Rules or Connections? The Past Tense Revisited

Kim Daugherty
Hughes Aircraft Company
and
University of Southern California
kimd@gizmo.usc.edu

Mark S. Seidenberg
University of Southern California
marks@neuro.usc.edu

Abstract

We describe a connectionist model of the past tense that generates both regular and irregular past tense forms with good generalization. The model also exhibits frequency effects that have been taken as evidence for a past tense rule (Pinker, 1991) and consistency effects that are not predicted by rule-based accounts. Although not a complete account of the past tense, this work suggests that connectionist models may capture generalizations about linguistic phenomena that rule-based accounts miss.

Introduction

A seemingly minor aspect of linguistic knowledge—the past tense of verbs—has generated considerable debate over the role of connectionist models in explaining language. Whereas Rumelhart & McClelland (1986) claimed that their model of past tense learning illustrated a new way of explaining linguistic phenomena, Pinker & Prince's (1988) critique of this work suggested that connectionism had little to add to standard linguistic accounts. Several developments in subsequent years have moved this debate forward. Pinker (1991), for example, now agrees that connectionist networks are needed in order to account for facts about irregular past tenses (e.g., SING-SANG) and generalization. Moreover, the models developed by MacWhinney & Leinbach (1991), Plunkett & Marchman (1989), and Cottrell & Plunkett (1991) show that connectionist networks can exhibit many of the characteristics of the past tense that were lacking in the original Rumelhart & McClelland (1986) work. Thus, there has been considerable movement toward theories of the past tense in which connectionist networks play a central explanatory role. Importantly, however, Pinker (1991) and Kim et al. (1991) retain the idea that a proper account of the past tense will have to include a rule governing regular forms such as LIKE-LIKED. This conclusion is based on a mass of behavioral and other types of evidence thought to implicate this rule. Insofar as connectionist models do not, by definition, incorporate this type of knowledge representation, connectionism cannot provide a complete account of the past tense. Pinker therefore opts for a mixed model employing both a rule and a net.

It would be important to determine whether a connectionist network can explain all of the relevant facts about the past tense or whether, as Pinker suggests, it will have to be supplemented by a rule, since these represent very different claims about linguistic knowledge. Although we cannot present a complete account of the past tense in this paper, we will describe a simple connectionist model that learns the past tense and use it to develop two central points. The first is that the kinds of data that Pinker (1991) sees as evidence for a rule of past tense formation may reflect very simple properties of connectionist nets. As such, these phenomena cannot be taken as uniquely compatible with the rule-based account. This argument is mostly negative: it says that the kinds of evidence that Pinker takes to differentiate between network and rule-based accounts are consistent with both approaches. Our second point, however, is that the network exhibits behaviors that are not predicted by the traditional (Pinker & Prince, 1988) or modified traditional (Pinker, 1991) linguistic approaches. Moreover, these behaviors are also observed in psycholinguistic studies of people (Seidenberg, in press; Seidenberg & Bruck, 1990). Therefore, the network model is not merely an alternative to a rule-based account; rather, it is to be preferred because it captures generalizations that rule-based accounts miss.

Architecture of the Model

The model is a simple feed-forward network. The input layer represents the present tense of a monosyllabic verb; the output layer, its past tense. The phonological representation is similar to one used by MacWhinney et al. (1989). There were 120 phonological units on each layer, representing a CCCVVCCC template for monosyllables in English. This is the maximum structure a syllable can take; the double V is needed to represent dipthongs. Each phonemic segment is represented by 15 binary articulatory features, using a scheme developed by Plunkett and Prince and modified by Mary Hare. The features are: back, tense, low, medium, high, glide, sonorant, fricative, stop, labial, coronal, velar, nasal, sibilant, and voiced. This scheme represents a plausible

C C C V V C C C

b l a k

b a k

Figure 1: Phonological Representation of a Syllable

compromise among various proposals within phonetics. If a feature exists for a segment, its value is 1.0; if not, its value is 0.0.

