From Rote Learning to System Building

Kim Plunkett
Virginia Marchman

CRL Technical Report 9020
September 1990

Center for Research in Language

University of California, San Diego
La Jolla, CA 92093-0126

From Rote Learning to System Building: Acquiring Verb Morphology in Children and Connectionist Nets

Kim Plunkett*
Institute of Psychology
University of Aarhus, Denmark

Virginia Marchman
Center for Research in Language
University of California, San Diego

Abstract

The traditional account of the acquisition of English verb morphology supposes that a dual mechanism architecture underlies the transition from early rote learning processes (in which past tense forms of verbs are correctly produced) to the systematic treatment of verbs (in which irregular verbs are prone to error). A connectionist account supposes that this transition can occur in a single mechanism (in the form of a neural network) driven by gradual quantitative changes in the size of the training set to which the network is exposed. In this paper, a series of simulations is reported in which a multi-layered perceptron learns to map verb stems to past tense forms analogous to the mappings found in the English past tense system. By expanding the training set in a gradual, incremental fashion and evaluating network performance on both trained and novel verbs at successive points in learning, we demonstrate that the network undergoes reorganizations that result in a shift from a mode of rote learning to a systematic treatment of verbs. Furthermore, we show that this reorganizational transition is contingent upon a critical mass in the training set and is sensitive to the phonological sub-regularities characterizing the irregular verbs. The optimal levels of performance achieved in this series of simulations compared to previous work derives from the incremental training procedures exploited in the current simulations. The pattern of errors observed are compared to those of children acquiring the English past tense, as well as children's performance on experimental studies with nonsense verbs. Incremental learning procedures are discussed in light of theories of cognitive development. It is concluded that a connectionist approach offers a viable alternative account of the acquisition of English verb morphology, given the current state of empirical evidence relating to processes of acquisition in young children.

1 Introduction

Several accounts of the development of English inflectional morphology are couched in terms of a three-phase U-shaped pattern of acquisition. These accounts derive primarily from analyses of naturalistic production data which indicate that early in development, children produce the correct forms of English past tense or plural irregular forms, such as *went* or *sheep*. Later, these forms are incorrectly inflected and errors occur, e.g. *goed* or *sheeps*. Finally, the tendency to make such errors decreases, as some forms are identified as exceptions to the predominant pattern in the inflectional system.

The production of these overgeneralization errors has been interpreted to indicate that learning a language primarily involves the acquisition of *rule systems*, i.e., explicitly representable generalizations about linguistic regularities which allow the productive generation of forms that are not (or have not yet been) encountered in the input. The three stages in U-shaped development have each been interpreted as manifesting the application of different *mechanisms* or *strategies* for forming past tense or plural forms, each representing different modes or periods within the course of rule acquisition. During the first stage, a *rote* learning mechanism stores all forms, both regular and irregular, as independent

*A substantial proportion of this work was carried out while the first author was a visitor at the *Center for Research in Language, University of California, San Diego* supported by a grant from the MacArthur Foundation. We would like to thank the PDPNLP discussion group for comments on this work, and especially Jeff Elman for making computing and office facilities available. We would also like to thank Steen Ladegaard Knudsen for assistance in running the simulations.

items in a mental lexicon. During this period, both irregular and regular forms are correctly produced. However, systematic patterns which might characterize the input are not generalized to novel forms encountered by the child. The second stage of acquisition reflects the child's identification of patterns of regularity, represented in such terms as "add /-ed/" to form the past tense of a verb or "add /-s/" to form the plural of a noun. Overly general application of such rules results in the production of incorrect forms like *goed* and *sheeps*, as well as the tendency to regularize nonsense forms such as *nibbed*. Finally, after prolonged exposure to their native language, children pass into a third stage which involves discovering the exceptions to the rules, isolating these forms as independent entries in a mental lexicon. This final phase thus supposes the existence of two distinct mechanisms underlying children's ultimate knowledge of the English inflectional morphological system. One mechanism controls the default application of a general rule, responsible for the generativity of the regular paradigm in a given inflectional system. The second mechanism identifies exceptions and prompts the child to consult a separate knowledge store in the production and comprehension of irregular forms. This second mechanism is typically assumed to be closely related to, if not identical with, the mechanism underlying the rote learning which characterizes the first stage of acquisition.

In general, then, traditional accounts can be seen to attribute the onset of the production of erroneous forms to the transition from a stage in which production is determined primarily by rote-governed processes to a stage of rule construction and refinement, i.e., *system building*. The triggering of this transition in the child is not well understood (that is, what are the necessary and/or sufficient circumstances under which the transition occurs?), although a requisite amount of linguistic experience is typically assumed (Karmiloff-Smith [1986]). With respect to the English past tense, the transition into system building is at least partially dependent on sufficient exposure to suffixed past tense forms in order for the systematicities which define the regular rule to be extracted. Maturational factors which control the emergence of an inflectional system building device might also determine a U-shaped profile of development (e.g., Bever [1982]). However, maturational factors must be interpreted, at least to some degree, in interaction with input factors in order to account for observed time lags in the onset of productive behavior in different linguistic domains, e.g., the relatively early acquisition of the English plural system and the typically late acquisition of the past tense system (Brown [1973]; de Villiers & de Villiers [1985])[1]. Other explanations of the time lag between the acquisition of these two major inflectional systems in English incorporate children's developing conceptual understanding of time and number (e.g., Carey [1982]), in interaction with the character (e.g., transparency of form-function mappings) of the inflectional system of the language to be acquired (e.g., Slobin [1985]). Hence, while details about the timing and nature of the transition are still under debate, rule-based accounts generally assume that U-shape profiles of development are best explained via *dual mechanism architectures*, in which one mechanism is responsible for rote learning processes, and hence ultimately for the representation of the exceptions to the regular paradigm. A second mechanism is responsible for extracting the rule (or rules) which characterize the regular paradigm, and for constructing the underlying system guiding productive application and rule-governed language usage.

In an attempt to evaluate this account, we have argued elsewhere (Plunkett & Marchman [In press]) that it is important to distinguish between *macro* and *micro* patterns of the onset of errors when characterizing children's acquisition of the English past tense. Macro U-shaped development refers to a rapid and sudden transition into the second phase of system building, resulting in indiscriminate application of the "add /-ed/" rule to whole classes or categories of verbs. In contrast, a micro U-shaped developmental pattern is characterized by *selective* suffixation of English irregular verbs, and results in a period of development in which some irregular verbs are treated as though they belong to the regular paradigm while others are still produced correctly. The basis for selective application of the suffix may be defined with respect to certain representational characteristics of the verb stem (phonological, semantic or otherwise), or may result from the operation of a probabilistic device which determines the likelihood that the suffix will be applied to an irregular verb.

In reviewing sources of evidence regarding the patterns of overgeneralization errors in children, Plunkett & Marchman [In press] conclude that a macro characterization of past tense acquisition is inaccurate. For example, there appears to be little evidence that children overgeneralize the /-ed/ suffix

[1] Unless we assume that different devices for controlling the acquisition of the English past tense and plural systems emerge at different points in development.

indiscriminately, i.e., to all irregular verbs in their current vocabularies. Nor is there evidence to suggest the existence of a single well-defined stage of development in which erroneous behavior is observed (see also Derwing & Baker [1986]). Rather, children are likely to overgeneralize the suffix to only some irregular verbs (typically a small number) while, at the same time, they correctly produce the past tense forms of other irregular verbs. Furthermore, errors may occur across a protracted period with some irregular verbs recovering from erroneous treatment only to be overgeneralized again at a later point in development. In general, then, the evidence suggests that the transition from the first to the second phases is best characterized as a micro phenomenon, in which the onset of overgeneralization errors is both selective and gradual. An understanding of U-shaped past tense acquisition thus requires an account of:

1. The mechanisms that trigger the transition from rote learning to system building.
2. The basis for the selective overgeneralization errors produced by children (featural or probabilistic).
3. The mechanism(s) by which overgeneralizations are eventually eliminated and appropriate performance on both regular and irregular verbs is ultimately achieved.

Two other factors further complicate an adequate account of children's acquisition of the English past tense. First, children's overgeneralizations do not always appear to result from the overapplication of an "add /-ed/" rule. For example, children produce stem and past tense forms like *eat* → *ated*, *sit* → *sit* or *pick* → *pack* (Marchman [1988]). In many respects, these *irregularizations* can be seen to be analogues to the mappings between stem and past tense of English irregular verbs (Bybee [1988]; MacWhinney [1987]). Second, as Pinker & Prince [1988] have noted, the set of irregular verbs in English (approximately 150 altogether) is not an unstructured list. Rather, it consists of a number of subregularities between the phonological form of the irregular stem and the type of transformation which relates the stem to its past tense form. For example, *all* English irregular verbs which have identical stem and past tense forms possess a stem final dental consonant (e.g., *hit* → *hit*). Similarly, verbs which undergo vowel suppletion tend to group into clusters of stem final vowel-consonant sequences, which form "family resemblance" patterns (e.g., *sleep* → *slept*, *weep* → *wept*, *keep* → *kept*).

The documented patterns of subregularity among English irregular verbs can be seen to predict the frequency of both overregularizations and irregularizations in children (Bybee & Slobin [1982]) and adults (Bybee & Moder [1983]). For example, Bybee & Slobin [1982] note that at certain periods in development, children tend to resist regularization of verb stems that already possess a dental final consonant, i.e., use the identical form in the past tense. Similarly, Derwing & Baker [1986], Marchman [1984] and Marchman [1988] note that vowel change errors on regular verbs often reflect the patterns of vowel suppletion which characterize the various clusters of irregular verbs. These findings suggest that children abstract and utilize more than one category of systematicity from among the corpus of verb stem/past tense pairs to which they are exposed. Given that children's early vocabularies contain a relatively large proportion of irregular verbs, it is not surprising that they may be misled into postulating alternative hypotheses of past tense formation based on the subregularities observed among the irregular verbs. A range of hypotheses for past tense formation may guide children's productions, and hence result in productive output which consists of both regularization and irregularization errors.

