Attention and U-Shaped Learning in the Acquisition of the Past Tense

Dan Jackson* (JACKSON@LING.UCSD.EDU)
Garrison W. Cottrell† (GARY@CS.UCSD.EDU)
*Cognitive Science & Linguistics 0108
†Computer Science & Engineering 0114
*,†Institute for Neural Computation
University of California, San Diego
La Jolla, CA 92093 USA

Abstract

Plunkett & Marchman (1993) showed that a neural network trained on an incrementally expanded training set was able to master the past tense and show the U-shaped learning pattern characteristic of children. In Jackson, Constandse & Cottrell (1996) we argued that Plunkett & Marchman's restriction of the training set was unrealistic and proposed a model of selective attention that enabled our network to master the past tense without external restrictions on its training set. Analysis in the present paper shows that the network in Jackson, Constandse & Cottrell (1996) does not exhibit appropriate U-shaped learning, however. We propose a modified model of selective attention that results in the mastery of the past tense as well as the kind of U-shaped learning observed in children.

Introduction

In the process of learning the past tense, children typically show what has been called a "U-shaped" pattern of development. The first past tense forms produced are generally correct, regardless of whether or not those forms are regular. After this period of correct performance, children go through a period of overregularization in which irregular forms are inflected with the regular suffix (e.g. *goed*). Finally, children seem to identify some forms as exceptions to the general regular pattern, and the overgeneralization errors decrease. This pattern of acquisition has been called "U-shaped," for obvious reasons—the performance starts off high, then goes down and finally comes back up again. Actually, this is something of a misnomer because it implies that children enter a period of development in which the regular rule is consistently applied to all verbs. In fact, children produce correct past tense irregulars at the same time as they overregularize others, and sometimes alternate within a short time between the correct and incorrect past tense form of the same irregular verb (Kuczaj, 1977, 1978; Bybee & Slobin, 1982; Plunkett & Marchman, 1991). At all points in development, overregularizations are a relatively small proportion of children's total past tense production (Marchman, 1988; Marcus, Pinker, Ullman, Hollander, Rosen & Xu, 1992). Marcus *et al.* (1992) investigated the rate of overregularization shown by the children in the CHILDES database (MacWhinney, 1990). They defined the overregularization rate as the proportion of tokens of irregular past tense forms that are overregularizations:

$$\frac{\text{\# overregularization tokens}}{\text{\# overregularization tokens + correct irregular past tokens}}$$

They graphed 1 overregularization rate for 4 children (Adam, Eve, Sarah and Abe). For all of the children but Abe, there was an initial period of no overregularization and the rate of overregularization was small throughout development (typically <10%). Figure 1 shows the overregularization rate for Adam. Note that the graph shows 1 - overregularization rate, so when the graph is at 100%, the overregularization rate is zero. Thus, the series of points at 100% in the initial part of the graph indicate the initial period of no overregularization. When the denominator in the overregularization rate is zero, the point is not plotted in the graph.

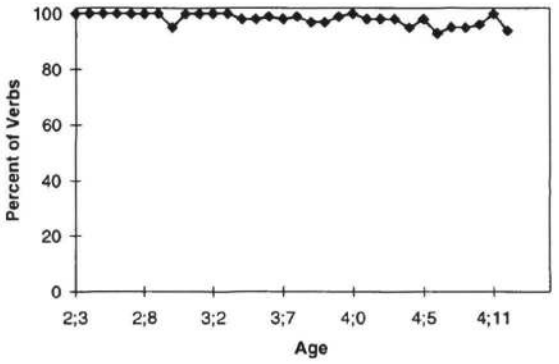

Figure 1: (1-overregularization rate) for Adam (reproduced from Marcus *et al.* (1992)).

Plunkett & Marchman (1991, 1993) (P&M hereafter) have shown that overregularization behavior can be modeled using a single mechanism in the form of a

connectionist network. In response to criticism of the discontinuous training set used in Rumelhart & McClelland's (1986) model of the acquisition of the past tense (Pinker & Prince, 1988; Lachter & Bever, 1988; Marcus *et al.*, 1992), P&M (1991) showed that a neural network will make overregularization errors without such discontinuities in its training set. Unfortunately, their network did not have the initial period of no overregularization that is characteristic of children. Furthermore, the final performance reported (after 50 epochs) was 100% correct for the arbitrary and identity mappings, but only 80% for vowel change verbs and 85% for regulars. Since adult humans are capable of correctly inflecting nearly 100% of regulars, the network performance left something to be desired.

