Learning the past tense in a recurrent network: Acquiring the mapping from meaning to sounds

Garrison W. Cottrell*
Department of Computer Science and Engineering
University of California, San Diego

Kim Plunkett
Institute of Psychology
University of Aarhus, Denmark

Abstract

The performance of a recurrent neural network in mapping a set of plan vectors, representing verb semantics, to associated sequences of phonemes, representing the phonological structure of verb morphology, is investigated. Several semantic representations are explored in attempt to evaluate the role of verb synonymy and homophony in deteriming the patterns of error observed in the net's output performance. The model's performance offers several unexplored predictions for developmental profiles of young children acquiring English verb morphology.

Introduction

Prior attempts to model the acquisition of English verb morphology in connectionist nets (Plunkett & Marchman [1991]; Plunkett, Marchman & Knudsen [1990]; Rumelhart & McClelland [1986]) have focused on the problem of learning the relationships between phonological representations of various forms of the verb. Phonological information is exploited by children and adults when prompted for the past tense form of a novel stem (Bybee & Slobin [1982]; Marchman [1988]). Nevertheless, the phonological form of the verb does not uniquely *determine* its past tense form. Although all verbs which have identical stem and past tense forms possess a dental final consonant (e.g. *hit* → *hit*), not all verbs that end in a dental consonant have identical stem and past tense forms (Pinker & Prince [1988]). Furthermore, connectionist models that learn purely *intra*-level phonological mappings cannot distinguish verb-stem homophones that take different past tense forms. For example, *to brake* and *to break* take past tense forms *braked* and *broke* respectively. Since the inputs to the network in these cases are identical, so will their outputs remain identical.

In this paper, we present a connectionist model of the acquisition of English verb morphology in which a network is trained to map a semantic representation of verbs to phonological representations of their stems and/or past tense forms (cf. Gasser & Lee [1990]). The mapping function to be learnt may be considered analogous to aspects of the production process in which meaning is mapped to sound representations. Phonological regularities between verb classes must be captured by the hidden unit representations generated through training. Homophones in this formulation are unproblematic, as they consitute a many-to-one mapping. However, two new potential problems arise. First, as Pinker & Prince [1988] point out, the semantics of a verb is *not* a good predictor of the type of inflectional mapping that it must undergo. The three verbs *hit, strike* and *slap* are closely related semantically but they have different mapping types relating their stem and past tense forms (*hit* → *hit, strike* → *struck, slap* → *slapped*). The network must learn to *ignore* this similarity in learning the mapping. Second, this same problem arises in general for the network, insofar as there is an arbitrary relationship between the meaning of the verb and its phonological form. Similar inputs do *not* lead to similar outputs. In particular, highly similar inputs, modeling synonyms, provide a potential source of error in these networks.

Our goals in this work are twofold:

1. To examine the performance of the network in solving a mapping problem that is analogous to that of mapping meaning to sound and determine its generalization characteristics.

2. To evaluate the pattern of outputs and errors produced by the network during the course of training and use these errors to predict those produced by children acquiring English verb morphology.

We report on three sets of simulations that differ either in the nature of the semantic representations used to encode verb meanings and/or in the number of meaning/form pairs that the network is required to learn. In each case, we provide an evaluation of the performance of the network on trained verbs and of the ability of the network to generalize to verb forms on which it has not been trained. In one set of simulations, we provide a detailed error analysis.

*We thank Steen Ladegaard Knudsen for his assistance in programming, analysis and running of simulations.

Methodology

All simulations utilize a simple recurrent network of the type developed by Elman [1990] (see Figure 1). In all simulations the output phoneme consists of a

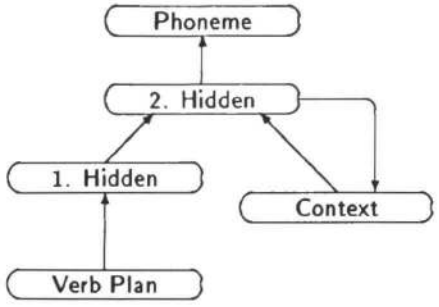

Figure 1: General Network Architecture

15 bit vector that reflects standard phonological contrasts. A noteworthy characteristic of this phonemic representation lies in its attempt capture the sonority relationships between vowels and consonants (see features O1–7 in Table 1).[1] The task of the network