However, it is surprising that irregularization errors tend to persist across development (Bybee & Slobin [1982]; Kuczaj [1977]), even past the point in lexical acquisition when the number of regular verbs greatly outnumbers irregular verbs. Thus, the tendency to irregularize may be just as *robust* as the tendency to regularize, even though suffixed forms may be more prevalent in children's productions when evaluated in terms of the number of regular verbs that are correctly inflected or the number of irregular verbs that are regularized. Note also that irregularization processes do not reflect the indiscriminate application of rule. Not all verbs that end in a dental are subject to a "no change" irregularization nor are all verbs that possess particular vowel/consonant sequences subject to vowel suppletion. Both processes of regularization and irregularization generate micro, as opposed to macro, profiles of U-shaped development. While these facts do not contradict the dual mechanism hypothesis, the timing and nature of the interaction between the two mechanisms requires further explication in order to account for these patterns of errors and recovery.

Recently, Rumelhart & McClelland [1986] have argued that a *single mechanism system* in the form of an *artificial neural network* is capable of extracting a range of regularities that characterize the English past tense system and producing patterns of regularization and irregularization analogous to the errors observed in children. Furthermore, this work has been interpreted as the first in a series of challenges to the widely accepted view that linguistic behavior should be understood in terms of *explicitly representable* rules and principles, and a separate store of exceptions to those rules. As a result, Rumelhart & McClelland [1986] has been subject to thorough analysis and criticism aimed both at the details of their model and at the applicability of artificial neural networks for models of language acquisition and processing. For our purposes, the crucial feature of Rumelhart & McClelland's [1986] work is the claim that a single mechanism is responsible for U-shaped acquisition and the representation of both the regularities and irregularities underlying the English past tense system. Here, the transition from rote learning to system building does not rely on a dual mechanism architecture to capture the distinction between regular and exceptional patterns. Rather, the model exploits the capacity of connectionist networks to simultaneously:

1. *Memorize* individual patterns and their transformations when the number of pattern types is sufficiently small.
2. *Generalize* on the basis of regularities observed in the input when the number of patterns is sufficiently large.

Rumelhart & McClelland [1986] initially trained their network on a subset of the vocabulary to which it would eventually be exposed. During the first 10 epochs of training only 10 verbs (8 of which were irregular) were presented to the network. Given the learning and representational characteristics of their network architecture (a single-layered perceptron), the model succeeded in learning the 10 verbs by "rote", i.e., without discovering any regularities among the individual verbs in the training set which were then generalized to new verbs or which interfered with the successful mapping of other verbs in the training set. After 10 epochs of training, Rumelhart & McClelland [1986] increased the size of the training set by 410 verbs. Most of new verbs were regulars. This sudden expansion in vocabulary size caused the learning algorithm (a probabilistic version of the Perceptron Convergence Procedure (Rosenblatt [1962])) to extract the "add /-ed/" regularity and to reorganize the mapping characteristics to reflect the dominant suffixation process. As a result, many irregular verbs displayed a sudden decrement in performance. Continued training on the input set allowed the network to gradually recover from the erroneous mappings. Analyses of the network's performance reveals that much of the success in modeling the classical U-shaped profile of development derives from the exploitation of the transition from item memorization to generalization that is inherent in these networks when the *number* and *structure* of mapping patterns is manipulated.

Pinker & Prince [1988] have pointed out that the discontinuities introduced into the training regime by Rumelhart & McClelland [1986] do not reflect plausible discontinuities in the input to children learning the past tense. First, they argue, there is scant evidence for such an abrupt increase in the total number of verbs to which children are exposed. Second, the evidence from children's productions (Brown [1973]) suggests that the actual relative proportions of regular and irregular verbs are less skewed than those represented in the Rumelhart & McClelland [1986] training set. For example, Pinker & Prince [1988] suggest that during early phases of acquisition, regular and irregular verbs are approximately evenly represented in children's production vocabularies. In general, current consensus has targeted the implausibility of the discontinuities in the original simulations, and hence the theoretical significance of the U-shaped learning demonstrated by the Rumelhart & McClelland [1986] model has been undermined.

However, Plunkett & Marchman [In press] and Plunkett & Marchman [1989] have demonstrated that several characteristics of micro U-shaped development can emerge in an artificial neural network which maps verb stems to past tense forms *in the absence of any discontinuities in the training regime*. Using a constant vocabulary size and structure throughout learning, several series of simulations were used to illustrate the conditions under which decrements in performance and subsequent recovery on individual verbs are observed. In all cases, the patterns of errors observed were attributed to the *competition* between the different types of mappings used in the simulations (arbitrary, suffixation, identity, vowel change), which are typical of the relationship between verb stem/past tense pairs in English. This work also showed that the capacity of these types of networks to learn inflectional verb morphology is highly sensitive to input parameters such as the *type* and *token frequency* of stems in the input set. For example, arbitrary mappings (*go* $\rightarrow$ *went*) tend to be mastered when they are few in number (low type frequency) and when each unique stem is repeated frequently (high token frequency).

Interestingly, the frequency parameters that support realistic verb learning in artificial neural networks tend to converge on frequency typologies associated with natural languages. Plunkett & Marchman [In press] also showed that the types of errors observed in artificial neural networks are constrained by the phonological characteristics of the distinct verb mapping classes. Thus, some regular verbs which end in a dental are identity mapped; while, at the same time, regular verbs with stem final vowel-consonant pairs that are characteristic of the vowel change class are subject to irregularization.

In summary, Plunkett & Marchman [In press] observe a range of errors in their simulations which can be documented in the child language literature and which comprise the selective patterns of micro U-shaped development observed in children acquiring the English past tense. These results are achieved, not by introducing discontinuities into the training set, but by manipulating type and token frequency parameters in ways which reflect the characteristics of verbs in English. The errors characteristic of micro U-shaped development are thus shown to be a natural outcome of learning in a network required to map competing classes of verbs which vary in systematic, but not absolute ways with respect to mapping type, type and token frequency, and phonological predictability.

In attempting to demonstrate the ability of artificial neural networks to solve the mapping of competing verb classes in the absence of discontinuous input, Plunkett & Marchman [In press] deliberately held vocabulary size constant throughout training, i.e., at 500 verbs. Although the type and token frequency parameters used in these simulations were characteristic of English, it is unlikely that children attempt to learn an entire lexicon all of a piece at any point in learning. Naturalistic production and comprehension measures suggest that verb acquisition in children is an incremental (albeit non-linear) learning process. Because the *size* of the vocabulary used by Plunkett & Marchman [In press] precluded the network from achieving complete mastery of the vocabulary early in training, the marked transition from an initial overall high performance to a performance decrement that was achieved in the original Rumelhart & McClelland [1986] model was not observed in these simulations. Indeed, one of the primary goals of the Plunkett & Marchman [In press] work was to demonstrate that an abrupt *transition* from conditions of rote learning to conditions of system building is *not necessary* for the emergence of systematic regularization and irregularization errors and the subsequent recovery from those errors. Thus, it is important to distinguish these findings from the observation that children pass from a period in which the past tense of all verbs are produced correctly, to a period in which regularization and irregularization errors are observed.

As noted above, the determinants of the transition from rote learning to system building in children are not well understood. The traditional view supposes that a *dual-mechanism* system is required to support the transition. An alternative, suggested by the work of Rumelhart & McClelland [1986], is that the transition is affected by *quantitative* and *structural* changes in the vocabulary of verbs that a *single mechanism* is required to learn. A sufficient evaluation of the single mechanism approach is hindered by the unrealistically abrupt vocabulary discontinuity that was introduced into Rumelhart and McClelland's training set. While they clearly demonstrate that a single mechanism can perform both rote learning and generalization, as well as make a transition between the two, further evaluation of learning in a single mechanism architecture under realistic learning conditions has so far been obscured. For example, how many verbs does the network need to experience before it attempts to extract regularities from the input and organize this knowledge base in a more systematic fashion? Is this event sudden or gradual? To what extent does the systematization of the current vocabulary impact on performance with verbs not yet acquired by the network? Are there characteristic patterns of errors associated with the transition from rote learning to system building? Does an early rote state of the knowledge base affect later patterns of acquisition? How do the developmental profiles observed in a network making the transition from rote learning to system building correspond to profiles of development in children acquiring the English past tense? How does such a single mechanism approach compare with the traditional dual mechanism architecture in accounting for the data?

In the remainder of this paper, we explore the performance of an artificial neural network required to learn mappings analogous to the English past tense when vocabulary size is expanded gradually, in an incremental fashion, across the course of learning. Our goal in this work is to determine the conditions under which an artificial neural network makes a transition from a stage of rote learning to a period of system building and to evaluate the characteristics of its performance in this period of transition and beyond. By comparing the characteristics of performance in the network with data from

English children acquiring the past tense, we hope to demonstrate that artificial neural networks offer an alternative explanatory framework within which to understand the mechanisms underlying children's acquisition of inflectional verb morphology.

2 Method

2.1 Overview

All simulations involve training a multi-layered perceptron to map phonologically represented verb stems to their corresponding past tense forms. After initial training on a subset of 20 verb stems, the vocabulary is gradually expanded until it reaches a size of 500 verbs. Vocabulary expansion is performed following two types of training schedules, criterial expansion and incremental expansion. Several conditions of incremental learning are tested (see below). Learning in all simulations is evaluated at regular intervals during training in the following ways. First, the network's performance on verb stems in the current training set is evaluated in terms of the percentage of forms output correctly. For those forms that are produced incorrectly, errors are coded and categorized. Second, at every evaluation point in training, the network is tested on a set of novel verbs. In these cases, the mapping performed by the network on each novel stem is categorized in relation to the mapping characteristics of the training set.

A variety of control simulations have also been conducted to evaluate the role of various input and learning parameters on observed performance and learning effects. The details of these controls will be reported in the results section where appropriate. Each set of simulations that is reported here uses identical vocabularies and identical initial starting weights. However, the results of *all* sets of simulations have been replicated using different vocabularies and starting conditions.