P&M (1993) showed that networks can achieve acceptable levels of performance and show the initial stage of no overregularization that characterizes U-shaped learning if their training set is expanded incrementally. Trained on an incrementally expanded set of verbs, the network described by P&M (1993) was able to master the vocabulary (correctly inflecting 100% of the irregular verbs and 97-98% of the regulars). The network was also able to model U-shaped learning. In particular, it showed the kind of overregularization behavior that Marcus *et al.* (1992) found for children: an initial period where no overregularization occurred, followed by a protracted period where low rates of overregularization were observed, followed by the correct production of both irregular and regular past tense forms.

In Jackson, Constandse & Cottrell (1996) we criticized P&M (1993), claiming that their training regime was unrealistic. At the outset of training, the network was given 20 verbs, on which it was trained to 100% accuracy before expansion began. After that, a new verb was introduced every 5 epochs until the size of the training set was 100. Then one new verb was introduced per epoch until the size of the training set was 500. Children are exposed to the entire adult language from the beginning of development, so this restriction of the network's training set is unjustified. We developed a model of selective attention in which the network trained on items which are most *salient*. The most salient items were defined as sampled items for which network error was highest. Networks with this mechanism of selective attention learned to inflect 100% of the verbs correctly without any external manipulation of their training set. Analysis in this paper, however, shows that these networks do not show appropriate U-shaped learning. They begin overregularizing at the beginning of training and their overregularization rate is much higher than is typical of children.

In the present work we argue that the use of error alone to define salience is unrealistic and leads to the model's inability to show appropriate U-shaped learning. We implement a new selective attention model that incorporates frequency information into the criterion for salience. This model learns the entire training set and shows U-shaped learning similar to what is seen in children, without any restrictions on its training set.

Selective Attention Model

The selective attention model is based on the method of active selection (Plutowski & White, 1993). This method was originally used for incrementally growing a training set by using a partially trained network to guide the selection of new examples. Instead of using active selection for incrementally growing the training set, we use it to select the training examples for each epoch.

In the selective attention model, the verb on which the network will train is chosen from a set of W items randomly sampled from the target language based on frequency. This set of W items is called the "sample window." To select a new example for training, the items from the sample window are compared with the verbs currently being trained on. The N most "salient" of these are selected for training. For each epoch, a new set of W items are selected for the sample window, and the training queue is updated. In this paper we use N=1 and W = 8. The nature of "salience" in this model will be addressed below.

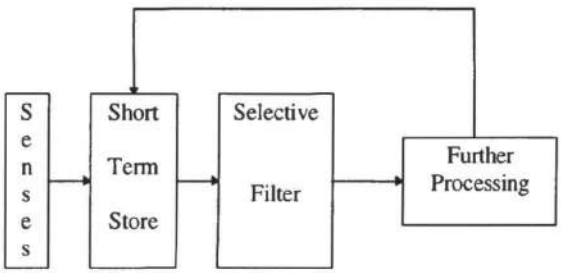

Figure 2: Broadbent (1958) model of selective attention.

This model of selective attention is similar to the general model proposed by Broadbent (1958). Broadbent's model (somewhat simplified) is illustrated in Figure 2. Items are sampled from the environment and held in a limited capacity short term memory. The items from this short term memory are filtered for further processing. The short term memory may hold information before filtering as well as retain information after it has passed through the selective filter and been processed further. The sample window in our selective attention model corresponds to the "short term store," and the verb selected for training corresponds to the item that has been selected for "further processing."

Salience

The notion of salience is central to the model described above. Plutowski *et al.* (1993) introduced the idea of using maximum error on an example as the criterion for selection (cf. Baluja & Pomerleau (1994), whose network ignores sections of the input with high prediction error). In Jackson, Constandse & Cottrell (1996), this criterion was used for selecting examples for weight adjustment. We implemented the selective attention model described above, with error used in place of "salience." Network error on a

particular verb may be thought of as a measure of the "novelty" of that verb to the network.

The novelty of a stimulus is well established as a factor influencing attentional response. Sokolov (1960, 1963, 1969) argued that the incongruity between an incoming stimulus and existing neuronal templates is the basis for the orienting response. Dunham (1990) showed that infants listening to temporal patterns show an increase in attention to unpredictable patterns and a decrease in attention to rhythmic, predictable patterns.