The phonological representation is centered on the nucleus of the syllable as shown in Figure 1. Hence, the rimes of the words **BLACK** and **BACK** receive the same representation. Aligning the representations on the rimes was thought to be desirable because of the perceptual salience of these units in English (e.g., Pinker & Prince, 1988; Seidenberg & McClelland, 1989). For unused segments in a word representation, the units are set to 0.0. There were 200 hidden units. Figure 2 shows the architecture of the entire model.

The phonological form of a present tense stem (e.g., BAKE) is activated on the input units and the model's task is to generate the phonological form of the past tense (e.g., BAKED) on the output units (we use orthography here to represent these phonological codes for typographical convenience). The model was presented with present-past tense pairs during training, with frequency of exposure determined by the logarithm of the verb's Francis & Kucera (1982) frequency. Weight correction was determined by standard back-propagation (Rumelhart, Hinton, & Williams, 1986). In scoring the model's performance, we determined for each phonemic segment whether the best fit to the computed output was provided by the correct target. The output pattern was scored as correct only if the correct targets provided the best fit for all segments in a word. We also calculated the total sum of squared error as a measure of goodness of fit.

Simulation 1: Initial Corpus

Two simulations using the above architecture but different training sets were performed. In the first simulation, the training set consisted of all monosyllabic present-past tense pairs with Francis and Kucera

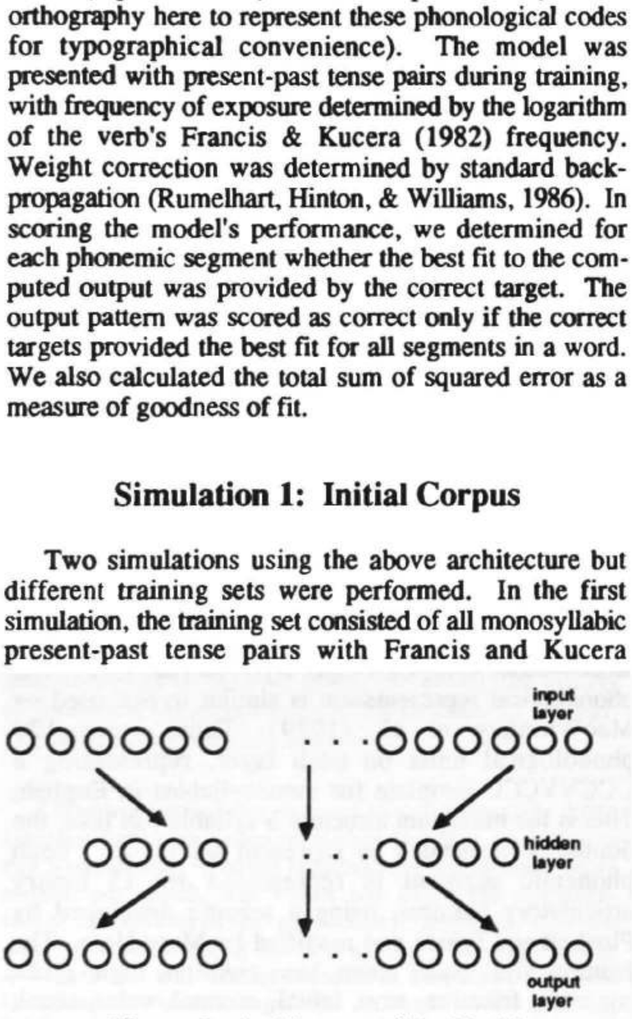

Figure 2: Architecture of the Model

frequencies greater than 1. This included 309 verbs with regular past tenses ("regular verbs") and 104 verbs with irregular past tenses. 112 regular verbs with frequency = 1 were reserved for testing the trained network's capacity to generalize on novel items. The present/past pairs were probabalistically presented during training according to their frequency. The most frequent pairs were presented once per epoch; the least frequent once per 100 epochs. The model was trained on this corpus for 400 epochs, at which point learning approached asymptote. The results below were averaged over three training sessions with random initial weights.