				LO	MD	HI	GL	SN	FR	ST	PLACE					
		BK	TN	O7	O6	O5	O4	O3	O2	O1	LA	CR	VL	NS	SB	VO
vat	æ	−1	−1	+1	+1	+1	+1	+1	+1	+1	−1	−1	−1	−1	−1	+1
cut	ˆ	+1	−1	+1	+1	+1	+1	+1	+1	+1	−1	−1	−1	−1	−1	+1
ate	e	−1	+1	−1	+1	+1	+1	+1	+1	+1	−1	−1	−1	−1	−1	+1
bet	ɛ	−1	−1	−1	+1	+1	+1	+1	+1	+1	−1	−1	−1	−1	−1	+1
boat	o	+1	+1	−1	+1	+1	+1	+1	+1	+1	−1	−1	−1	−1	−1	+1
bought	O	+1	−1	−1	+1	+1	+1	+1	+1	+1	−1	−1	−1	−1	−1	+1
beat	i	−1	+1	−1	−1	+1	+1	+1	+1	+1	−1	−1	−1	−1	−1	+1
bit	I	−1	−1	−1	−1	+1	+1	+1	+1	+1	−1	−1	−1	−1	−1	+1
boot	u	+1	+1	−1	−1	+1	+1	+1	+1	+1	−1	−1	−1	−1	−1	+1
foot	U	+1	−1	−1	−1	+1	+1	+1	+1	+1	−1	−1	−1	−1	−1	+1
	y	−1	−1	−1	−1	−1	+1	+1	+1	+1	−1	−1	−1	−1	−1	+1
	w	+1	−1	−1	−1	−1	+1	+1	+1	+1	−1	−1	−1	−1	−1	+1
	h	−1	−1	−1	−1	−1	+1	+1	+1	+1	−1	−1	−1	−1	−1	−1
	r	−1	−1	−1	−1	−1	+1	+1	+1	+1	−1	+1	−1	−1	−1	+1
	l	−1	−1	−1	−1	−1	−1	+1	+1	+1	−1	+1	−1	−1	−1	+1
	m	−1	−1	−1	−1	−1	−1	+1	+1	+1	+1	−1	−1	+1	−1	+1
	n	−1	−1	−1	−1	−1	−1	+1	+1	+1	−1	+1	−1	+1	−1	+1
	ŋ	−1	−1	−1	−1	−1	−1	+1	+1	+1	−1	−1	+1	+1	−1	+1
	f	−1	−1	−1	−1	−1	−1	−1	+1	+1	+1	−1	−1	−1	−1	−1
	v	−1	−1	−1	−1	−1	−1	−1	+1	+1	+1	−1	−1	−1	−1	+1
	s	−1	−1	−1	−1	−1	−1	−1	+1	+1	−1	+1	−1	−1	+1	−1
	z	−1	−1	−1	−1	−1	−1	−1	+1	+1	−1	+1	−1	−1	+1	+1
theatre	θ	−1	−1	−1	−1	−1	−1	−1	+1	+1	−1	+1	−1	−1	−1	−1
mother	δ	−1	−1	−1	−1	−1	−1	−1	+1	+1	−1	+1	−1	−1	−1	+1
	p	−1	−1	−1	−1	−1	−1	−1	−1	+1	+1	−1	−1	−1	−1	−1
	b	−1	−1	−1	−1	−1	−1	−1	−1	+1	+1	−1	−1	−1	−1	+1
	t	−1	−1	−1	−1	−1	−1	−1	−1	+1	−1	+1	−1	−1	−1	−1
	d	−1	−1	−1	−1	−1	−1	−1	−1	+1	−1	+1	−1	−1	−1	+1
	k	−1	−1	−1	−1	−1	−1	−1	−1	+1	−1	−1	+1	−1	−1	−1
	g	−1	−1	−1	−1	−1	−1	−1	−1	+1	−1	−1	+1	−1	−1	+1
silence	-	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1

