Shultz, T. R., & Bale, A. C. (2006). Neural networks discover a near-identity relation to distinguish simple syntactic forms. *Minds and Machines, 16*, 107-139. The original publication is available at www.springerlink.com

Neural networks discover a near-identity relation to distinguish simple syntactic forms

Thomas R. Shultz

Department of Psychology and School of Computer Science

and

Alan C. Bale

Department of Linguistics

McGill University

Abstract Computer simulations show that an unstructured neural-network model [Shultz, T. R., & Bale, A. C. (2001). *Infancy, 2*, 501–536] covers the essential features of infant learning of simple grammars in an artificial language [Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). *Science, 283*, 77–80], and generalizes to examples both outside and inside of the range of training sentences. Knowledge-representation analyses confirm that these networks discover that duplicate words in the sentences are nearly identical and that they use this near-identity relation to distinguish sentences that are consistent or inconsistent with a familiar grammar. Recent simulations that were claimed to show that this model did not really learn these grammars [Vilcu, M., & Hadley, R. F. (2005). *Minds and Machines, 15*, 359–382] confounded syntactic types with speech sounds and did not perform standard statistical tests of results.

Keywords artificial grammars, cascade-correlation, connectionism, generalization, neural networks, representation, sonority, syllables

Author e-mail: thomas.shultz@mcgill.ca

Address: Thomas Shultz, Department of Psychology, McGill University, 1205 Penfield Ave., Montreal, Quebec, Canada H3A 1B1

Introduction

One of the fundamental debates in cognitive science over the last 20 years concerns the proper theoretical account of human cognition. Is cognition better interpreted in terms of symbolic rules or subsymbolic neural networks? A study of infant familiarization to sentences in an artificial language claimed to have produced data that only a rule-based account could explain (Marcus, Vijayan, Rao, & Vishton, 1999). Those results showed that 7-month-old infants attend longer to sentences with an unfamiliar syntax than to sentences with a familiar syntax. Learning of such artificial languages is thought to reveal the language-learning capabilities of infants (Gómez & Gerken, 2000).

Partly because of their own unsuccessful attempts at modeling these data with neural networks, Marcus et al. (1999) concluded that infants possess a rule-learning capability unavailable to neural networks that do not employ symbolic variables and rules. A

background, companion article suggested that rule learning might be an innately provided capacity, distinct from the associative learning mechanisms in neural networks (Pinker, 1999).

The challenge laid down by Marcus et al. and Pinker was promptly accepted by a number of neural-network modelers, many of whom produced neural models that seemed to capture the infant data (cf. review by Shultz & Bale, 2001). However, recent papers by Vilcu and Hadley (2001, 2003, 2005) reported, somewhat surprisingly, that several of these simulation results could not be replicated. The one simulation they replicated was by Shultz and Bale (2001). Although they acknowledged that this model could cover the infant data, they argued that it was learning numerical contours of the artificial sentences rather than the underlying grammatical pattern. In support of their claim, they noted that their own extensions of the Shultz and Bale model failed to generalize, both in terms of interpolation and extrapolation.

The purpose of the present paper is to examine Vilcu and Hadley's argument in detail. We present nine new simulations and three new analyses to show that Shultz and Bale's (2001) model does in fact learn to discriminate the simple grammars used in the Marcus et al. (1999) infant experiments. The paper begins with a brief review of the infant experiments and the original Shultz and Bale simulations. The contours involved in the sentences learned by the infants and the network models are then analyzed, and a new simulation explains cross-experiment differences based on contour patterns. Then the ability of the networks to generalize within and beyond the range of training stimuli is examined in additional simulations. This is followed by simulations and analyses of the knowledge representations learned by the networks. Finally, other new simulations explore the role of sonority contours in identifying syllables in continuous speech. It is concluded that contour learning and grammar learning are not incompatible, that learning sound contours can help in learning grammars, and that the Shultz and Bale (2001) model still offers a viable account of the infant data.