In terms of overall performance, the model learned all (100%) of the regular past tenses and 86/102 (84.3%) of the irregular forms. Errors on the irregulars included regularizations (FALL-FALLED), no change errors (GET-GET, analogous to HIT-HIT), and vowel errors (HIDE-HED).

After training, the model's performance on 112 additional verbs was used to assess its capacity to generalize. The model produced correct output for 84/112 (76.8%) of these items. The two most frequent errors were no change (PEEK-PEEK) and assimilation with phonologically-similar irregular past tenses in the training set (e.g., SEEP-SEPT, which is similar to SWEEP-SWEPT, and WRITHE-WROTHE, which is similar to WRITE-WROTE).

We then examined two theoretically-important phenomena, consistency and frequency effects. Consistency effects have been identified in previous work on spelling-sound correspondences and were simulated in the Seidenberg & McClelland (1989) model of word pronunciation. Briefly, networks trained using learning algorithms such as backpropagation (Rumelhart, Hinton, & Williams, 1986) pick up on the consistency of the mapping between input and output codes. The mapping between the present and past tenses is highly consistent in English because most past tenses obey the regular rule. However, the mapping is not entirely predictable because of irregular cases such as TAKE-TOOK and SIT-SAT. Standard accounts such as Pinker & Prince's (1988) distinguish between rule-governed cases and exceptions. However, connectionist models predict the existence of intermediate cases, so-called "regular but inconsistent" patterns such as BAKE-BAKED and FLIT-FLITTED, which obey the rule but have inconsistent "neighbors" (Seidenberg, in press). Thus, even though BAKE-BAKED is rule-governed, performance may be impaired because the model must also encode MAKE-MADE and TAKE-TOOK. Specifically, it should be worse than on a completely regular pattern such as LIKE-LIKED (all of the -IKE verbs have regular past tenses). Thus, the standard theory predicts that BAKE-BAKED should pattern with LIKE-LIKED, because both are rule governed. However, a backprop net might be expected to perform more poorly on inconsistent

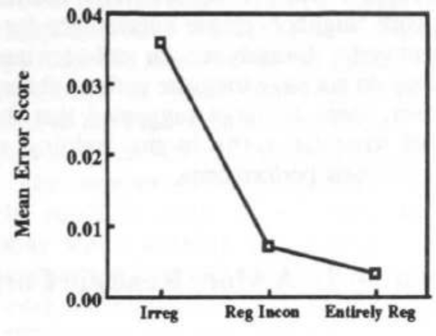

Figure 3: Performance on Matched Subsets of Items

items such as BAKE-BAKED, owing to TAKE-TOOK and MAKE-MADE.

To test this prediction, we identified sets of 60 entirely regular, 60 regular inconsistent, and 60 irregular verbs, equated in terms of frequency Performance on these words is shown in Figure 3. As predicted, the entirely regular verbs yielded better performance than inconsistent verbs, which in turn generated better performance than irregular verbs. Analogous results have been reported in the domain of spelling-sound correspondences (e.g., MUST is regular, HAVE is irregular, GAVE is regular but inconsistent) and simulated by the Seidenberg & McClelland (1989) model. Importantly, Seidenberg & Bruck (1990) and Seidenberg (in press) observed these effects in a study of past tense generation. The subjects in their experiment (college students) were presented with a present tense stem on each trial and had to generate the past tense. Response latencies followed the pattern illustrated in Figure 3: Irregular >> Regular Inconsistent > Entirely Regular. Thus, the model picks up on an important fact about peoples' knowledge of verbs: they encode the degree of consistency in the mapping between present and past tenses. It is important to recognize that these intermediate, regular-but-inconsistent cases are not predicted by either the traditional (Pinker & Prince, 1988) or modified traditional (Pinker, 1991) theories of the past tense. They are, however, a consequence of learning in a multi-layer backprop net. The reason for this is simple. The traditional accounts have the regular and irregular verbs processed by separate mechanisms. Hence there is no basis for predicting that they will interfere with each other. Our network, however, encodes both regular and irregular past tenses in the same set of connection weights. Hence the processing of a "rule-governed" item is affected by whether it has an irregular neighbor or not.