Given the interpretation of error as a measure of novelty, it seems reasonable to use error for salience, as we did in Jackson, Constandse & Cottrell (1996). Using error as salience, the networks mastered the training set without any external manipulation of their training sets. They did not, however, show the type of U-shaped learning discussed above. As we show below, these networks begin overregularizing early in development, contrary to children's pattern of development. By using error as its criteria for what is salient, the network is reducing the importance of frequency in the training set. At the beginning, the network is more likely to train on irregular verbs, because their frequency in the training set is higher. As soon as the error on these examples starts to improve, however, the network will begin training on regulars, for which its error is worse. Thus, all verb types are kept on approximately equal footing with respect to error, and the frequency of the items in the training set is, in a sense, "ignored." Children, on the other hand, do not ignore frequency. Presumably, the failure of children to overregularize early in development arises because they initially memorize forms that are highly frequent (the irregulars), and only later learn and overapply the majority, regular mapping. To the extent that our model eliminates the effects of frequency in the training set, its early overregularization is to be expected.

There is a large body of research showing that children pay preferential attention to things that are more frequent or more familiar. Jusczyk, Cutler & Redanz (1993) showed that American 9-month-olds (but not 6-month-olds) listen significantly longer to words that follow the predominant stress pattern of English words. Jusczyk, Friederici, Wessels, Svenkerud & Jusczyk (1993) showed that at 9 months (but not at 6 months) American infants listen longer to lists of English words than lists of Dutch words. Jusczyk, Luce & Charles-Luce (1994) showed that, among words allowed by the constraints of English, 9 month olds listen significantly longer to items with phonetic patterns that occur frequently than items with infrequent (but still allowable) phonetic patterns.

This may seem like a paradox: on the one hand, children pay attention to what is novel, and on the other they pay attention to what is familiar or frequent. These conflicting findings may reflect different levels of processing. Novel stimuli tend to provoke an orienting response, as Sokolov showed. All other things being equal, more attention will be paid to the stimulus that provoked the response. This makes sense, because in the process of learning about one's environment it is important to correct false expectations as well as learn about stimuli that are occurring for the first time. Conversely, stimuli that are frequent in the environment are also focused on. This also makes sense, because mastering the appropriate responses to the most frequent stimuli maximizes the likelihood of dealing appropriately with the average stimulus.

In an effort to take both of these factors into account, we developed a variant of the selective attention model in which the criteria for salience includes information about both novelty (error) and familiarity (frequency). In this model, the salience of a particular example is defined as the product of the network's error on that example and the log frequency[1] of the example. This salience measure can be thought of as "moderate novelty" and/or "moderate familiarity." Supplying the network with frequency information is justified by the fact that children have knowledge about the frequency of items in their language before they start to learn the past tense. Sensitivity to the frequency of phonetic patterns is already present in infants of 9-months (Jusczyk, Luce & Charles-Luce, 1994), while children do not typically start producing the past tense until around 20 months of age (Cazden, 1968).

In the experiments described below, we will test the behavior of three types of networks: (a) networks without selective attention (replicating P&M (1991)), (b) networks with selective attention that use error for salience and (c) networks with selective attention that use the product of error and log frequency to determine what is salient.

Methods

Our input-output pairs were taken from the database of artificial verbs used by P&M. The interested reader should refer to P&M (1991, 1993) for details about the representations. The network is given a verb stem as input and must produce the inflected verb as its output. The transformations from the stems to the past tense forms are classified into four possible classes: arbitrary, identity, vowel change, and regular. Each of these corresponds to a possible English past tense transformation.

For the arbitraries, there is no relation between the stem and the past tense form, e.g. 'go→went.' For the identities, the past tense form is identical to the verb stem. This mapping requires that the verb stem end in a dental consonant (/t/ or /d/), e.g. 'hit→hit.' For the vowel changes, a vowel in the stem may be replaced by a different vowel in the inflected form of the verb, depending on the original vowel and the consonant that follows. We had 7 different types of vowel changes in our vocabulary, analogous to 'ring→rang,' 'blow→blew,' etc. Finally, for the regulars, a suffix is appended to the verb stem. The form of the suffix depends upon the final vowel/consonant in the stem. If the stem ends in a dental (/t/ or /d/), then the

[1] Specifically, what was used was log2((token frequency) + 1). The addition of one was made because the token frequency of regulars is 1, and the log of 1 is 0. We do not want to multiply by zero, so a factor of one was added to all of the token frequencies. The log of frequency is used because, as Marcus et al (1992) note, "a frequency difference of 1 versus 10 would have a greater effect than a frequency difference of 1,001 versus 1,010 (p. 118)."