2.2 Vocabulary

A vocabulary of 500 verb stems is constructed from a dictionary of approximately 1000 stems. Each verb in the dictionary consists of a Consonant-Vowel-Consonant (CVC) string, a CCV string or a VCC string. Each string is phonologically well-formed, even though it may not correspond to an actual English word. Each vowel and consonant is represented by a set of phonological contrasts, such as voiced/unvoiced, front/center/back[2]. Table 1 summarizes the phonological representations for all consonants and vowels used in the simulations. Verb

Table 1: Phonological representation

	PHONOLOGICAL FEATURE UNITS					
	CON/VOW	VOICING	MANNER		PLACE	
	#1	#2	#3	#4	#5	#6
/b/	0	1	1	1	1	1
/p/	0	0	1	1	1	1
/d/	0	1	1	1	1	0
/t/	0	0	1	1	1	0
/g/	0	1	1	1	0	0
/k/	0	0	1	1	0	0
/v/	0	1	1	0	1	1
/f/	0	0	1	0	1	1
/m/	0	1	0	0	1	1
/n/	0	1	0	0	1	0
/η/	0	1	0	0	0	0
/δ/	0	0	1	0	1	0
/θ/	0	1	1	0	1	0
/z/	0	1	1	0	0	1
/s/	0	0	1	0	0	1
/w/	0	1	0	1	1	1
/l/	0	1	0	1	1	0
/r/	0	1	0	1	0	1
/y/	0	1	0	1	0	0
/h/	0	0	0	1	0	0
/i/ (eat)	1	1	1	1	1	1
/I/ (bit)	1	1	0	0	1	1
/o/ (boat)	1	1	1	0	1	1
/ˆ/ (but)	1	1	0	1	1	1
/u/ (boot)	1	1	1	1	0	1
/U/ (book)	1	1	0	0	0	1
/e/ (bait)	1	1	1	1	1	0
/ε/ (bet)	1	1	0	0	1	0
/ai/ (bite)	1	1	1	0	0	0
/æ/ (bat)	1	1	0	1	0	0
/au/ (cow)	1	1	1	1	0	0
/O/ (or)	1	1	0	0	0	0

stems are assigned to one of four classes. Each class corresponds to a different type of transformation analogous to classes of past tense formation in English. The four classes of transformation are[3]:

Arbitrary mappings: There is no apparent relationship between the stem and its past tense form, e.g., '*go* $\rightarrow$ *went*'.

Identity Mapping: Past tense forms are identical to their corresponding verb stems. Such mappings are contingent upon the verb stem ending in a dental consonant i.e. /t/ or /d/, e.g., '*hit* $\rightarrow$ *hit*'.

Vowel Change: Certain vowels can be changed under the condition that they precede partic-

[2] See Plunkett & Marchman [In press] for a more thorough discussion of the phonological representation used here.

[3] A more fine-grained classification of the past tense of English is provided by Bybee & Slobin [1982], Pinker & Prince [1988]. However, the current four-way distinction serves to capture many of the phenomena of interest.

ular consonants. The following four vowel-consonant cluster changes are permitted:

1. /iz/ → /ez/ '*fiz* → *fez*'
2. /it/ → /ɛt/ '*kit* → *kɛt*'
3. /ais/ → /es/ '*lais* → *les*'
4. /ail/ → /Ol/ '*rail* → *rOl*'

Regular mappings: A suffix is appended to the verb stem. The form of the suffix depends upon the final vowel/consonant in the stem:

1. If the stem ends in a dental (/t/ or /d/), then the suffix is /-id/, e.g., '*pat* → *pat-id*'.
2. If the stem ends in a voiced consonant or vowel, then the suffix is voiced /d/, e.g., '*dam* → *dam-d*'.
3. If the stem ending is unvoiced, then the suffix is unvoiced /t/, e.g., '*pak* → *pat-t*'.

The suffixes on the regular past tense forms are represented non-phonologically as three distinct patterns across two output units, i.e., 0 1, 1 0, and 1 1. A fourth pattern (0 0) corresponds to the absence of a suffix, as is the case for stems in the irregular classes (i.e., arbitrary, identity and vowel change).

Stems are assigned randomly from the dictionary to each of the four classes, with the constraint that stems possess the appropriate characteristics of a given class. The resulting 500 verb vocabulary contains 2 stems in the arbitrary class, 458 stems in the regular class, 20 stems in the identity class and 20 stems in the vowel change class. Each of the four vowel-consonant clusters defining the vowel change class contains 5 members. Stem assignment to the arbitrary and regular classes are not contingent upon any particular criteria, and these classes may contain stems which have phonological characteristics of identity mapping or vowel change stems. The number of stems assigned to each verb class, i.e., the vocabulary configuration, is based on results from Plunkett & Marchman [1989] in which a wide range of verb frequencies were evaluated for their learning consequences, in light of their similarity to supposed vocabulary configurations of English. The current choice parallels a configuration which demonstrated a high level of learning in the earlier work.

Appropriate past tense forms are constructed for each vocabulary item in each of the four classes. In the case of stems in the arbitrary class, a past tense form is chosen that does not share any consonants or vowels with the stem, nor corresponds to the stem or past tense form of any other verb in the training set.

2.3 Training Schedule

After 500 verbs have been assigned to the four class types, a subset of 20 verbs is randomly selected from the vocabulary for use in the initial phase of training. The initial training set is comprised of 2 arbitrary stems, 10 regular stems, 4 identity stems and 4 vowel change stems. The token frequencies (i.e. the frequency with which any given stem is likely to be repeated during a single training epoch) for this initial phase of learning are 15 for the arbitrary stems, while regular, identity and vowel change stems have a token frequency of 5[4]. The relatively high token frequency of the arbitrary stems is based on results from Plunkett & Marchman [1989] in which arbitrary mappings are only acquired under similar token frequency conditions.

The network is trained on this small vocabulary until all verb stems are mapped to their appropriate past tense forms. The vocabulary is then gradually expanded in size. Two general types of expansion schedules have been tested. On the first schedule, *criterial expansion*, the vocabulary is expanded one verb stem at a time and trained until the new verb is successfully mapped by the network. At this point, a new verb is introduced, training continues until that verb is mapped correctly, another verb is added and trained until it is mapped correctly, and so on. On the second training schedule, *epoch expansion*, a new verb is introduced to the vocabulary and trained for a set number of epochs. Another verb is then introduced into the training set *irrespective of the level of performance on the mapping of the previously introduced new verb*. This process is repeated until the vocabulary has reached 500 verbs. In the current set of simulations, incremental schedules of 1, 2, 5, 10 and 25 epochs per new verb have been evaluated. In all epoch expansion simulations, training is reduced to 1 epoch per verb after 80 new verbs have been introduced (i.e., the total vocabulary has reached 100 verbs).

The order in which new verbs enter the vocabulary is determined by a weighted random selection process which is based on an 80% likelihood that the new verb is taken from the regular class and a 10% likelihood that the verb is taken from the identity or vowel change classes. Each new verb entered into the training set after the initial set of 20 are assigned a token frequency of 3, until the vocabulary size reaches a total of 100 verbs. Thereafter, verbs that are introduced (predominantly regulars)

[4] Pilot simulations indicated that the network would fail to learn all of the regular stems if a token frequency of less than 3 was used.

are trained using a token frequency of 1. This frequency profile is chosen to accommodate the data set to the observation that children are more likely to hear, and thus have a greater opportunity to learn, verbs with a high token frequency. A summary of the changing structure of the vocabulary is provided in Table 2.

Table 2: Vocabulary Structure by Verb Class

Total	Arbs	Regs	Ids	Vcs	Tok
20	2	10	4	4	5[a]
80	2	61	8	9	3
100	2	77	10	11	1
140	2	112	11	15	1
200	2	163	15	20	1
260	2	221	17	20	1
269	2	227	20	20	1
380	2	338	20	20	1
500	2	458	20	20	1

[a]Arbitrary verbs have a token frequency of 15.

2.4 Novel Verbs

A further 100 stems were selected from the dictionary for testing the generalization properties of the network. Of the 100 novel verb stems, 10 end in a dental final consonant (/t/ or /d/), 10 stems possess the characteristics of each of the 4 clusters defining the vowel change class (40 verbs), and 50 stems are legal stems but possess none of the previously mentioned characteristics. These sub-classes of novel stems permit evaluations of the manner in which the network has tuned its response characteristics to stems which are candidate (though not definitive) members of the various implicit stem classes making up the training set. The network's performance on the novel verbs is evaluated at regular intervals during training.

2.5 Network Configuration

All the simulations are run using the Rlearn simulator (Center for Research in Language, UCSD) using a back propagation learning algorithm. Back propagation involves the adjustment of weighted connections and unit biases when a discrepancy is detected between the actual output of the network and the desired output specified in a teacher signal. In multi-layered perceptrons (containing hidden units), error is assigned to non-output units in proportion to the weighted sum of the errors computed on the output layer.

All networks contain 18 input units, 30 hidden units and 20 output units. All layers in the network are fully connected in a strictly feed-forward fashion. Previous work (Plunkett & Marchman [In press]) demonstrates that a multi-layered perceptron is required to solve mapping problems of the type encountered here. There is no generally acknowledged criterion for selecting appropriate numbers of hidden units for an arbitrary problem. The modeler must, therefore, experiment with network capacities in order to find a configuration suited to the problem. The final choice of 30 hidden units reflects a compromise between the attempt to achieve an optimal level of performance and the aim to maximize the generalization properties of the network. Minimizing the number of hidden units in a network encourages the system to search for regularities in the input stimuli.

Training in the simulations follows a pattern update schedule, i.e., a pattern is presented to the net, a signal propagates through the net, the error is calculated, and the weights are adjusted. Pattern update is preferred to batch update (in which error signals are averaged over a range of input patterns before the weights are adjusted) for this problem since children are unlikely to monitor an average error on their output, but are more likely to monitor the error associated with individual pattern tokens. Learning rate and momentum are held constant throughout the simulation at values of 0.1 and 0.0, respectively. (As with the choice of network configuration, learning rate and momentum parameters are typically determined through experimentation rather than principled criteria). Verb stems are presented randomly to the network within each epoch of training.