The Infant Data

Marcus et al. (1999) performed experiments in which 7-month-old infants were familiarized with three-word sentences of monosyllabic words in an artificial language and were then tested on novel sentences that were either consistent or inconsistent with those to which the infant was familiarized. In one experiment, infants were presented with sentences exhibiting an ABA pattern, for example, *ga ti ga* or *li na li*. There were 16 of these ABA sentences, created by combining four A-category words (*ga*, *li*, *ni*, and *ta*) with four B-category words (*ti*, *na*, *gi*, and *la*). Then the infants were presented with two novel sentences that were consistent with the ABA pattern (*wo fe wo*, and *de ko de*) and two novel sentences that were inconsistent because they followed an ABB pattern (*wo fe fe*, and *de ko ko*). In a control condition, infants were familiarized with sentences having an ABB pattern, e.g., *ga ti ti* and *ga na na*. Again, 16 such sentences were created by combining the four A-category words with the four B-category words. The test sentences were the same in this second condition, although here the novel ABB sentences were consistent and the novel ABA sentences were inconsistent with the familiar ABB pattern.

The dependent measure in these studies was looking time. During the test phase, if an infant looked at a flashing light to her left or right, a test sentence was played from a speaker near the light. A test sentence was played for 15 s or until the infant looked away.

A second experiment had the same structure except that the words were chosen more carefully so that phoneme sequences were more different in the familiarization and test patterns. A third experiment used the same words as did Experiment 2, but had contrastive syntactic patterns that each duplicated a consecutive word, i.e., AAB vs. ABB. This was to rule out the possibility that infants might have used the presence or absence of consecutively-duplicated words to distinguish syntactic types. Here both syntactic types had consecutively-duplicated words.

In all three experiments, infants attended more to inconsistent than to consistent novel sentences, suggesting that they had learned something about these simple grammars. The issue that has perplexed the literature ever since concerns the proper theoretical account of this syntactic processing. Is this processing based on rules and variables or on the mechanisms employed in neural networks, i.e., unit activations and connection weights? Marcus et al. (1999) argued that even these simple grammars could not be learned by a computational system that did not employ rules and variables, such as *If the first word of a sentence matches the third word of a sentence, then this sentence is grammatical*. In this rule, *sentence*, *first-word*, and *third-word* are variables that can be bound to different instances. Computational models by Shultz (1999) and Shultz and Bale (2001), among others, showed that unstructured neural networks could cover most features of the infant data. In this context, *unstructured* neural networks are those that are not engineered to explicitly employ rules and variables. Unstructured networks pass activation signals from unit to unit and modify connection weights, and some even grow new units and weights, but they do not bind values to variables within symbolic rules.

Cascade-correlation Simulations

The models of Shultz (1999) and Shultz and Bale (2001) both used an encoder version of the cascade-correlation (CC) learning algorithm. CC is a constructive algorithm for learning from examples in feed-forward neural networks (Fahlman & Lebiere, 1990; Shultz, 2003). As with other constructive algorithms, CC builds its own network topology as it learns. It does this by recruiting new hidden units as needed, thus searching in network-topology space as well as in connection-weight space for a solution to the problem on which it is being trained. Network-topology space is the space of possible network topologies, a space ordinarily searched by modelers themselves as they design static (unchanging) network topologies. Connection-weight space is the space of possible patterns of network weights, ordinarily searched automatically by a learning algorithm except in the case of programmer-designed weights.

Unlike the more standard, back-propagation networks with designed and static topologies, CC networks grow as they learn. They grow during what are called *input phases* by recruiting new hidden units into the network as needed to reduce error. New hidden units are recruited one at a time and installed each on a separate layer, receiving input from the input units and from any existing hidden units. The candidate hidden unit that is recruited is the one whose activations correlate most highly with the current error of the network. After recruiting a new hidden unit, the algorithm returns to an *output*

phase in which weights feeding the output units are adjusted to reduce error. CC has been used to simulate many aspects of cognitive development (Shultz, 2003).

Encoder networks are feedforward networks whose task is to learn to reproduce their inputs on their output units. Discrepancy between inputs and outputs is network error, which the learning algorithm attempts to reduce. Such encoder networks are particularly well suited for simulating familiarization experiments. Just as infants are imagined to build up a model of stimuli to which they are being exposed, and start to attend to more novel stimuli that deviate from their existing models, so do encoder networks build a model of the stimuli to which they are exposed. Network error can be taken as an index of stimulus novelty and interest.

Ordinary CC networks have many cross connections that bypass hidden layers. An encoder option within CC freezes direct input-output connections at zero in order to prevent trivial solutions in which weights of about 1.0 are learned between each input unit and a corresponding output unit (Shultz, 1999). Such trivial solutions can solve an encoder problem very quickly in the sense of error-free performance. However, they tend not to develop knowledge representations that could enable completion of partial patterns or generalization to similar, but novel patterns.