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

The type and token frequencies of each of these classes in our vocabulary are shown in Table 1. We calculated the average frequencies reported in Kucera and Francis (1967) for each of the past tense types, and created a data set that mirrored the ratios of type frequencies we found there. As Marcus *et al.* (1992) note, however, this database "should predict children's behavior less well than parental frequency counts, of course, because it is from written English addressed to adults (p. 118)." They reported that only slightly more than one-quarter of parental verb tokens in the CHILDES database were regular (p. 80). In order to make our training set resemble the input children receive more closely, we increased the token frequencies of the irregulars, keeping their relative proportions the same, so that just over one-quarter of the total tokens available to the network were regular.

Table 1 - Frequency distribution of the training set.

	Arb	Reg	ID	VC
Type Frequency	2	458	8	32
Token Frequency	216	1	9	8

We also trained all three types of network on two other vocabularies—one which had the type and token frequencies used in P&M's (1991) "Phone 34" simulation, and one with frequencies based directly on Kucera & Francis (1967). In all of these simulations, only the networks with selective attention, using both error and frequency for salience, showed appropriate U-shaped learning. Furthermore, as the training sets were made more similar to the input children receive, the networks' overregularization behavior became more similar to what is observed for children. Because of limitations on space, we will only report the results from the training set that provides the closest approximation to the input to children.

Our networks were trained with the back propagation algorithm. Each network was initialized with the same set of random initial weights. Other simulations were run with different sets of initial weights, and the results were virtually identical to those reported here. The network architecture consisted of 18 input units (each verb stem was formed from 3 phonemes each requiring 6 units to represent), 30 hidden units and 20 output units (2 suffix units were needed in addition to the transformed stem). The choice of 30 hidden units was made to parallel the architecture used by P&M (1993). The learning rate and momentum were also set according to the values used by P&M (1993), namely a learning rate of 0.1 and a momentum of 0.0. To evaluate network performance, the output for each phoneme in the stem was mapped to the closest legal phoneme (using Euclidean distance). Then the output was compared with the target.

Results

The networks with selective attention were able to master the past tense mappings completely. The network without selective attention never reached 100% correct on the regular mapping (final performance was 97% for regular verbs, 100% for irregulars). The overregularization behavior of the network without selective attention is shown in Figure 3. Like the simulation reported in P&M (1991), this network began overregularizing early in training and does not show the kind of U-shaped learning seen in children (cf. Figure 1). The overregularization behavior of the selective attention network that used error alone for salience is shown in Figure 4. This network overregularizes too early and too much. Thus, using error alone to define salience leads to learning behavior that is unlike what has been observed in children, as noted above. The selective attention network that used both error and frequency information in its criterion for salience showed overregularization behavior similar to what Marcus *et al.* (1992) reported for children (shown in Figure 5). It had an initial period with no overregularization errors, followed by a prolonged period of low overregularization rates (typically <10%), after which overregularization errors ceased.

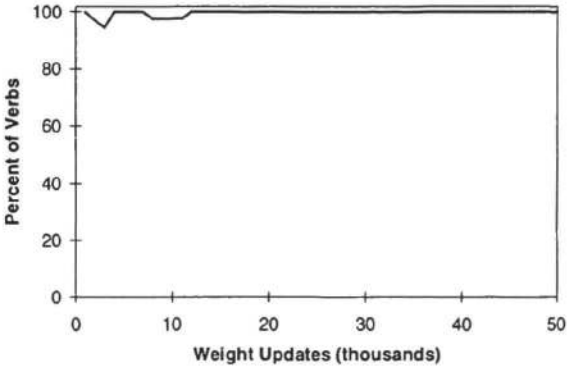

Figure 3: (1 - overregularization rate) for the network without selective attention.

This network was also tested on its ability to generalize. There were three types of novel stems: stems ending in a dental (novel dental), stems from each of the vowel change types (novel vowel change) and stems that did not fall into either of the first two categories (novel indeterminates). By the end of training, the majority of the novel indeterminate verbs (ranging from 84-88%) were inflected as regulars, showing that the network has learned that the regular mapping is the default. The other responses at the end of training were 2-4% identity (the same stem as in the input with no suffix), 4-6% wrong suffix, 2% regular suffix (appropriate for the final phoneme) accompanied by a bad vowel change (one that did not correspond to any of the seven legal vowel change types) and 4% unclassified (mostly changes to the consonants in the stem). Combining