We also examined a second important phenomenon, frequency effects. Pinker (1991) has accumulated several types of evidence thought to support the existence of a past tense rule. The question that arises is whether such phenomena could also be captured by a connectionist net. This issue can be illustrated as follows. Prasada, Pinker & Snyder (1990) observed that the frequency of a past tense form (how often it is used in the language) affects the generation of irregular past tenses, but not regulars. TOOK, for example, is higher in frequency than RANG and takes longer for subjects to generate. However, there is no frequency effect for regular past tenses; BIKED (low frequency) is as easy to generate as LIKED (high frequency). Pinker (1991) interprets this pattern as follows. Regular past tenses are generated by rule; hence they are not affected by frequency. All that matters is how long it takes to recognize the present-tense stem and apply the rule. Irregular pasts are different, however. Either they have to be looked up in a list (traditional theory) or generated by a connectionist net (modified traditional theory). Both processes are thought to be affected by frequency. Thus, the interaction between frequency and regularity of the past tense was thought to implicate two separate mechanisms, a rule and a net.

We thought it likely, however, that our net would also produce this interaction, mainly because we observed the same effect in the Seidenberg & McClelland (1989) model. In that model, frequency has a bigger effect on words with irregular pronunciations (e.g., DEAF, SHOE) than words with regular, rule-governed pronunciations (LIKE, MALE). The explanation for this effect is also simple. Regular, "rule-governed" words contain patterns that occur repeatedly in the corpus. The weights reflect exposure to all these patterns. Mastering a rule-governed instance does not depend very much on its frequency because performance benefits from exposure to neighbors that contain the same pattern. Mastering an irregular instance, however, is highly sensitive to frequency; performance on DEAF (irregular pronunciation) or TAKE-TOOK (irregular past tense) depends on how often the model is exposed to these patterns because the correct output cannot be derived from exposure to neighbors. Thus, we expected that at least one of the behavioral phenomenon that Pinker takes as evidence for a rule would be exhibited by our model.

To test this prediction, we examined the 50 highest frequency regular verbs, the 50 lowest frequency

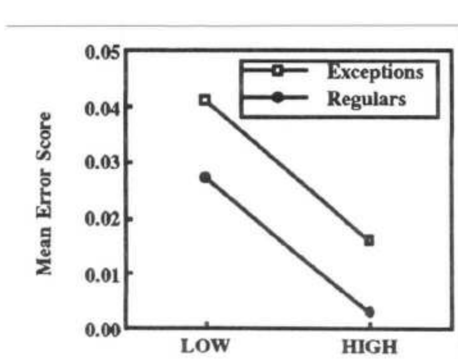

Figure 4: Frequency and Regularity Effects

regular verbs, the 50 highest frequency irregular verbs, and the 50 lowest frequency irregular verbs from the training set. The model's performance on these items is summarized in Figure 4. The results are not as predicted. There are main effects of verb type and frequency but no interaction between the two.

To summarize, the results of the initial simulation were mixed. The model was able to learn all of the regular items in the training set and a high proportion of the irregular items; it produced plausible errors and correct output on most generalization trials. It also showed the consistency effects seen in the behavioral studies, which are not predicted by traditional accounts. However, the model's performance is problematical in two respects. First, it performed more poorly than people on irregular items and generalization trials. Second, it did not exhibit the interaction between frequency and regularity.