2.6 Output Analysis

On each presentation of an input pattern, any error on the output units is recorded and the weights adjusted accordingly. The weight matrix for the network is saved at regular intervals; first, when the net has just mastered the initial 20 verbs and then each time a new verb is introduced but before any training on the new verb has occurred. These weight matrices provide snapshots to evaluate the accuracy of the network in producing the correct past tense form for each unique stem at different points in the network's development. The output of the network is evaluated in terms of the "closest fit" (in Euclidean space) to the set of phonemes that map the output space, defined by the teacher signal to the network (see Table 1). For each class of stems, error analyses provide an overall hit rate (i.e. % correct) and error types are classified by verb class to determine the proportion of stems that are incorrectly mapped as identity stems, vowel change stems, blends, etc. Similarly, novel verb stems are tested on the saved weight matrices. The different categories of novel verbs are analyzed separately to determine their output tendencies, i.e. whether they are regularized, irregularized or handled in some other fashion by the net.

3 Results

3.1 Criterial Training

Under the criterial expansion condition, vocabulary size is increased one verb at a time and training is continued on each new verb (as well as the initial set) until that new verb is successfully mapped by the network. Typically, training on the initial set of 20 verbs requires 15 to 40 epochs to reach criterion, depending on the initial configuration of random weights. Performance on subsequent training is also sensitive to the initial set of random weights. Several initial configurations were tested. However, in none of the criterial learning simulations was it possible to continue vocabulary expansion beyond approximately 27 verb stems. Training continued for a considerable number of epochs (up to 1000) or until it was clear that the error gradient had reached asymptote at a non-zero level.

The inability of networks in the criterial expansion condition to learn a large number of verb stem/past tense mappings reflects the propensity of networks of this type to be caught in "local minima". In order for networks to avoid entrenchment in specific areas of weight space, training must ensure that a variety of weight changes occur. If the network is repeatedly trained on a limited and fixed number of patterns, where a series of similar weight changes occur, further training may fail to promote necessary reorganizations or may even enhance the network's entrenchment in a particular region in weight space. This training schedule was therefore abandoned as a method of vocabulary expansion that is appropriate to the current task.

The following sections report on results from simulations in which verbs are added to the vocabulary *irrespective* of the level of performance of the network on the previously added verb.

3.2 Epoch Expansion

3.2.1 Overall Performance

Figure 1 summarizes the hit rates (%) on stems in the regular and irregular classes (arbitrary, identity and vowel change combined) as a function of vocabulary size in simulations which use the following expansions schedules: (a) 1 verb per epoch, (b) 1 verb every 5 epochs, and (c) 1 verb every 25 epochs[5].

First, note that a high level of performance is achieved on both regular and irregular verbs by the end of training (i.e., a total vocabulary size of 500 verbs) in all expansion schedules. This consistently high level of performance contrasts with that reported in earlier work (Plunkett & Marchman [In press]) in which performance on similar vocabulary configurations did not exceed 80% for the regular and vowel change classes. It is noteworthy that final level of performance appears to vary depending on the particular expansion schedule condition used. In particular, level of mastery on the set of regular verbs decreases as a function of the epoch increment, i.e., overall performance on regulars tends to diminish slightly as the network is trained more on each new verb before yet another new verb is introduced. For example, overall performance on the regular verbs in the 1 epoch condition reached 99%, whereas, only 95% of the regular stems were correctly mapped in the 25 epoch increment condition.

Interestingly, the relationship between expansion schedule and final level of performance on regular stems contrasts to the pattern of learning observed early in training. When vocabulary size is increased at a rapid rate (e.g., 1 epoch increment), performance on regulars does not improve as quickly and greater decrements are observed, compared to the other expansion schedules. Thus, while exposure to a constant vocabulary for a substantive period (i.e., 25 epochs) results in a more even level of performance across training, this appears to be at the expense of overall mastery of the set of regular mappings. It is possible that this pattern of results derives from the general inability of these networks to master a large number of mappings when training follows a criterion learning schedule (see section 3.1 above). When the training set is fixed for a number of training epochs, the network may have difficulty recovering from erroneous configurations of the weight matrix. Thus, the criterion learning condition may be seen as corresponding to an indefinitely long expansion schedule.

[5] Similar patterns of results are found for the 2 epoch and 10 epoch expansion schedules.

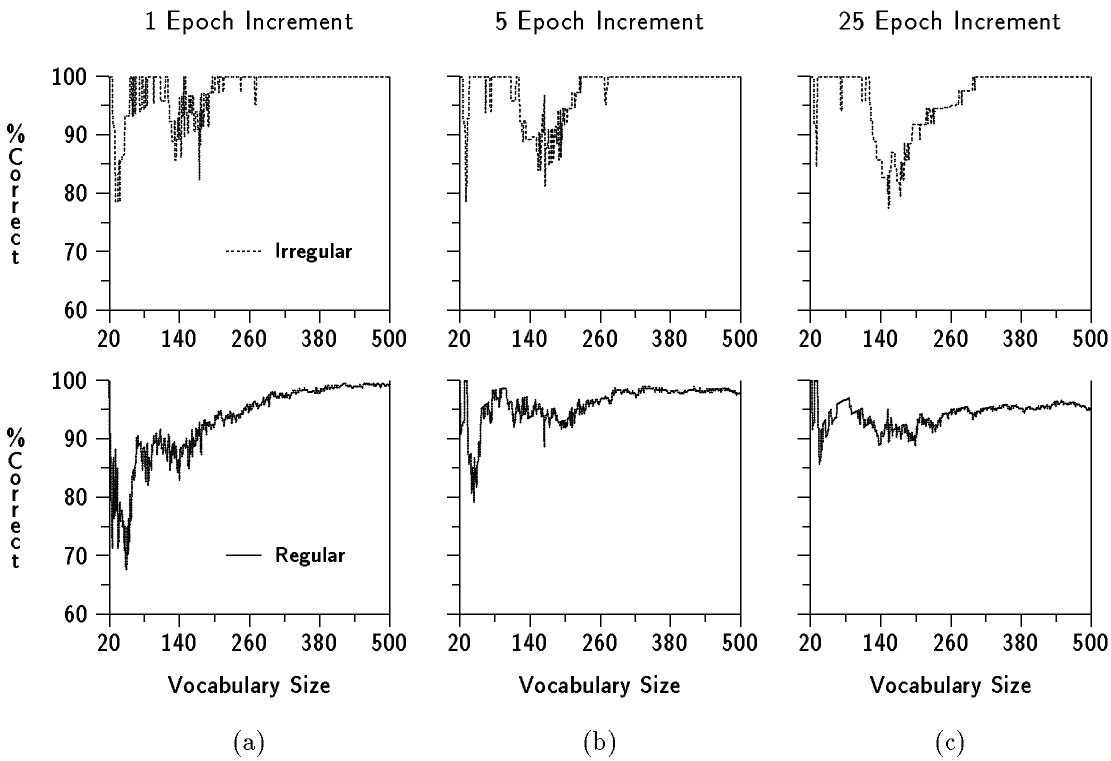

Figure 1: Hit Rates for Regular and Irregular Verbs

In order to rule out the effect of initial set size on final level of performance, a control simulation was performed in which the network was trained on an initial vocabulary of 1 verb (compared to 20), and the vocabulary was expanded incrementally every two epochs to a total of 500 stems. The results from this control simulation revealed a similar level of performance on the regular verbs, achieving 100% performance for all verb classes. While more complete analyses are required, these results suggest that the size of the initial subset of the vocabulary *per se* (e.g., 20 verbs) is not a necessary determinant of the high level of performance on the regular verbs observed in these simulations.

Second, Figure 1 also reveals that, in all conditions, performance on both regular and irregular verbs suffers a marked decrement in the period immediately following the initial period of training. This is the point in training when vocabulary expansion is first initiated, and hence, the network is now faced with the task of learning a continually increasing vocabulary. Note, however, that comparisons across the expansion conditions illustrated in Figure 1 reveal that expansion schedule influences the *degree* to which performance deteriorates during this period. In particular, the higher the rate of expansion, the greater the observed decrement in performance. Thus, the 1 epoch expansion schedule results in a deterioration to 68% for regular verbs. In contrast, performance on the regular verbs in the 25 epoch expansion schedule only deteriorates to 88%. The level of deterioration in performance on the 2 epoch and 10 epoch conditions (not shown in Figure 1) round out this pattern of results at 77% and 84%, respectively. Given the dynamics of learning in these networks, this result is not surprising. When the rate of expansion is decreased and the network is trained for a greater number of epochs on each new pattern, the opportunity to accommodate the weight matrix to the new pattern is correspondingly increased. A high rate of verb introduction allows only a minimal amount of training on each new verb and hence, performance on the overall set of verbs is decreased.

More surprising, however, is that recovery from this period of initial decrement is first manifest when vocabulary size reaches around 50 verbs *irrespective of the expansion schedule condition*[6]. The fact that

[6]It should be stressed that decrements in performance plotted in Figure 1 do *not necessarily* indicate the U-shaped "un-learning" of individual verbs. New verbs are continually introduced into the training set and may contribute to the overall decrement in performance even though old verbs continue to be mapped appropriately (c.f., the criterial expansion

recovery occurs at approximately the same level of vocabulary size regardless of the depth of the performance decrement, and hence, training schedule, suggests that absolute vocabulary size, i.e., *critical mass*, may be an important factor in determining the network's recovery from erroneous performance.

The simulations described in Figure 1 deliberately confound the size of the training set with the number of learning trials that each verb stem/past tense pair receives. In order to investigate a "critical mass" interpretation of these results, we conducted a series of control simulations in which expansion of the vocabulary is halted at various points. Training is then continued with a constant vocabulary size in order to observe whether performance recovers in the manner observed in the test simulations. Using a 2 epoch expansion schedule, performance was evaluated in networks with final vocabulary sizes of 30, 40, 50, 60, 70 and 80 verbs. For all vocabulary sizes, the network eventually recovers from its initial erroneous performance to achieve 100% correct performance on all verb classes. Indeed, similar patterns of trajectories of recovery were observed for networks learning vocabularies of these sizes. Thus, we conclude that the size of the training set does not appear to influence the ability of the network to recover from erroneous performance on verbs in the training set, i.e., on *trained* verbs. However, the results from these control simulations will illustrate an important contrast between trained vs. novel verbs. As presented below, the final size of the training set determines the generalization properties of the network, and hence can be seen to contribute to an explanation of the network's treatment of verbs introduced later in training.