Table 1: Phonological representation

is to output a sequence of phonemes that correspond to the stem or past tense of the verb whose semantic representation is presented at the input. The distinction between stem and past tense forms is encoded by a 2-bit vector at the input level. The inventory of verbs (both stems and past tense forms) that the network is required to produce at the output is taken directly from previous simulations conducted by Plunkett, Marchman & Knudsen [1990]. In this work 500 stem/past tense pairings are used. Each stem 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. Verbs are assigned to one of four classes. Each class corresponds to a different type of transformation analogous to a distinct past tense form in English as in (Plunkett & Marchman [1991]). The first class follows the regular rule in English. The three irregular classes are: (1) ***arbitrary verbs*** have no systematic relationship between the stem and past tense form (*go* → *went*), (2) ***identity verbs*** are unchanged between forms (*hit* → *hit*), and (3) ***vowel change verbs*** undergo a change in vowel in CVC forms (*strike* → *struck*). Verbs are assigned randomly to each of the four classes, with the constraint that stems possess the appropriate characteristics of a given class.

Semantic representations of verbs are of two types. In the first set of simulations, each verb is represented in a localist manner in a 500-bit vector. An additional two units encode whether the network is to produce a sequence of phonemes corresponding to the stem of the verb or the past tense of the verb at the output. In the second and third set of simulations a similarity structure is imposed on the semantic representations by using distortions of several prototype vectors (Chauvin [1988]). Distortions may vary in their distance from the prototype. We use 9 or 50 prototype vectors (and thus as many categories) depending on vocabulary size. Two extra inputs specify the stem or the past tense form. For the large simulations, 10 distortions each of the 50 prototypes are generated at 3 levels of distortion.

Training proceeds by randomly selecting a plan vector (the verb's semantic representation) and a tense bit (stem or past tense). This composite vector is then presented at the input units over a number of time steps that correspond to the number of phonemes in the output form. At each time step, the discrepancy between the actual output of the network and the target phoneme is used as the error signal to a back propagation learning algorithm. We use the TLEARN simulator developed by Jeff Elman at UCSD. As part of the teaching signal, the verb plan is trained to produce an end-of-form signal (corresponding to the *silence* phoneme in Table 1). The "context units" are reset between forms.

Analysis

The performance of the network is analysed at regular intervals in training. In this paper we present two types of analysis. First, we determine the hit rate for stems and past tense form, both on the entire training

[1] The phonological representation presented here was originally designed by Alan Prince and Kim Plunkett.

set and on a class-by-class basis. Hit rate is evaluated by determining which of the phoneme vectors (as defined in Table 1) is closest to the output vector using a least squares measure at each time step. This yields a sequence of output phonemes for each verb plan. In the first two sets of simulations we report on whether the output sequence is a hit or a miss.

We analyse the generalization characteristics of the network by first training the network with 25 verb plans to produce only the stem form of the verb and with another 25 verb plans to produce only the past tense form of the verb. Each verb plan is then tested on the phonological form of the verb to which it has not been trained i.e. 25 stem forms and 25 past tense forms. The output of the network on these novel inputs is used to evaluate the net's generalization properties.

Finally, in the third set of simulations we provide a detailed analysis of the output of the network when trained on just 50 verb plans i.e. 100 phoneme sequences. These analyses relate the role of semantic similarity to the similarity of the phoneme sequences across different verbs, and the syllabic structure that the network extracts over the sequence of output phonemes.

Experiment One

This experiment reports the results of simulations using a 500 word vocabulary and orthogonal representations of the verb plan. Figure 2 (a) provides a summary of the network performance on all past tense forms and all stem forms while Figure 2 (b) compares the generalization characteristics of the network in predicting the past tense forms of the verb when it has only be trained on the stem and *vice versa*. Figure 2 (a) shows that the network is equally fast at learning both stem and past tense forms and that learning undergoes a spurt in growth around the 20 epoch mark. In contrast, the test verbs differ with respect to their performance on stems and past tense forms. Figure 2 (b) shows that when a verb plan is trained to a past tense form, the network is quite accurate in predicting the correct stem ($\geq$ 90% after 70 epochs of training). On the other hand, generalization from stem to past tense never exceeds 55%. It should be noted that we use a very strict criteria for generalization: All past tenses are assumed to be regular. Over several simulations, we find that performance on past to stem generalizations is always good, while stem to past varies. This result is to be expected given that the form which the past tense takes is a better predictor of the stem than the stem is of the past tense form (e.g. if the past tense of a verb is *talked* then its stem form is unambiguous, but if the stem is *hit* then, in principle, its past tense form is underdetermined). Indeed the discrepancy between the generalization curves in Figure 2 (b) can be accounted for in this fashion.