the regular responses with those given the wrong suffix gives us 90-92% suffixation for novel indeterminates, which is what P&M (1993) report for their simulations. None of these novel indeterminate verbs were mapped with a legal vowel change or the combination of a legal vowel change and a regular suffix (blends). By the end of training, the network's responses to novel dental verbs were 20% identity, 20% regular, 10% vowel change (these stems fell into vowel change class 2, which ends in a dental), 20% wrong suffix, 20% regular suffix accompanied by a bad vowel change and 10% unclassified. Obviously the network has not learned the dental mapping as well as the regular mapping. The fact that these novel verbs ended in a dental did change the network's response to them, however, increasing the likelihood that they would be inflected as regulars or according to vowel change class 2, and also increasing the likelihood of errors—inflections that do not correspond to one of the canonical past tense mappings. At the end of training the responses to novel vowel change verbs were 21% vowel change, 29% regular, 9% identity, 10% blend, 4% bad VC, 10% bad VC with regular suffix and 17% unclassified. When confronted with novel vowel change stems, the network is more likely to inflect them as vowel change verbs or blends than when it sees either novel indeterminates or novel dentals. In fact, none of the novel indeterminates or novel dentals were treated as blends. The network's likelihood of making a response that does not look like any of the normal inflections of English also increases.

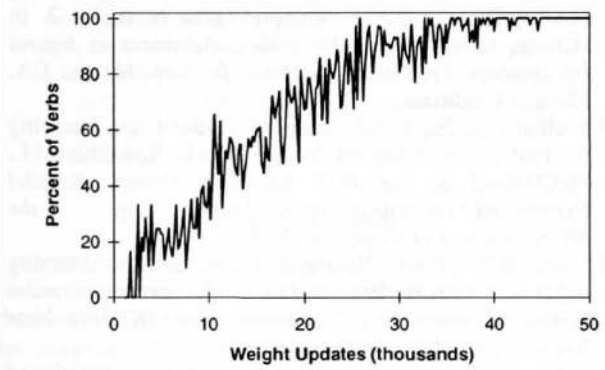

Figure 4: (1 - overregularization rate) for the network with selective attention, using error alone for salience.

These results resemble the behavior of humans when faced with a "wug test" (Berko, 1958) where they are asked to give the past tense form of nonsense words. Most words are inflected as regulars, but words that are similar to irregular English words may be inflected as irregulars. Furthermore, words that are similar to irregulars take longer to process. The fact that the network is more likely to give responses that do not correspond to any of the normal inflection types when confronted with words that look like the irregulars it has learned may be seen as analogous to this difficulty experienced by humans. The unclassified responses may potentially be eliminated by using a phonological attractor at the output (Plaut, McClelland, Seidenberg & Patterson, 1996). This is a direction for future research.

Conclusion

Selective attention was shown to be a powerful aid to learning in neural networks. Both types of selective attention networks mastered the training set completely, while the network without selective attention did not. The mechanism of selective attention allows the network to guide itself through a form of "incremental learning" (Elman, 1993) so that difficult mappings can be learned without the experimenter controlling the presentation of training examples.

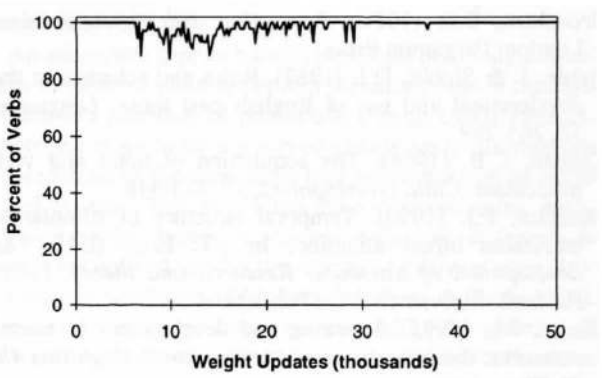

Figure 5: (1 - overregularization rate) for the network with selective attention, using both error and frequency for salience.

We also showed that using error alone as the criterion for salience leads to learning behavior unlike what is found in children. The networks that used error alone failed to go through a process of U-shaped learning. Rather, they produced overregularizations too soon and too often. When frequency information was added to the criterion for salience, however, the network's learning behavior provided a good model of what is seen in children. Both novelty (which corresponds to error in the network) and familiarity (or frequency) have been shown to play a role in determining what children pay attention to, so the use of both of these sources of information in the model of selective attention is justified. Of course, other things may play a role in determining what is salient as well. Although only frequency and novelty were utilized here, we do not mean to imply that they are the only possible determinants of attention.