Finally, Figure 1 reveals that for all training schedules, irregular verbs undergo a substantial decrement in performance during the middle period of training, i.e., when vocabulary size reaches approximately 100 verbs. During this period in training, the number of irregular verbs increases from 23 to 37 (including an extra 5 IDs and 9 VCs). Two factors contribute to this decrement in performance. First, recall that *all* verbs introduced after the 100 vocabulary mark are introduced with a token frequency of 1. We have shown in previous work (Plunkett & Marchman [In press]) that irregular mappings (in particular, arbitrary and vowel change verbs) are difficult for the network to master in the context of other types of mappings if those stems have a low token frequency. Second, as discussed in more detail in the following section, the generalization properties of the network during this stage of training are such that at least 30% of novel verbs are treated as regulars, regardless of whether they possess phonological properties characteristic of verbs in the irregular classes.

Note, however, that performance decrements on the irregular verbs are not identical in duration and magnitude across the expansion schedules. For example, in the 1 epoch increment schedule, the deterioration in performance is not as great nor as persistent compared to that observed in the 25 epoch increment condition. Further discussion of this finding and a detailed breakdown of the type of errors observed during this period of training is provided in section 3.2.2 (Error Analysis).

As mentioned above, the ability of these networks to map certain classes of verbs (especially arbitrary and vowel change verbs) is greatly facilitated by high levels of token frequency relative to the dominant mapping type (Plunkett & Marchman [In press]). It is possible, however, that the overall improved learning under the incremental expansion conditions obviates the need to construct training vocabularies in light of token frequency differences across verb classes. Therefore, a series of control simulations were conducted in which all verbs were introduced with the same token frequency (i.e., 1 token) and trained following a 2 epoch incremental expansion schedule. Briefly, the results indicate that incremental learning does not result in an improved level of performance on the arbitrary and vowel change verbs (measured by hit rate). Thus, the effects of token frequency outlined in earlier work appear to be robust across the range of training schedules tested here. These results, therefore, further justify the attempt to carefully configure training vocabularies in light of the general token frequency characteristics of English.

3.2.2 Error Analysis

All incorrectly generated mappings are analyzed in terms of the proportion of errors by verb class. We refer to these errors as *general learning errors* in subsequent discussion. (Patterns of micro U-shaped learning, i.e., errors on verbs which are successfully mastered followed by erroneous mapping and subsequent recovery are presented below.) Table 3 summarizes the categories and timing of errors, as well as the hit rate (%) for each verb class[7] for the 5 epoch expansion condition. Successive rows in the

schedule above). Analyses of the patterns of U-shaped learning in these simulations are discussed in section 3.2.2.

[7] An analysis of the arbitrary class is not presented here since they perform at optimal level throughout the expansion schedules.

Table 3: Errors by Class

		REGULARS				IDENTITY			VOWEL CHANGE				
VOC	EPO	HIT	SUF	ID	BLD	HIT	SUF	VC	HIT	SUF	ID	BLD	VC
20	40	100	0	0	0	100	0	0	100	0	0	0	0
30	90	90	0	0	0	100	0	0	71	0	0	0	0
40	140	80	20	0	0	100	0	0	100	0	0	0	0
50	190	85	20	20	20	100	0	0	100	0	0	0	0
60	240	95	50	50	0	100	0	0	100	0	0	0	0
70	290	94	0	33	0	100	0	0	100	0	0	0	0
80	340	98	0	100	0	100	0	0	100	0	0	0	0
90	390	97	0	100	0	100	0	0	100	0	0	0	0
100	440	97	0	50	0	100	0	0	100	0	0	0	0
110	490	94	20	20	0	90	100	0	100	0	0	0	0
120	540	94	40	20	0	100	0	0	100	0	0	0	0
130	590	97	33	66	0	100	0	0	85	0	50	50	0
140	640	93	57	28	0	100	0	0	80	0	66	33	0
150	690	95	40	20	0	100	0	0	81	0	33	0	33
160	740	93	37	37	0	91	0	100	88	0	50	0	0
170	790	94	28	42	0	84	0	50	88	50	0	0	50
180	840	95	14	57	0	92	100	0	78	0	25	25	50
190	890	92	9	54	0	92	100	0	78	0	25	25	50
200	940	92	15	53	0	93	100	0	95	0	0	100	0
210	990	94	10	60	0	93	100	0	95	0	0	100	0
220	1040	95	0	66	11	93	100	0	100	0	0	0	0
230	1090	95	11	55	0	100	0	0	100	0	0	0	0
240	1140	96	0	71	14	100	0	0	100	0	0	0	0
250	1190	96	0	62	12	100	0	0	100	0	0	0	0
260	1240	96	12	62	12	100	0	0	100	0	0	0	0
270	1290	97	0	83	0	95	100	0	100	0	0	0	0
280	1340	96	0	62	12	100	0	0	100	0	0	0	0
290	1390	98	25	75	0	100	0	0	100	0	0	0	0
300	1440	98	25	75	0	100	0	0	100	0	0	0	0
310	1490	98	20	60	0	100	0	0	100	0	0	0	0
320	1540	97	14	42	28	100	0	0	100	0	0	0	0
330	1590	98	20	60	0	100	0	0	100	0	0	0	0
340	1640	98	25	50	0	100	0	0	100	0	0	0	0
350	1690	98	25	75	0	100	0	0	100	0	0	0	0
360	1740	98	40	60	0	100	0	0	100	0	0	0	0
370	1790	98	40	40	0	100	0	0	100	0	0	0	0
380	1840	97	28	57	14	100	0	0	100	0	0	0	0
390	1890	98	16	66	16	100	0	0	100	0	0	0	0
400	1940	98	14	71	14	100	0	0	100	0	0	0	0
410	1990	98	16	66	16	100	0	0	100	0	0	0	0
420	2040	98	16	66	16	100	0	0	100	0	0	0	0
430	2090	98	20	40	20	100	0	0	100	0	0	0	0
440	2140	97	12	50	25	100	0	0	100	0	0	0	0
450	2190	98	28	28	42	100	0	0	100	0	0	0	0
460	2240	98	16	33	49	100	0	0	100	0	0	0	0
470	2290	98	16	33	49	100	0	0	100	0	0	0	0
480	2340	98	16	33	49	100	0	0	100	0	0	0	0
490	2390	97	22	33	33	100	0	0	100	0	0	0	0
500	2440	97	20	30	30	100	0	0	100	0	0	0	0

table represent expanding vocabulary levels (and increasing number of epochs). Error coding categories are:

SUF: The stem is regularized. For regular stems this indicates that an inappropriate suffix is affixed.

ID: The stem and past tense have the same form.

VC: The stem undergoes a vowel change. For vowel change stems this indicates that an inappropriate vowel change occurs.

BLD: The stem is blended i.e., it undergoes both vowel suppletion and suffixation.

These categories account for the overwhelming majority of errors observed. Residual errors are mostly incorrect mapping of consonants. Scores indicate percentage of total errors in a given class. The exact pattern and timing of errors differs across expansion training schedules. However, only those findings applicable to all expansion conditions, represented by the epoch 5 condition, are presented.

First, note that the overall level of errors is low and circumscribed to a limited range of error types. Second, different verb classes are susceptible to different patterns of errors. For example, regular stems are typically identity mapped, given an inappropriate suffix, or blended. While identity mapping and inappropriate suffix errors occur throughout the training period, blends are more likely to occur later in training. Following the phonological systematicities in the vocabulary, regular stems which are identity mapped are likely to end in a dental final consonant. However, in some cases, non-dental final regular stems are also identity mapped. Identity stems are typically suffixed or undergo a vowel change. Interestingly, all identity stems which undergo an erroneous vowel change possess the requisite vowel/consonant stem final combination. It is also noteworthy that identity stems *never* undergo blending. Finally, vowel change stems display the greatest range of error types, generally consisting of identity mapping, inappropriate vowel changes, and blending. Regularizations are generally rare on vowel change stems.

The previous analysis summarized the proportion of all errors generated by the network for each class, irrespective of whether the network had mastered any of those forms at an earlier point in training. In order to investigate *reorganizational* processes in these networks, Table 4 presents the proportion of stems in each verb class that undergo U-shaped development in each of the five epoch increment conditions. Here, a stem is defined as undergoing U-shaped development if it is correctly produced by the network, then at some subsequent point in training, it is incorrectly mapped and finally, again correctly mapped by the network.

Table 4: Proportion (%) of U-Shaped Verbs by Verb Class

	EPOCH EXPANSION				
CLASS	1	2	5	10	25
ARB	0.0	0.0	0.0	0.0	0.0
REG	20.7	18.3	17.7	12.7	14.0
ID	35.0	25.0	25.0	20.0	20.0
VC	35.0	40.0	30.0	30.0	30.0

These data suggest the following generalizations: First, the proportion of each class of stems which is correctly output and then subsequently erroneously produced is greater for the irregular than regular verbs. Second, as the rate of vocabulary expansion increases, the proportion of U-shaped stems tends to increase for both regular and irregular verbs.

Analyses of error types for the U-shaped errors on regular stems reveal patterns that are consistent with the general learning error analysis presented in Table 3. That is, when a regular verb is U-shaped, it is most likely to be identity mapped, blended and inappropriately suffixed. Interestingly, comparisons across training condition suggest that as expansion rate is slowed (i.e., 10 and 25 epoch increments), the tendency for blending increases and the tendency for inappropriate suffixation decreases. For irregular verbs, U-shaped errors are most likely to result from suffixation, blending, or incorrect vowel changes. These patterns appear to be consistent across training condition. However, the small absolute number of U-shaped irregular stems precludes the identification of stable trends.