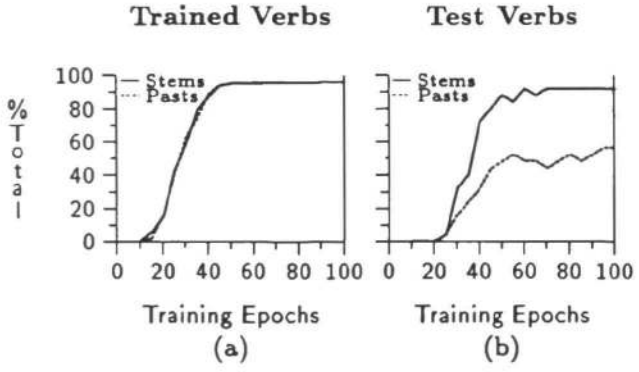

Figure 2: Overall Orthogonal Results

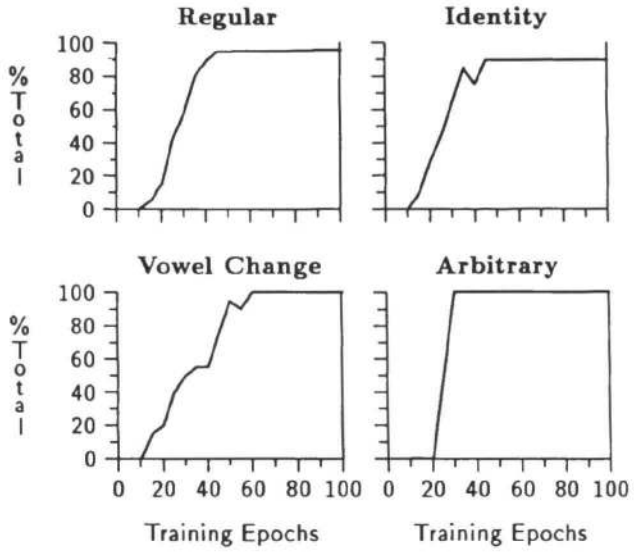

Figure 3: Stem Forms by Class (Orthogonal)

Figure 3 provides a class-by-class breakdown of network performance on stems. These result indicate that the regular, identity mapping and vowel change classes are learned first, while arbitrary mappings are delayed. Figure 4 reveals a similar rank ordering of classes with past tense forms.

Experiment Two

This experiment reports the results of simulations using a 500 word vocabulary and semantically structured representations of the verb plan. Figure 5 (a) provides a summary of the network performance on all past tense forms and all stem forms while Figure 5 (b) compares the generalization characteristics of the network in predicting the past tense forms of the verb when it has only be trained on the stem and *vice versa*. As with Experiment One, Figure 5 (a) shows that the network is equally fast at learning both stem and past tense forms. Learning undergoes a spurt in growth around the 25 epoch mark. Similarly, there is a contrast between the test stems and the test past tense forms. However, the generalization characteristics for

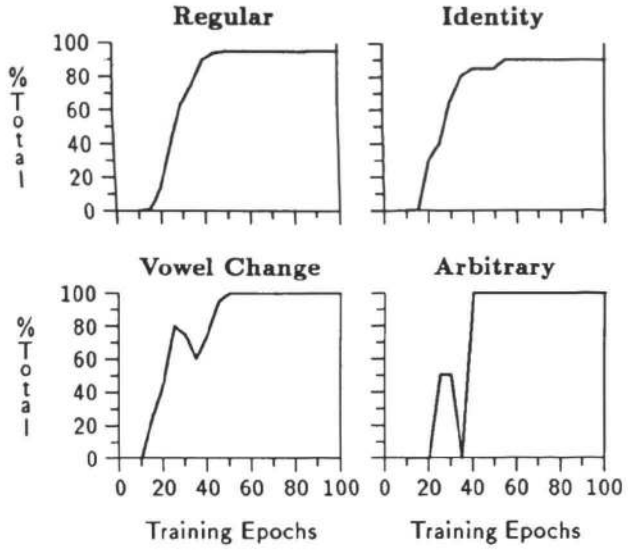

Figure 4: Past Forms by Class (Orthogonal)