At this point, we noticed several hints that the defects in the model's performance were principally due to the large number of irregular items in the training corpus. The model failed to master all of the irregulars in the training corpus, and produced unexpected frequency effects on regular items. Moreover, many of the errors on the generalization trials seemed to occur because the model was affected by phonologically-similar irregular past tenses in the training corpus. In order to assess the effects of the proportion of irregular items on performance, we conducted simulations in which we varied the number of irregular verbs in the training set, while keeping the number of regular verbs constant. The results are shown in Figure 5. As seen in the figure, the number of errors on generalization trials was related to the number of irregular verbs in the training set. The stimuli used in the generalization tests all require the regular past tense; however, some of them are entirely regular with respect to the training corpus whereas some are regular but inconsistent. Figure 6 presents the results for these two types of generalization trials separately. The number of errors on entirely regular novel verbs remained largely invariant as the number of irregular verbs was increased. However, the number of errors on regular inconsistent novel verbs was affected by the number of irregular verbs in the training set. This indicates that irregular verb neighbors create interference for regular inconsistent verbs. Entirely regular verbs are unaffected because they do not have irregular verb neighbors.

Together, these findings suggested that the large number of irregular verbs in the training set was adversely affecting performance.

Simulation 2: A More Realistic Corpus

We then compared the type and token frequencies in our corpus to those in the language at large. An analysis of the Francis & Kucera (1982) sample revealed that irregular verbs comprise 5% of all verb types listed there and 22% of the verb tokens. In the corpus employed in the first simulation, 25% of the verbs were irregular and they accounted for 65% of the tokens presented during training. Thus, irregular items were overrepresented in our training corpus compared to the language as a whole. Other factors also contributed to the overrepresentation of irregulars as well. The model's phonological representation only permitted the use of monosyllabic words, and the proportion of irregular verbs is higher among the monosyllabic words than the multisyllabic words (this is because most irregular verbs are monosyllabic). Finally, regular verbs predominate in the lower frequency range; the training corpus was restricted to items with frequency > 1, meaning that the many regular but very low frequency verbs were excluded. In sum, the large number of irregular items in the training corpus had a negative impact on the model's performance, but these words were overrepresented in the training corpus.

A new training set was then constructed with the goal of maintaining more realistic proportions of regular and irregular verbs. We also attempted to represent the different types of irregular verbs accurately. Pinker & Prince identified 25 classes and sub-classes of irregular verbs which we collapsed into 5 major classes reflecting the most important subtypes. The classes were no

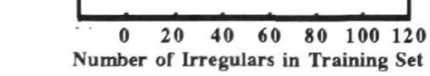

Figure 5: Generalization Errors vs. Number of Irregulars in Training Set

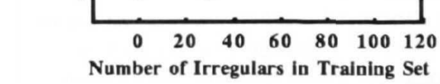

Figure 6: Generalization Errors vs. Number Irregulars in Training Set

change (HIT-HIT), vowel internal change (MEET -MET), vowel internal change plus consonant (LEAVE-LEFT), suppletion (GO-WENT), and consonant change (SEND-SENT). We then devised the training set so that the proportions of these subtypes matched the proportions in the Francis & Kucera corpus. The new training set included 309 regular verbs and 24 irregular verbs. The number of irregular verbs had to be relatively small in order to maintain the correct proportions while keeping the overall size of the training corpus within manageable limits.

The model was trained as in the previous simulation for 500 epochs, at which point performance approached asymptote. The following results reflect averages over three training sessions with random initial weights. All regular verbs in the training set were learned, as before. 22 of the 24 irregular verbs in the training set (92%) were learned, better than in the previous simulation. The 2 errors on irregular verbs were FALL-FELLED an overregulaization error and WIN-WAUN a vowel error.