U-shaped errors were also analyzed in terms of the point during training in which they occurred, i.e., at what point in training after the verb was correctly mapped did the first *incorrect* output occur. In general, this analysis reveals that the onset of U-shaped errors tends to be distributed relatively evenly across the entire training cycle. However, the data also suggest that there is an interaction between the proportion of verbs undergoing U-shapes during any given period of training and the training condition. In particular, the data suggest that increasing the rate of vocabulary expansion tends to increase the tendency to make an U-shape error early in training. Table 5 summarizes this finding, presenting the proportion of U-shape *onsets* observed in four training periods, for each of the five epoch expansion schedules.

Table 5: Proportion (%) of U-Shape Onsets on Regular and Irregular Verbs by Training Period

	EPOCH EXPANSION				
PERIOD	1	2	5	10	25
20–140	36.7	19.6	17.2	19.1	5.4
141–260	32.1	40.2	38.7	42.6	51.4
261–380	15.6	20.6	21.5	19.1	24.3
381–500	11.9	16.5	21.5	20.6	21.6

An error type by training period analysis on these onsets indicates a relatively even distribution of types of U-shapes across the training period for the *regular verbs.* That is, blends and identity mapping U-shapes were equally likely to occur early and late in the training period. However, there is a tendency for inappropriate suffix U-shapes to occur early in training.

In contrast to the regular verbs, the great majority of the U-shapes on irregular verbs (52 of 58 or 89.7%) occur during the first half of training. The occurrence of irregular U-shapes is split fairly evenly across this period in the five training conditions, although there is some evidence to suggest that U-shapes are more likely to occur during the later phases of learning as the rate of epoch expansion is decreased (i.e., in the 10 and 25 epoch increment condition). Thus, the pattern of U-shaped errors on stems in the regular and irregular classes is quite similar across epoch expansion conditions.

3.2.3 Novel Verbs

The preceding analyses concentrate on the network's ability to produce the appropriate past tense forms of stems that were members of the training set, i.e., trained verbs. The following analyses investigate network performance when it is required to produce the past tense forms of stems that it has never seen. Figure 2 outlines responses to novel stems in the 5 epoch expansion condition. As with the error analyses presented above, the findings discussed are applicable to all epoch expansion conditions. Figure 2(a) plots the tendency of the network to treat novel stems that end in a dental (Dental-Final) as if they were regular, identity mapping, or vowel change forms. Figure 2(b) plots performance on novel stems which possess stem final vowel-consonant clusters characteristic of vowel change verbs (Vowel Changes). Lastly, Figure 2(c) plots network performance on novel stems which do not possess any particular phonological sub-regularities (Indeterminates). These data are relevant to understanding the network's sensitivity to the predictable phonological characteristics of a stem (or lack thereof) when generating past tense forms, as well as changes in the tendency of the network to generalize regularities across learning.

These figures reveal that the network is indeed sensitive to the phonological properties of stems when generating the past tense forms of novel verbs. First, note that dental final verb stems are more likely to be identity mapped (28%)[8] than verb stems which do not end in a dental (7%). Similarly, novel stems that possess particular Vowel/Consonant stem final clusters are more likely to undergo vowel suppletion (37%) than those which do not possess those sub-regularities (11%). In addition, the network appears to be sensitive to a lack of phonological

[8] The percentages quoted here are means for the whole training period.

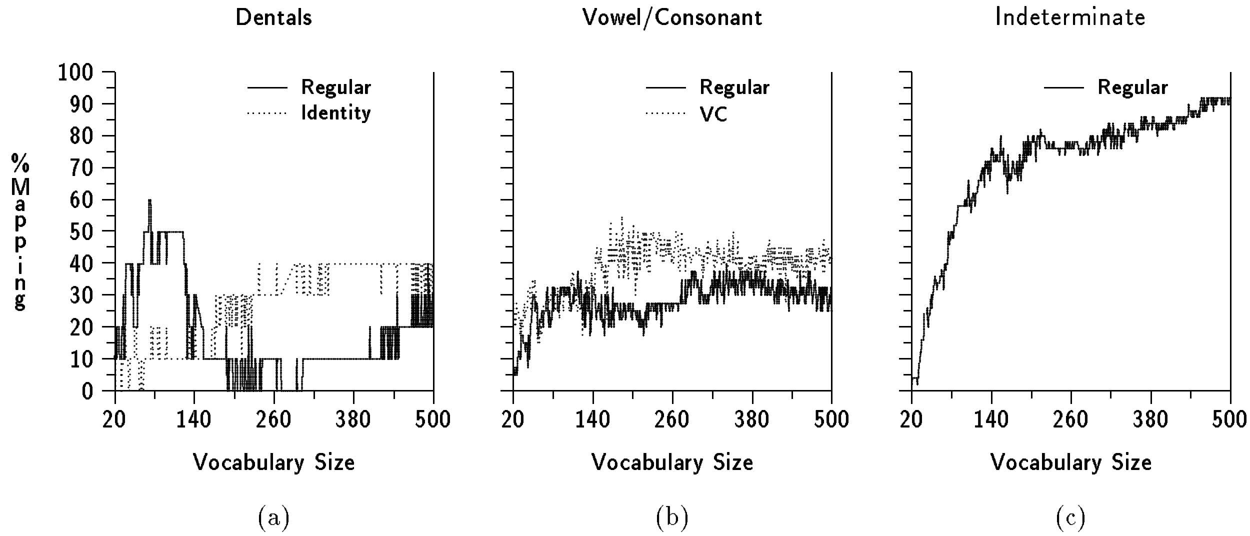

Figure 2: Performance on Novel Stems

properties in its tendency to regularize novel stems, although this pattern is less absolute. For example, there is a striking tendency for indeterminate novel stems to undergo regularization (71%); however, a substantial proportion of the dental and vowel/consonant novel stems are also regularized (18% and 26% respectively), especially towards the end of training. Note that very few indeterminate stems undergo identity mapping (3%) or vowel suppletion (0.4%).

Focusing on changes in the tendency to regularize indeterminate novel stems across learning, we observe that the regularization properties of the network alter dramatically between early and late training. Early in training, i.e., immediately after the network has mastered the initial vocabulary of 20 verbs, indeterminate novel stems are treated in an unsystematic fashion. Network output is unclassifiable in terms of the mapping categories used in Table 3[9]. However, as vocabulary size expands (and training continues), the tendency for the network to add a suffix to indeterminate novel stems increases quickly and substantially. Interestingly, this rapid increase occurs at about the same point in training (50 verbs) at which performance on trained verb stems begins to show recovery from the initial performance decrement (cf., Figure 1(b) with Figure 2(c)). Furthermore, while a major proportion of indeterminate stems are regularized by the network by the 140 vocabulary mark (76%)[10], the tendency of the network to regularize novel stems continues to increase, albeit at a less marked rate, as vocabulary size (and training) also increases.

The sudden onset of the systematic treatment of novel stems indicates that abrupt reorganization processes are occurring in the weight matrix of the network. However, it is unclear whether these changes are the result of prolonged training or the result of the network's exposure to an increasing number of different stems. In section 3.2.1, we report the sudden recovery from error on *trained* verbs. Figure 3 plots the performance of the network in mapping *indeterminate novel verbs stems* to regular past tense forms as a function of training (by epoch) on a series of vocabulary sizes[11]. These generalization curves indicate that final vocabulary size, rather than amount of training (number of epochs), is a better predictor of final level of generalization. That is, the ability of the network to generalize the suffix to novel stems continues at a low rate when the final vocabulary size is small. However, the relationship between changes in final vocabulary size and final level of generalization is non-linear. In particular, a substantial increase in generalization tendencies is observed when final vocabulary size is increased beyond 30 verbs.

[9] During this early period of training, only one of the novel indeterminate stems is treated in a systematic fashion. It is regularized. The tendency to treat Dental-Final novels and Vowel/Consonant novels systematically is greater, including 3 regularizations, 3 identity mappings and 11 vowel changes.

[10] Note that this point in training corresponds to a period in which performance on *trained* irregular verbs is low (see Figure 1).

[11] Bullets (•) indicate the point on the generalization curve, where vocabulary expansion is halted for each control simulation. The open circle (∘) indicates where vocabulary expansion starts for each control simulation. Vocabulary sizes are indicated in Figure 3.

4 Discussion

In comparison to previous neural network models of the acquisition of the English past tense, the simulations reported here achieve a high level of performance within a reasonable period of training. For example, using similar network architectures and token frequency manipulations on an identical vocabulary configuration, Plunkett & Marchman [In press] report only an average of 80% mastery on verbs in the regular and vowel change classes. The current set of simulations result in overall learning levels of 98% and 100% on these verb classes, respectively, averaged across the five epoch increment conditions.

The improved performance in the current simulations can be attributed directly to the use of an incremental epoch expansion learning schedule, rather than training the network on the entire set of 500 verbs from the initial point in training. On problems that are radically different from the mapping of verb stems to their past tense forms, other researchers (Cottrell & Tsung [1989]; Elman [1989]) have also noted improved overall learning in networks that are trained on subsets of the data prior to expanding to the full set. However, the facilitatory effect of an expansion training procedure is not well-understood. Given the statistical nature of these systems, it is likely that limiting the size and/or sampling of a problem domain reduces the probability that the network will extract spurious correlations in the data set. Provided that initial data sets are sufficiently representative of the overall problem space, training on limited data sets increases network efficiency in uncovering the *principal components* of variation that define the problem domain. Once these principal components are encoded in the weight matrix of the network, training on an expanded sample serves to reinforce the initial organization of the weight matrix and reduces the identification of spurious correlations. In addition, the network is better equipped to extract lower order components of variation. In the current context, lower order correlations might correspond to the phonological sub-regularities that characterize the irregular classes.