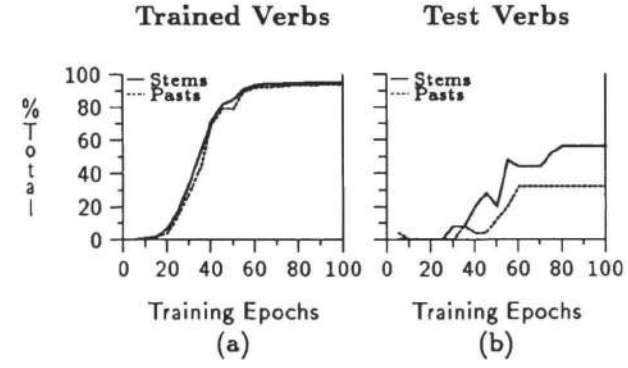

Figure 5: Overall Orthogonal Results

this network are more restricted than in the simulations where an orthogonal verb plan is used. An examination of the output for test stems indicates that the network has difficulty generating the epenthisized form of the *ed* suffix.

A class-by-class analysis of past tense forms and stem forms reveal results similar to the class-by-class analysis in Experiment One and so will not be reported here.

Experiment Three

This experiment reports the results of analysis of the effects of input and target similarity structure on the forms the network learns. The semantic classes in this experiment are designed to highlight these effects. In this experiment, we use a 54-stem subset (with one new arbitrary) of the larger set, with 31 regular forms, 3 arbitraries, 8 identities and 12 vowel-change verbs. We use 9 prototype vectors, where each exemplar within a class has the same amount of distortion from the prototype. Three of the classes use high distortion, three medium distortion, and three low distortion. Two out of three of these classes has one lexical item more frequently represented than the rest, a regular in one, and an arbitrary in the other.

We perform two kinds of analyses on this network:

1. We measure the changes in the stem output strings during learning from the point of view of the syllabic structure of the target language.
2. We measure the changes in similarity of stem and past output strings during learning with respect to the semantic clusters.

Syllabic structure changes

In order to assess the vocabulary development of the network, we divide the stem output strings of the network into three classes:

Words: Strings that belong to the target vocabulary.

Pseudo-words: Strings that are not in the target vocabulary but conform to the syllabic structure of the language — CVC, VCC, or CCV.

Non-words: Strings that do not fit the above criteria — CCC, CVV, VCV, VVC and VVV.

A graph of the numbers of unique forms of each kind over learning, along with the total number of unique forms, is shown in Figure 6. Interestingly , long before

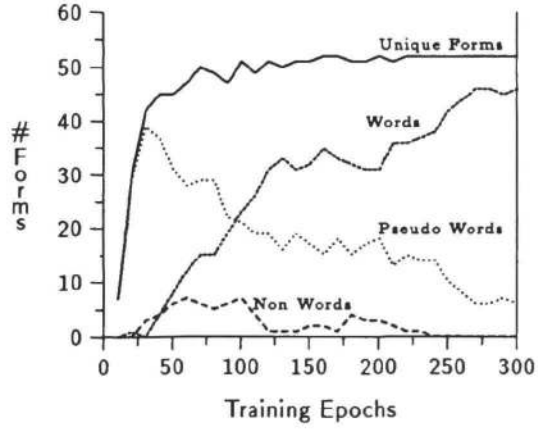

Figure 6: Vocabulary Structure

it has acquired any words in the language, the network has captured the syllabic structure, as evidenced by the high proportion of pseudowords in the set of unique forms. Over learning, there is an inverse relationship between these forms and the correct forms as the pseudo-words migrate into the target vocabulary. This set of curves is reminiscent of a similar set of curves found by Plunkett [1990] in a case study of language acquisition in two Danish children between the ages of 12 and 26 months. The total number of consistently produced and applied non-(adult) Danish phonological forms was inversely related to the number of Danish words over the period studied. That is, the children had their own vocabulary early in development that was eventually replaced by target forms.