On the generalization trials, 105 of 112 regular past tenses (94%) were correctly generated, an improvement over the first simulation and a rate that compares well with people. The 7 past tenses that were incorrectly generated were MERGE-MERGT, WAKE-WAKED, WHINE-WOOND, CLINK-CLANGT, WANE-WONE, MEW-VIEW, BROOK-BROOK. The first reflects a substitution of /t/ for /d/ (i.e., incorrect voicing) and the second is an overregularization error. The others are a variety of vowel, consonant and no change errors. As before, some of these can be described as assimilation with irregular verbs in the training set. WHINE-WHOUND is similar to WIND-WOUND, CLINK-CLANGT is similar to CLING-CLANG. Subjects in behavioral experiments produce some of these responses as well (e.g., Bybee & Moder, 1983).

We then re-examined the consistency and frequency effects described earlier. For the consistency effects, we constructed sets of 20 entirely regular, 20 regular inconsistent, and 20 irregular verbs from the training set equated in terms of frequency. The model's performance on the three types is given in Figure 7. As in the previous simulation, the model showed the graded

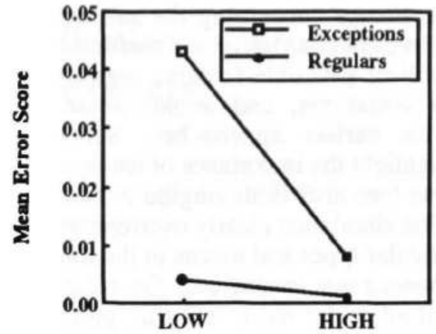

Figure 8: Frequency and Regularity Effects

effects of the consistency of the mapping between present and past tense (the difference between the two types of "regular" verbs) that is not predicted by rule-based accounts. The consistency effect was actually larger than in the previous simulation. In addition, the model now exhibits the predicted interaction of frequency and regularity. We constructed sets of the 10 highest frequency regular verbs, 10 lowest frequency regular verbs, 10 highest frequency irregular verbs, and 10 lowest frequency irregular verbs from the training set. The model's performance on these items is shown in Figure 8. For regular verbs, frequency has little effect on performance. For irregular verbs, performance is better on high frequency items than on low frequency items. This is the pattern that was reported by Pinker (1991) and Prasada et al. (1990) and taken as evidence for a rule-based mechanism.

Thus, changing the corpus so that it better reflected the facts about the distribution of verb types in the language yielded better simulation results. The model continued to master the regular items and over 90% of the irregulars; there was better generalization on novel verbs; and the consistency effect and frequency by regularity interaction seen in behavioral studies were obtained.

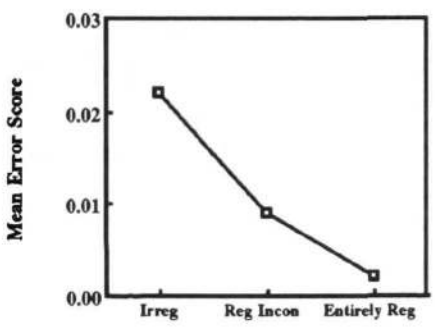

Figure 7: Performance on Matched Subsets of Items

Discussion

Our results, like those of MacWhinney & Leinbach (1991), Plunkett & Marchman (1989), and Cottrell & Plunkett (1991), suggest that connectionist models can exhibit the phenomena that Pinker & Prince (1988) see as central to an understanding of the past tense. We are in no way restricted to the mechanisms employed by Rumelhart & McClelland (1986) and are not doomed to the same failures. Two aspects of our models contributed to their relatively better performance. First, the phonological representation that we employed addresses many of the concerns that Pinker & Prince expressed concerning the Wickelphonology that Rumelhart & McClelland had used. Our representation of segments is motivated by articulatory constraints and the slot positions in the representation are motivated by independent evidence concerning the salience of the rime. This representation is by no means complete; however, it utilizes plausible featural, segmental, and syllabic representations, and avoids some of the problems with earlier approaches. Second, the simulations highlight the importance of using a realistic training regime (see also Hetherington & Seidenberg, 1989). Our first simulation clearly overrepresented the number of irregular types and tokens in the training set and its performance was inadequate. Once the training set was modified to be more realistic, performance improved greatly. This result suggests the interesting possibility that for English, the past tense is learnable only if the proportion of irregular pasts is limited.