Several prominent theories of cognitive development have explored the relationship between the current knowledge state of the child and the nature of the problem domain to which the child is exposed with respect to determining the child's success or failure. For example, Piaget [1953] introduces the notion of *moderate novelty* to summarize the finding that children display the most advancement in those

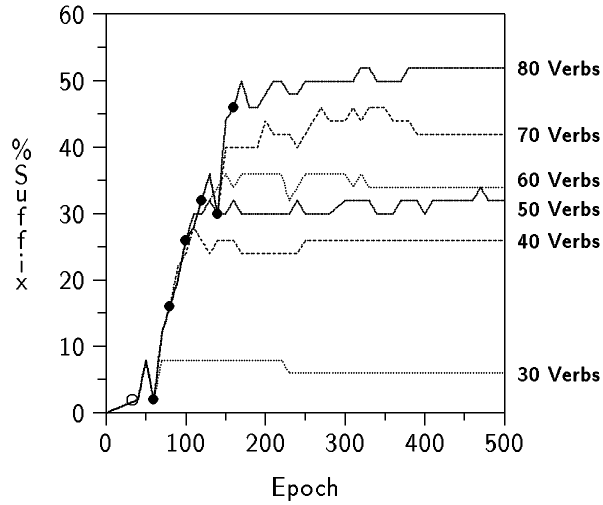

Figure 3: Generalization on Indeterminate stems by Vocabulary Size

conditions where new problems only exert moderate demand on their current knowledge state. Incorporating slightly different components, Vygotsky [1962] utilizes the notion of *zone of proximal development* for similar purposes. In these simulations, the interaction between the current knowledge state of the network, as encoded in the weight matrix, and the nature and size of the problem space to which the network is exposed can be seen to account for many of the observed behaviors. Networks, like children, appear to benefit from a moderate novelty effect, such that if the initial knowledge state of the network is undifferentiated with respect to the problem at hand (e.g., a random weight configuration), overall performance is enhanced if the learning set is initially restricted and then gradually expanded.

However, recall that these networks have considerable difficulty mastering the past tense mapping problem if the conditions for expansion involve learning each verb to criterion before any additional verbs are added to the vocabulary. Indeed, the upper limit on the number of verbs that the network was able to master was several orders of magnitude smaller in the criterial expansion condition (i.e., 25–30 stems) than in the epoch expansion conditions (i.e., 500 stems). Criterial learning results in weight changes which are derived from only one training pattern (i.e., the new verb), given that the error signal resulting from old verbs has already been minimized. In effect, the network is no longer attempting to solve a global mapping problem (i.e., at the level of the overall problem domain), but it is instead attempting to adjust the weight matrix to the mapping characteristics of a single pattern. Eventually, the network gradually entrenches itself in an increasingly isolated partitioning of the weight space, such that it is impossible for the gradient descent learning algorithm to accommodate to the mapping demands of the new verb stem. The network is trapped in a local minimum, and even indefinite training on the single new verb has little or no effect on performance.

Many theories of cognitive development emphasize the importance of continuous variation of environmental stimulation for cognitive advancement (though also acknowledging the role of maturational factors and internal reorganization processes (Karmiloff-Smith [1986])). It is unlikely that children attempt to learn verb morphology, or other types of tasks for that matter, by limiting the domain of exemplars to just one greater than what has already been mastered. Most current account of lexical acquisition assume that children learn many items in parallel, and *partial* mastery of a lexical items or inflectional systems may best characterize much of development. The gradual but continuous expansion of the problem space that was illustrated in these connectionist simulations can be seen to parallel those aspects of learning in children which ensure that they do not become entrenched in the irrelevant details of a particular problem and are forced to satisfy a multiplicity of constraints characterizing the problem domain.

One major goal of this work was to determine the nature of the mechanisms that trigger the transition from rote learning to system building in children acquiring the English past tense. Two sets of results from the simulations presented here indicate that artificial neural networks can also be observed to pass from a period of rote learning to a period of system building early in learning. A third set of results illustrates that, in these networks, this transition is triggered by quantitative factors relating to the number of verbs which share specific mapping properties and characteristic phonological features.

First, after the initial set of 20 verbs are learned to criterion, the network does not generalize the regularities in that set to new stems. In the period immediately following initial training, both regular and irregular verbs are prone to incorrect mapping (see Figure 1). In addition, the errors produced by the network during this period are not systematic, i.e., they do not appear to reflect the mapping characteristics of the initial learning set (see error tabulations in Table 3). Interestingly, however, the initial 20 verbs are likely to continue to be mapped correctly by the network, even in the presence of the other erroneous mappings[12]. This pattern of results indicates that the network has failed to extract any pattern of regularities from the initial set and treats each stem/past tense pair as an independent mapping problem, i.e., the network has *memorized* the initial set of 20 stem/past tense pairs. Overall performance on new verbs (particularly on regular verbs) continues to deteriorate as new verbs are added to the training set (see Figure 1).

The deterioration in performance on newly introduced verbs is reversed when the vocabulary reaches a specific size (around 50 verbs). At this point in training, performance on new verbs improves rapidly. In addition, errors are now observed on several irregular verbs that were members of the initial training set. The timing of this turnaround in performance appears to be stable across all epoch expansion schedules tested here, and the same patterns of learning have been replicated in networks learning a different random selection of the total vocabulary. These findings suggest that the network has reorganized its representation of the mapping problem, to the extent that the network now treats new verbs entering the vocabulary in a systematic fashion. The fact that the timing of the turnaround is remarkably consistent across the training conditions tested here suggests that the trigger for the transition from rote learning to system building in these networks is associated with the *quantity* of verbs in the training set which undergo systematic mapping processes.

Secondly, the performance of these simulations on novel verbs also confirms the interpretation that the network can be seen to pass from a stage of rote learning to system building. In the period following the initial training on the 20 verbs, responses to novel stems is unsystematic, irrespective of their phonological characteristics (see Figure 2). Thus, correct performance on the initial 20 verbs does not transfer to novel verbs, and the corresponding successful output on verbs in the training set again cannot be attributed to the generalization of mapping characteristics. As vocabulary size is expanded (and training continues), the tendency to treat novel verbs in a systematic fashion increases rapidly. For example, there is a clear tendency to regularize those novel stems which do not possess phonological subregularities characteristic of the irregular classes (i.e., the indeterminates). In addition, novel dental or vowel change stems are now likely to be mapped in accordance with their phono-

[12] It will be recalled (see footnote 4) that regular verbs in the initial training set are only correctly mapped if they have a high token frequency — another symptom of rote learning in these networks.

logical shape, or they are regularized.

Crucially, the onset of the systematic treatment of novel verbs stems *coincides with the turnaround in the performance on trained verbs.* Thus, the network is consistent in its treatment of trained and novel stems. At this point in training, the network appears to have modified its representation of the verb mapping problem from one in which individual stem/past tense form mappings are encoded independently to one in which classes of mappings can be differentiated into categories. As a result, the process of regularization can be seen to act like a "default" transformation, in that it is likely to be applied to the majority of novel stems, especially when a stem lacks a characteristic phonological shape. The various processes of irregularization, in contrast, are applied to a smaller subset of novel stems and are contingent on the presence of phonological subregularities.

A third set of simulations demonstrates that the *quantity* of stems in the training set which undergo systematic mapping serves to trigger the transition from the period of rote learning to the period of system building in these networks. The control simulations illustrated in Figure 3 suggest that size of training set is the best predictor of the regularization of indeterminate novel verb stems. In particular, the size of the training set must be increased beyond 30 verbs before substantial regularization tendencies are observed. This finding clearly does warrant the conclusion that an absolute critical mass of verbs can be defined as a prerequisite for generalization. However, the non-linear relationship between the tendency to generalize the regular mappings and size of training set suggests that repeated training on a small set of mappings will not in itself lead to generalization. The network must be exposed to a sufficient range (types) of mappings. Yet, adequate sizes of training sets cannot be defined independently from the specific network architecture employed and the overall set of mapping characteristics that the network must learn.

In general, these results support the view that a *single mechanism* learning system can offer an alternative account of the transition from rote learning processes to system building in children's acquisition of English verb morphology. In contrast to the traditional view which posits an interaction between two qualitatively distinct mechanisms supporting different modes of representation (i.e., rote and rule), a connectionist account posits a single mechanism driven by a general learning algorithm which is capable of both memorization and generalization processes. The network's transition from rote-like to rule-like processes are not triggered by the interaction of qualitatively distinct mechanisms, but instead by *quantitative* increments in the size of a structured training set.

The behavior of these networks can be seen to mimic several aspects of the type and timing of children's pattern of morphological acquisition. However, it is as yet impossible to determine the degree to which children's transition from rote learning to system building is driven by similar quantitative changes in the size of the learning set. There is abundant evidence that many children pass first through a prolonged period of development in which they only produce a limited number of verbs and their corresponding appropriate past tense forms. This period is then superceded by one in which verb vocabulary expands at a fairly rapid pace, and errors are likely to occur on irregular (and sometimes regular) past tense forms. However, sufficient longitudinal evidence is not yet available to determine whether vocabulary size operates in a critical mass fashion for children. Furthermore, it is difficult to anticipate the contributions of comprehension or other linguistic processing factors in determining the nature and timing of this transition in children.

However, it is interesting to note that similar relationships between set size and the onset of reorganizational processes have been hypothesized with respect to the "vocabulary (or naming) spurt" that is observed in young children during the second half of their second year. Several researchers (Bates, Bretherton & Snyder [1988]; Nelson [1973]; Plunkett [1990]) have noted that there is a shift in the rate of vocabulary acquisition after the point in learning when overall vocabulary reaches around 50 words. Although the interpretation of this finding is controversial, many researchers (e.g., Barrett [1982], Bowerman [1982]) argue that such a transition is associated with a *reorganization* in the structure of children's vocabularies. Prior to this point, word meanings tend to be encoded independently as separate and distinct lexical items; whereas, the vocabulary spurt signals that children's lexicons are now organized into coherent semantic systems. This interpretation is consistent with the view that the transition from rote learning to system building can derive from processes inherent in a single mechanism connectionist-like system, driven by quantitative changes in the size of the learning set. A similar view has been espoused by Karmiloff-Smith [1986].