Note that in the simulation, the number of non-words actually *increases* during the acquisition of the target vocabulary. Further analysis reveals a simple explanation for this effect. We assign each string of the network's stem outputs to one of 8 classes given by the possible combinations of {C∨V}, {C∨V} and {C∨V}. Three of these classes characterize the target language's syllabic structure which consists of 41 CVC, 7 CCV, and 5 VCC strings (there is one less than 54 due to a homophone). A plot of the number of network outputs that belong to each of these classes over training is shown in Figure 7. The network quickly learns

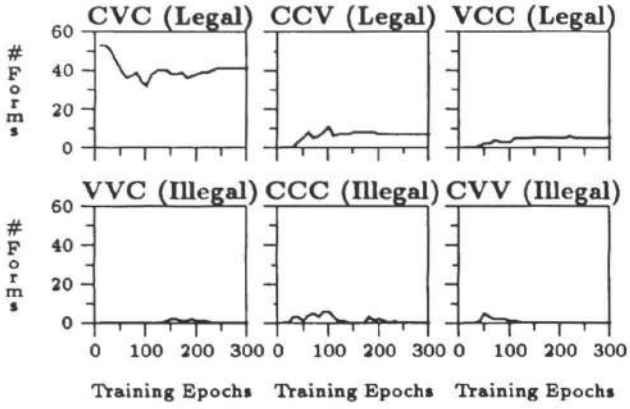

Figure 7: Syllabic Structures

the dominant CVC form of the language, overgenerating strings in this class initially. In order to extend to the target language syllabic structure these strings have to mutate from CVC to CCV and VCC during epochs 30 through 120. Strings changing from CVC to CCV have to change the mid-vowel to a consonant and the *coda* consonant to a vowel. The possible intermediate forms are CCC and CVV. Similarly for CVC to VCC, the possible intermediate forms are CCC and VVC. Indeed, a graph of the number of these forms produced by the network shows that they occur only during the cross-over from pseudo-words to words. The (logically possible) forms VVV and VCV never occur.

Similarity Effects

We hypothesized that the synonym groups would produce outputs that were more similar to one another than the non-synonym groups and the vocabulary as a whole. In order to test this hypothesis, we use the following measure of within-group similarity:

$$\text{Similarity}(\mathcal{G}) = 1 - \frac{1}{N(N-1)} \sum_{\substack{i,j \in \mathcal{G} \\ i \neq j}} \mathcal{RD}_{i,j}$$

where i, j range over the members of $\mathcal{G}$, and $\mathcal{G}$ is of size N. $\mathcal{RD}_{i,j}$ is a relative distance measure given by:

$$\mathcal{RD}_{i,j} = \frac{dist^{O}_{i,j}}{dist^{T}_{i,j} + 1}$$

This is the distance between the output phonemes in a string *relative* to the distance they should be after learning. $\mathcal{RD}_{i,j}$ should tend to 1 as the network learns, so Similarity($\mathcal{G}$) should tend to 0.

We apply this similarity measure to the strings produced by the network for each of the 9 prototype classes. For comparison purposes, we subtract the within-group similarity of the *total* output of the network from each score. Figure 8 (a) shows the average of this measure across the low-distortion (synonym) classes and the average across the high- and medium-distortion classes. The curves show that, in general,

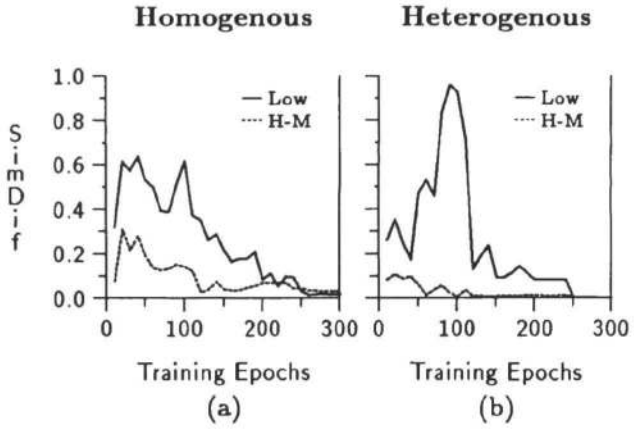

Figure 8: Synonym Analyses

semantic classes produce surface forms that are more cohesive than the forms of the network are as a whole. However, this effect disappears over training. This is the result of the network overcoming the false cue of input similarity. As expected, the synonym classes have higher within-group similarity throughout training than non-synonym classes. It is noteworthy that during the period that the network is actually acquiring the target vocabulary, i.e. from 70–120 epochs (cf. Figure 6), the within-group similarity of the synonyms *increases* compared to the rest of the classes. That is, synonyms are forced to be near-homonyms.