By understanding the nature of the input representation, the learning algorithm, the phenomena we were trying to capture, and the architecture of the model, we were able to make predictions about the difficulty of learning different types of verbs. In the second simulation, we observed the expected consistency and frequency effects. These effects have also been observed in experiments with human subjects and replicate results that have been obtained in another domain (spelling-sound correspondences; Seidenberg & McClelland, 1989). The frequency effects (i.e. the fact that frequency only affects irregular pasts, not regular pasts) indicate that the kind of phenomena that Pinker (1991) cites as evidence for a rule may be simply captured within connectionist nets. Of course, it remains to be seen whether all of the phenomena he cites can be accommodated in the same way.

The consistency effects are not predicted by the earlier theories and strongly implicate the connectionist alternative. Of course, it might be possible to modify the traditional (Pinker & Prince, 1988) or modified traditional (Pinker, 1991) theories to accommodate these results. Pinker does not present implemented computational models; he describes models and assigns computational properties to them by stipulation. Working at this level of description, it might be possible to formulate a new theory that preserves the idea of a rule but accommodates the new behavioral phenomena. For example, one might introduce ad hoc assumptions about interactions between the rule and network mechanisms. If that were the case, however, the rule-based theory could be said to be useful insofar as it is able to "implement" the connectionist model. Thus, whereas Pinker & Prince (1988) presented a view in which connectionist accounts of language are parasitic on symbolic accounts, we view the present results as suggesting the opposite relation. Of course, it will be necessary to develop our models further, so as to account for the entire range of verb-related phenomena. We take the present results as one step along the route toward an explanatory computational theory.

References

Bybee, J.L. & Moder, C. (1983). Morphological classes as natural categories. *Language, 59*, 251-270.

Cottrell, G. & Plunkett, K. 1991. Using a recurrent net to learn the past tense. In Proceedings of the Cognitive Science Society Conference. Erlbaum.

Francis, W.N. & Kucera, H. 1982. *Frequency Analysis of English Usage*. Houghton-Mifflin.

Hetherington, P., & Seidenberg, M.S. 1989. Is there "catastrophic interference" in connectionist networks? In Proceedings of the Cognitive Science Society Conference. 26-33. Erlbaum.

Kim, J.J., Pinker, S., Prince, A. & Prasada, S. 1991. Why no mere mortal has ever flown out to center field. *Cognitive Science*, 15:73-218.

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

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

Pinker, S. 1991. Rules of language. *Science*, 253:530-534.

Pinker, S. & Prince, A. 1988. On language and connectionism. *Cognition*, 28:73-194.

Plunkett, K & Marchman, V. 1991. U-shaped learning and frequency effects in a multi-layered perceptron. *Cognition*, 39:43-102.

Prasada, S., Pinker, S., & Snyder, W. 1990. Some evidence that irregular forms are retrieved from memory but regular forms are rule generated. Psychonomic Society meeting, November.

Rumelhart, D., Hinton, G., & Williams, R.J. 1986. Learning internal representations by error propagation. In Rumelhart, D. & McClelland, J. (Eds.), *Parallel Distributed Processing, vol. 1*. MIT Press.

Rumelhart, D. & McClelland, J.L. 1986. On learning the past tenses of English verbs. In Rumelhart, D.E. & McClelland, J.L. (Eds.), *Parallel Distributed Processing, vol. 2*. MIT Press.

Seidenberg, M.(In press). Connectionism without tears. In S. Davis (Ed.), *Connectionism: Theory and Practice*. Oxford University Press.

Seidenberg, M., & Bruck, M. 1990. Frequency and consistency effects in the past tense. Psychonomic Society meeting, November.

Seidenberg, M.S. & McClelland, J.L. 1989. A distributed, developmental model of word recognition and naming. *Psychological Review*, 96:523-568.