The categories of general learning errors reported for these simulations resemble those reported in earlier work (Plunkett & Marchman [In press]). In particular, patterns of multi-lateral interference are observed between the mapping classes resulting in both regularizations and irregularizations. Furthermore, corroborating earlier findings, different verb classes are susceptible to different types of errors. When errors are made, regular verbs are most likely to be identity mapped; identity verbs are most likely to be suffixed. Vowel change verbs are subject to the widest range of errors. The patterns of timing of errors are also similar to that reported in earlier work, e.g., blending errors tend to be a characteristic of late training. However, because of the high level of performance in these simulations, errors on irregular verbs in the second half of training are virtually absent. Instead, errors on irregular verbs tend to be clustered around the beginning and middle periods of training. Errors on regular verb stems are observed throughout training.

We have argued elsewhere (Plunkett & Marchman [In press]) that the timing and pattern of errors produced in these networks resemble those produced by young children acquiring English verb morphology. Children (and adults) produce incorrect past tense forms over a wide span of development and the types of errors produced suggest a sensitivity to the phonological properties of the irregular classes. Such sensitivities can sometimes interfere with the appropriate production of regular past tense forms, just as the pattern of suffixation interferes with the appropriate production of irregular forms. These multi-lateral patterns of interference are an inherent property of an interconnected connectionist system attempting to map several competing classes of verb types.

In our analyses of the errors produced by these networks, we distinguish errors produced across the course of *learning* from errors that resulted from an apparent *unlearning* of stem/past tense mappings (i.e., those stems that have been produced correctly at some point earlier in training). Much of the empirical foundation for the traditional view of acquisition has been concerned with the latter type of error, perhaps because the phenomena of U-shaped regressions in performance is more theoretically striking than the hypothesized processes which guide the gradual linear acquisition of a verb. Clearly, however, the results from these simulations predict that children's erroneous output results from both types of errors. In these networks, general learning errors typically occur early in training; whereas the occurrence of U-shaped errors is distributed fairly evenly throughout the training period. Interestingly, as with children, the proportion of irregular verbs that are U-shaped is greater than the proportion of regular verbs that are U-shaped. However, irregular mappings are all mastered by the 300 vocabulary size mark. Unfortunately, empirical data comparing U-shaped errors to patterns of general learning errors are not available for children.

Lastly, we tested several types of learning schedules, focusing primarily on the rate at which vocabulary size is expanded across training. The findings suggest that high expansion rates (i.e., 1 epoch increment) are likely to result in:

1. Greater decrements in performance on both regular and irregular verbs early in training.
2. A higher final level of performance on regular verbs.
3. A shorter and less marked performance decrement on irregular verbs in the middle portions of the training period.
4. A greater proportion of stems undergoing U-shaped learning, though these are more likely to be restricted to early periods in training.

At this point, these results can only offer speculative predictions for the study of language acquisition. However, we suggest that the rate of vocabulary expansion that we have modeled in these networks might prove to be analogous to the rate with which young children attempt to assimilate new verbs into their vocabularies. It is possible, for example, that children who vary in their rate of vocabulary acquisition (e.g., "late talkers" vs. "early talkers" (Bates & Thal [In press])) may also vary in the type and timing of overgeneralization errors. In particular, the results from these series of simulations predict that early language learners would be more likely to produce errors earlier than later in acquisition, and U-shaped errors would be less likely to occur later in development. Late language learners, in contrast, would be likely to achieve a slightly lower level of final performance and produce a greater proportion of errors on verbs that do not match the dominant pattern, i.e., the irregular verbs. These predictions are clearly only speculations which must await empirical findings. Nevertheless, the simulation work does suggest that qualitatively different patterns of learning can arise from quantitative manipulations of the input and training schedules.

5 Conclusion

Our current understanding about children's acquisition of English inflectional morphology is typically based on naturalistic observations of children's actual productions, and experimental studies using nonsense words (e.g., Berko [1958]). Naturalistic studies typically report that overgeneralization errors comprise a small portion of children's productive vocabularies, however, children do produce a range of regularization and irregularization errors. In contrast, studies using nonsense forms typically demonstrate an overwhelming preference for regularization. However, several more recent studies suggest that children are indeed sensitive to the phonological aspects of nonsense forms. For example, (Derwing & Baker [1986]) have shown that novel stems which end in a dental tend to be identity mapped and novel stems with specific vowel/consonant clusters will sometimes undergo vowel suppletion.

In these simulations, analyses of network performance on *trained* verbs is analogous to naturalistic observations of children's performance on real English words. Correspondingly, network performance on novel verbs is best viewed to correspond to the experimental elicited production studies with nonsense verbs. Interestingly, the overall pattern of results suggests a considerable correspondence between children and networks across these two types of measures. For example, the pattern of errors observed across learning for verbs in the training set are multi-lateral i.e., irregular verbs are regularized and regular verbs are irregularized. In contrast, responses to novel verbs are overwhelmingly more likely to involve regularization. For those novel stems that phonologically resemble irregular verbs, the tendency to regularize is still prominent but identity mappings and vowel suppletions do occur. Thus, evaluations of network performance on novel verbs (in comparison to patterns of errors across learning) suggests that the network, like the child, is best characterized as a rule-governed system. We suggest that evaluations of acquisition using experimental studies using nonsense verbs, while clearly illustrating the generalization abilities of young children, may have succeeded in biasing our view of the nature of the phenomenon to be explained. While we accept that experimental findings contribute important insights into the developmental process, they do not obviate the need for detailed naturalistic studies of children's acquisition of verb morphology. Indeed, it is precisely such data that are needed to evaluate the predictions of the present model in a stringent fashion.

References

Barrett, M. D. [1982], "Distinguishing between Prototypes: The early acquisition of the meaning of object names," in *Language Development*, S. A. Kuczaj, ed. #1: Syntax and Semantics, Lawrence Erlbaum Associates, Hillsdale, NJ.

Bates, E., Bretherton, I. & Snyder, L. [1988], *From First Words to Grammar: Individual Differences and Dissociable Mechanisms*, Cambridge University Press, Cambridge, MA.

Bates, E. & Thal, D. [In press], "Associations and dissociations in child language development," in *Research on child language diorders: A decade of progress*, J. Miller & R. Schiefelbusch, eds., Pro-Ed, Austin, TX.

Berko, J. [1958], "The child's learning of English morphology," *Word* 14, 150–177.

Bever, T. G., ed. [1982], *Regressions in mental development: Basic phenomena and theories*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Bowerman, M. [1982], "Reorganizational process in lexical and syntactic development," in *Language Acquisition: The State of the Art*, E. Wanner & L. Gleitman, eds., Cambridge University Press, Cambridge, MA.

Brown, R. [1973], *A first language: The early stages*, Allen & Unwin, London.

Bybee, J. & Moder, C. L. [1983], "Morphological classes as natural categories," *Language* 59, 251–270.

Bybee, J. & Slobin, D. I. [1982], "Rules and schemas in the development and use of the English past tense," *Language* 58, 265–289.

Bybee, J. L. [1988], "Morphology as lexical organization," in *Theoretical morphology*, M. Hammond & M. Noonan, eds., Academic Press, New York, NY.

Carey, S. [1982], "Semantic Development: The state of the art," in *Language Acquisition: The State of the Art*, E. Wanner & L. Gleitman, eds., Cambridge University Press, Cambridge, MA.

Cottrell, G. & Tsung, F. [1989], "Learning simple arithmetic procedures," *Proceedings of the Eleventh Annual Cognitive Science Society Conference, Ann Arbor, MI*, Hillsdale, NJ.

Derwing, B. L. & Baker, W. J. [1986], "Assessing morphological development," in *Language acquisition: Studies in first language development. Second edition*, P. Fletcher & M. Garman, eds., Cambridge University Press, Cambridge.

Elman, J. L. [1989], "Representation and Structure in Connectionist Models," Center for Research in Language, University of California, San Diego, Technical report #8903.

Karmiloff-Smith, A. [1986], "From meta-processes to conscious access: Evidence from children's metalinguistic and repair data," *Cognition* 23, 95–147.

Kuczaj, S. [1977], "The acquisition of regular and irregular past tense forms," *Journal of Verbal Learning and Verbal Behavior* 16, 589–600.

MacWhinney, B. [1987], "The competition model," in *Mechanisms of language acquisition*, B. MacWhinney, ed., Lawrence Erlbaum Associates, Hillsdale, NJ.

Marchman, V. [1984], "Learning not to overgeneralize," *Papers and reports on child language development* 24, 69–74.

Marchman, V. [1988], "Rules and regularities in the acquisition of the English past tense," *Center for Research in Language Newsletter* 2.

Nelson, K. [1973], "Structure and strategy in learning to talk," *Monographs of the Society for Research in Child Development* 38.

Piaget, J. [1953], *The origin of intelligence in the child*, Routledge & Kegan Paul, London.

Pinker, S. & Prince, A. [1988], "On language and connectionism: Analysis of a parallel distributed processing model of language acquistion," *Cognition* 28, 73–193.

Plunkett, K. [1990], "Identifying formulaic expressions in early language acquisition: The case for articulatory/fluency criteria," *Center for Research in Language Newsletter* 4.

Plunkett, K. & Marchman, V. [1989], "Pattern association in a back propagation network: Implications for language acquisition," Center for research in language, University of California, Technical Report # 8902, San Diego.

Plunkett, K. & Marchman, V. [In press], "U-Shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition," *Cognition*.

Rosenblatt, F. [1962], *Principles of neural dynamics*, Spartan, New York.

Rumelhart, D. E. & McClelland, J. L. [1986], "On learning the past tense of English verbs," in *Parallel distributed processing: Explorations in the microstructure of cognition, #2: Psychological and Biological Models*, J. L. McClelland, D. E. Rumelhart & PDP Research Group, eds., MIT Press, Cambridge, MA, 216–271.

Slobin, D. I. [1985], "Crosslinguistic evidence for the language-making capacity," in *The crosslinguistic study of language acquisition #2: Theoretical issues*, D. I. Slobin, ed., Lawrence Erlbaum Associates, Hillsdale, NJ.

de Villiers, J. G. & de Villiers, P. A. [1985], "The Acquisition of English," in *The Crosslinguistic Study of Language Acquisition (Vol 1)*, D. I. Slobin, ed., Lawrence Erlbaum, Hillsdale, NJ.

Vygotsky, L. [1962], *Thought and language*, MIT Press, Cambridge, MA.