This "squeezing" effect is greatly magnified when there is only one synonym (Low distortion) cluster out of the nine classes (Figure 8 (b)). The explanation in terms of the network's organization is that patterns that are more easily learned (because they have less input similarity) are dominating the error gradient. The patterns may be characterized as *competing for representational resources* at the first hidden layer.

Examination of the network outputs over the training period reveals that the output strings for synonym classes during the steepest rate of target acquisition are within 2 or 3 features of one another. An interesting question here is: What is the string the outputs of a synonym class are pushed towards? Is it a blend of all of the strings of the class, or does one string in the class "capture" the output for that class? We tentatively find that when there is one lexical item that

is more frequent than the others, it captures the class. Interestingly, this is not the case if the more frequent item in a synonym class is an arbitrary verb, probably due to the fact that they map similar inputs to very different outputs (see Bartell, Cottrell & Elman, this volume for a thorough discussion of this issue). In the case that all are equally frequent, a blend of all of the outputs for the class is produced.

Taken in combination, these results suggest an unorthodox account of the source of the non-word forms found in Plunkett's subjects. These consistently-used pseudo-words are the result of two constraints or pressures on the child's language production: (1) A pressure to produce forms that are in keeping with the syllabic structure of the language at the output level, and (2) a pressure to produce similar forms based on input similarity. The child is thus producing the best approximation to a word in the language that is a blend of *all* of the words for that semantic class, with a tendency for this blend to be similar to the most frequent element of that class. A second counter-intuitive prediction of this work is that, during acquisition of the correct forms, the child will produce strings that may be inappropriate for the target language because they are *between* a common (over-acquired) form and a less common form.

Conclusions

We have described a connectionist model of morphology acquisition in which input forms representing the semantics of words are mapped to sequences of outputs representing their phonological forms. The network is successful in producing appropriate forms, even in the case where the input forms have a similarity stucture that is independent of the output similarity structure. Furthermore, the learning curves indicate a spurt-like acquisition profile. There is ample evidence for the spurt-like nature of vocabulary growth (McShane [1979]). It is unclear whether the acquisition of inflectional morphology in children shows a similar non-linear growth to that observed in the network.

The network experiences difficulty in generalizing from the stems to past tense forms. The model predicts that children are better at generalising from past tense forms to stems than *vice versa*. Further analysis is needed to investigate what modifications must be made to the model in order to achieve good generalization in the structured input case.

The analysis of the influence of input and target similarity on the acquisition of phonological form suggests some radical predictions. Children's non-adult forms may be a result of blending words for the same category. Looked at another way, words are distorted by their neighbors in a semantic class. The effects of similarity at the phonological level suggest that children will produce forms that do not belong to the syllabic structure of their language if these forms are between the most common form in the language and other forms. Finally, the model suggests that during the vocabulary burst, synonyms will be forced to be homonyms.

One problem that we have avoided addressing in this work, and suggested by other research (McClelland, personal communication) is that such models have difficulty learning uneven length strings. We plan to investigate ways to overcome this limitation in future research.

References

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

Chauvin, Y. [1988], "Symbol Acquisition in Humans and Neural (PDP) Networks," University of California, San Diego, PhD Thesis.

Elman, J. L. [1990], "Finding structure in time," *Cognitive Science* 14, 179–211.

Gasser, M. & Lee, C. [1990], "Networks that Learn about Phonological Feature Persistence," *Connection Science* 2, 265–278.

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

McShane, J. [1979], "The development of naming," *Linguistics* 17, 879–905.

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], "The segmentation problem in early language acquisition," CRL, University of California, Center for Research in Language Newsletter, San Diego.

Plunkett, K. & Marchman, V. [1991], "U-Shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition," *Cognition* 38, 1–60.

Plunkett, K., Marchman, V. & Knudsen, S. L. [1990], "From Rote Learning to System Building: Acquiring Verb Morphology in Children and Connectionist Nets," in *Proceedings of the 1990 Connectionist Models Summer School*, D. S. Touretzky, J. L. Elman, T. J. Sejnowski & G. E. Hinton, eds., Morgan Kaufmann, San Mateo, CA.

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.