belonging to the same cryptotype are scattered all over the branches of the cluster tree without being grouped at all; at intermediate stages, when the network has developed some structures, the verbs tend to come together as groups, forming small clusters at several levels; and only later do meaning clusters clearly emerge in groups. Results from the intermediate stages are especially helpful in that they present us with snapshots of the network's efforts in discovering semantic categories.

To summarize, our network is able to solve the prefixation problem in a simple task in which the learning mechanism has very little built-in structure, and the speed with which a particular verb is learned is a function of how well the verb fits the cryptotype and its prefixation pattern. Moreover, our cluster analysis indicates the dynamic process in which the network discovers cryptotypes or covert semantic categories inherent in the form-meaning mappings.

4.2 *Simulation 2: Overgeneralization and recovery*

In this simulation, the network was first 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 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 either (a) generalizes correctly on the new items on the basis of the learned pattern, because the new items fit in with the learned pattern; (b) overgeneralizes on the new items, because the new items do not fit in with the learned pattern but the child thinks they do. 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 3 presents the results of three different verbs when the network was trained on 85 of the 105 verbs and tested on the remaining 20 verbs. It shows the RMS scores in the testing phase (after 20000 cycles, at which point all verbs in the training phase have been learned). It can be seen that similarly to what was found in Simulation 1, *appear* displayed high initial learning errors. However, in this simulation, *appear* has not been seen by the network until a later stage, when the network has 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

appear. 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 in with 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 fall into a cluster of rotating verbs that all take the prefix *un-* (except *turn*). In Simulation 1 we found that the network quickly discovered the category of rotating verbs, so that the learning errors of *fold* decreased rapidly. The same holds true for the verb *clasp*, which belongs to a large cluster of verbs which share a locking or binding meaning, all of which take the prefix *un-*.

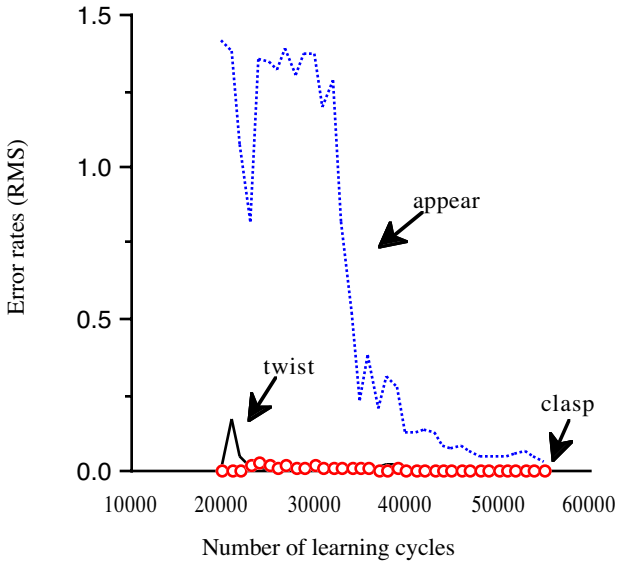

Fig. 3 Overgeneralization and recovery as a function of the verb's fit to cryptotypes and prefixation patterns (1)

The above results are also observed in other trainings using a different combination of training vs. test items. Figure 4 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). These results are completely consistent with results presented in Figure 3. *Buckle* is similar to *clasp* in Figure 3, 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 more discussions below). Looking into the network's internal representations we found that the difficulty with *entangle* is due to its close semantic similarity 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). 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.

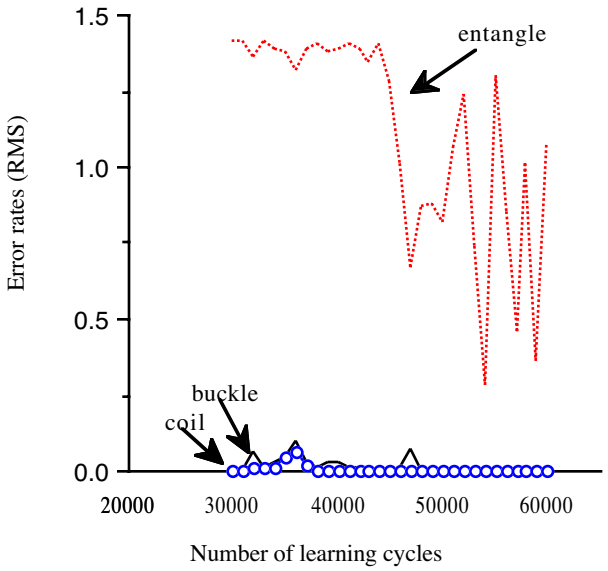

Fig. 4 Overgeneralization and recovery as a function of the verb's fit to cryptotypes and prefixation patterns (2)

To summarize, when novel input is consistent with the prefixation patterns of its semantic associates in a cryptotype, no overgeneralization occurs. Overgeneralization occurs when the new input is inconsistent with the prefixation of members of the cryptotype. 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.