These simulations also vindicate the claim, originally made by Rumelhart & McClelland (1986), and then defended by P&M (1991, 1993), that a connectionist network can provide an explanation for U-shaped learning within a single-mechanism learning system. The network with selective attention, using both error and frequency information, masters the all of the past tense mappings,

learns to generalize to novel verbs, and shows U-shaped learning behavior qualitatively similar to children. It is important to reiterate that this is accomplished without external manipulation of the training set. The mechanism of selective attention, rather than external manipulation, is responsible for the network's ability to learn the entire training set *and* show U-shaped learning.

References

Baluja, S. & Pomerleau, D.A. (1994). Using a saliency map for active spatial selective attention: implementation & initial results. In Tesauro, Touretsky & Leen (Eds.), *Advances in Neural Information Processing Systems 7.* Cambridge, MA: The MIT Press.

Berko, J. (1958). The child's learning of English morphology. *Word, 14,* 150-177.

Broadbent, D.E. (1958). *Perception and communication.* London: Pergamon Press.

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

Cazden, C.B. (1968). The acquisition of noun and verb inflections. *Child Development, 39,* 433-448.

Dunham, P.J. (1990). Temporal structure of stimulation maintains infant attention. In J.T. Enns (Ed.), *The development of attention: Research and theory.* North Holland: Elsevier Science Publishers.

Elman, J.L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition 48,* 71-99.

Jackson, D., Constandse, R.M. & Cottrell, G.W. (1996). Selective attention in the acquisition of the past tense. *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (pp. 183-188). Hillsdale, NJ: Lawrence Erlbaum Associates.

Jusczyk, P.W., Cutler, A., & Redanz, L. (1993). Infants' sensitivity to predominant stress patterns in English. *Child Development, 64,* 675-687.

Jusczyk, P.W., Friederici, A.D., Wessels, J., Svenkerud, V.Y., & Jusczyk, A.M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language, 32,* 402-420.

Jusczyk, P.W., Luce, P.A. & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language, 33,* 630-645.

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

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

Kuczaj, S.A. (1978). Children's judgements of grammatical and ungrammatical irregular past tense verbs. *Child Development, 49,* 319-326.

Lachter, J., & Bever, T.G. (1988). The relation between linguistic structure and associative theories of language learning—a constructive critique of some connectionist learning models. *Cognition, 28,* 195-247.

MacWhinney, B. (1990). *The CHILDES Project: Computational Tools for Analyzing Talk (Version 0.88).* Pittsburgh, PA: Carnegie-Mellon University, Department of Psychology.

Marchman, V. (1988). Rules and regularities in the acquisition of the English past tense. *Center for Research in Language Newsletter, 2* (4).

Marcus, G.F., Ullman, M., Pinker, S., Hollander, M., Rosen, T.J., & Xu, F. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development, 57* (4), Serial No. 228.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition, 28,* 73-193.

Plaut, D.C., McClelland, J.L., Seidenberg, M.S., & Patterson, K.E. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103,* 56-115.

Plunkett, K., Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition 38,* 43-102.

Plunkett, K., Marchman, V. (1993). From Rote Learning to System Building: Acquiring Verb Morphology in Children and Connectionist Nets. *Cognition 48,* 21-69.

Plutowski, M., White, H. (1993). Selecting concise training sets from clean data. *IEEE Transactions on neural networks, 3*:1.

Plutowski, M., Cottrell G.W., White, H. (1993). Learning Mackey-Glass from 25 examples, plus or minus 2. In Hanson, Cowan and Giles (Eds.), *Advances in Neural Information Processing Systems 6.* San Mateo, CA: Morgan Kaufmann.

Rumelhart, D. E., McClelland, J.L. (1986). On Learning the Past Tense of English Verbs. In D.E. Rumelhart, J.L. McClelland & the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol 2.

Sokolov, E.N. (1960). Neuronal models and the orienting reflex. In M.A.B. Brazier (Ed.), *The central nervous system and behavior* (3rd conference, pp. 187-286). New York: Josiah Macy, Jr. Foundation.

Sokolov, E.N. (1963). *Perception and the conditioned reflex.* Oxford: Pergamon Press.

Sokolov, E.N. (1969). The modeling properties of the nervous system. In M. Cole & I. Maltzman (Eds.), *A handbook of contemporary Soviet psychology.* New York: Basic Books.