4.3 *Simulation 3: Semantic similarity effects*

The last simulation is particularly concerned with how semantic similarity may affect the recovery process. We have already observed some semantic similarity effects in Simulation 2, in which the difficulty of recovery with *entangle* was attributed to its close similarity with the verb *tangle* in the training set. In this simulation, we systematically investigate pairs of synonymous verbs by training on one member of the pair and testing on another. These pairs include: *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 similarity, the same kind of results were found for all pairs, even though the exact form of recovery may be different for different pairs.

Figures 5a and 5b show the results of two different pairs of verbs 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.

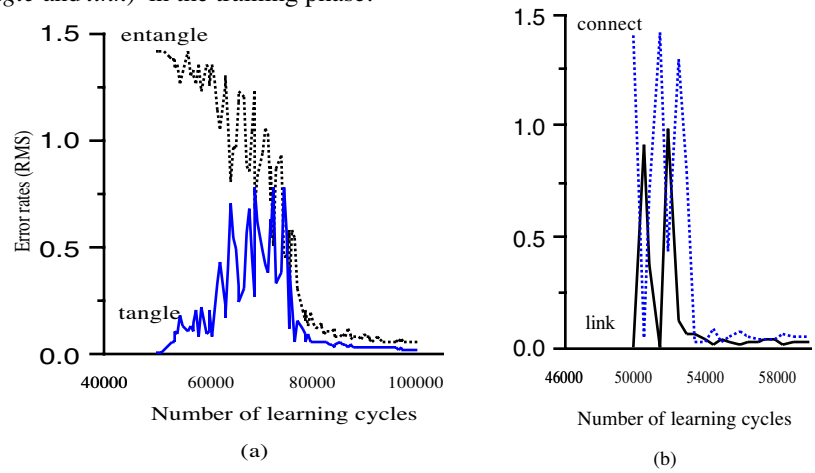

Fig. 5 Contrastive patterns in learning synonymous pairs

These graphs show clearly 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 5a, 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 5b, the same kind of contrastive pattern was found, 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. The contrastive pattern of performance 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, synonymous verbs *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 weights need to be adjusted. In this case, learning proceeds as a function of suppressing the strength of mapping similar words 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.

Although Figures 5a and 5b show very different forms of recovery, the fact that both show mirror images indicates that members of the synonymous pair 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), but competition for different mapping patterns between items in the same category. The different forms with the two pairs in these graphs are the results of two different ways of getting out the problem by competition: in Figure 5a, 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, whereas in Figure 5b, the system was eager to get out of trouble, resulting in radical swings of learning errors, but also consequently in rapid recovery.

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 (e.g., *link, tangle, load,* and *put*). Apparently, the recurrent errors with the old learned verbs (i.e., the synonyms) in the new test phase are due to the disturbance or interference from the network's learning of novel verbs. This disturbance may take different forms, as shown in Figures 5a and 5b. 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 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 synonyms at all.

In this connection, we also wish to note that the recurrent errors constitute part of an interesting U-shaped pattern (see errors with *tangle* in

Figures 5a, less obviously with *link* in Figures 5b), 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, we have found that the learning of new verbs may disturb already learned patterns of old verbs, in particular, synonymous pairs that undergo different prefixation patterns. The mirror images reflect effects of attraction and competition between the old and the new items. Overgeneralization in this case is the result of semantic similarity and recovery is the result of competition.

5. General discussion

The goal of the study is to uncover the psycholinguistic mechanisms underlying overgeneralization and recovery in language acquisition. 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 have shown that neural networks are useful tools for investigating child language phenomena. Contrary to proposals of symbolic advocates, neural networks are not only sensitive to aspects of acquisition in the form-form mappings, they are also sensitive to the complex interactions between meanings and forms.

In the connectionist perspective, Whorf's cryptotype problem can be solved relatively easily by the use of distributed representation. 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 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 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 becomes a function of how well the verb fits in with the cryptotype and its prefixation pattern.

One important aspect of overgeneralization 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. Results from Simulation 3 shows that the learning of new verbs may disturb already learned patterns of old verbs. 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 a highly dynamic and interactive system in neural network learning.

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 how the network makes its effort in discovering relevant cryptotypes distributed among the meaning patterns of verbs and how learning proceeds on the basis of this. 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. However, 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.

References

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

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.

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

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.

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.

Li, P. 1992. Overgeneralization and recovery: Learning the negative prefixes of English verbs. Technical Report 9203, Center for Research in Language, University of California, San Diego.

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

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

Pinker, S., & Mehler, J. (Eds). 1988. *Connections and symbols.* The MIT 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.

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.