In the Shultz (1999) model, the A-category training syllables were coded as the real numbers 1, 3, 5, and 7, while B-category training syllables were coded as 2, 4, 6, and 8. The test syllables were coded with the interpolated values of 2.5, 3.5, 5.5, and 6.5. These networks simulated the consistency effect found with infants, and also generalized well, both inside and outside of the range of training sentences.

The coding scheme was considerably more realistic in the Shultz and Bale (2001) simulation. Coding there employed a continuous sonority scale inspired by phonological research (Vroomen, van den Bosch, & de Gelder, 1998). Sonority is the quality of vowel likeness and can be defined by openness of the vocal tract during speech production (Selkirk, 1984). The coding scheme for phonemes in single-syllable words in the infant experiments is presented in Table 1. The precise sonority numbers are somewhat arbitrary, but their ordering is based on research by Vroomen et al. (1998), who in turn based their sonority scale on Selkirk (1984).

Table 1 Phoneme sonority scale used in the Shultz and Bale (2001) simulations[a]

Phoneme category	Examples	Sonority
Low vowels	/a/ /æ/	6
Mid vowels	/ɛ/ /e/ /o/ /ɔ/	5
High vowels	/I/ /i/ /U/ /u/	4
Semi-vowels and laterals	/w/ /y/ /l/	-1
Nasals	/n/ /m/ / ŋ /	-2
Voiced fricatives	/z/ /ʒ/ /v/	-3
Voiceless fricatives	/s/ /ʃ/ /f/	-4
Voiced stops	/b/ /d/ /g/	-5
Voiceless stops	/p/ /t/ /k/	-6

[a]Example phonemes are represented in International Phonetic Alphabet. From "Infant familiarization to artificial sentences: Rule-like behavior without explicit rules and

variables." By T. R. Shultz and A. C. Bale. In L. R. Gleitman & A. K. Joshi (Eds.), Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society (p. 461), 2000. Mahwah, NJ: Erlbaum. Copyright 2000 by the Cognitive Science Society, Inc. Adapted by permission.

Sonorities ranged from -6 to 6 in steps of 1, with a substantial gap and change of sign between the principal categories of consonants and vowels. Each word was coded on two input units for the sonority of its consonant and the sonority of its vowel. For example, the sentence *ga ti ga* was coded as (-5 6 -6 4 -5 6).

Both CC models, whether using arbitrary (Shultz, 1999) or sonority (Shultz & Bale, 2001) coding of phonemes, captured the essential features of the infant data (Marcus et al., 1999) including:

1. exponential decreases in attention to a repeated sentence pattern,
2. more interest in sentences inconsistent with the familiar pattern than in sentences consistent with the familiar pattern,
3. occasional familiarity preferences,
4. more recovery to consistent novel sentences than to familiarized sentences,
5. and generalization both outside and inside the range of the training patterns.

Sonority Contours

Vilcu and Hadley (2003, 2005) argued that these CC models only learn numerical contours and not grammatical relations. As an example of what they mean by numerical contours, Vilcu and Hadley referred to the peaks or valleys formed by ABA sentences, depending on whether the B-category words have higher (peak) or lower (valley) values than the A-category words.

However as shown in Figure 1, plots of the raw sonority scores used as network inputs by Shultz and Bale (2001) reveal a sawtooth-shaped pattern reflecting the fact that vowels have higher sonorities than consonants. This difference between vowels and consonants is, of course, built into the sonority scale and does not by itself help a learning system to understand the grammatical nature of the sentences.

The contours of most relevance to CC networks were identified in Shultz and Bale's (2001) extensive knowledge-representation analysis of their networks. Shultz and Bale discovered via principle-component analyses of network contributions (sending activations x connection weights entering output units) that networks learned to represent on their hidden units sums (or equivalently, differences) of the consonant and vowel sonority values represented in the input. Further evidence for this claim is presented in detail in Simulation 8 and the subsequent section. For now, we focus on the shapes of sonority contours of sounds used in the infant experiments of Marcus et al. (1999).

Plots of sonority sums (sums of consonant and vowel sonority values) of training sentences in the three experiments, corresponding to a network's hidden-unit representations, reveal a mixture of sonority contours in each condition. As illustrated in Figure 2, for example, in the ABA condition of Experiment 1, there is a not a single contour to learn, but rather an equal mixture of peaks and valleys. The consistent test

patterns follow a subset of these trained contours, whereas the inconsistent test patterns (e.g., ABB) violate them.

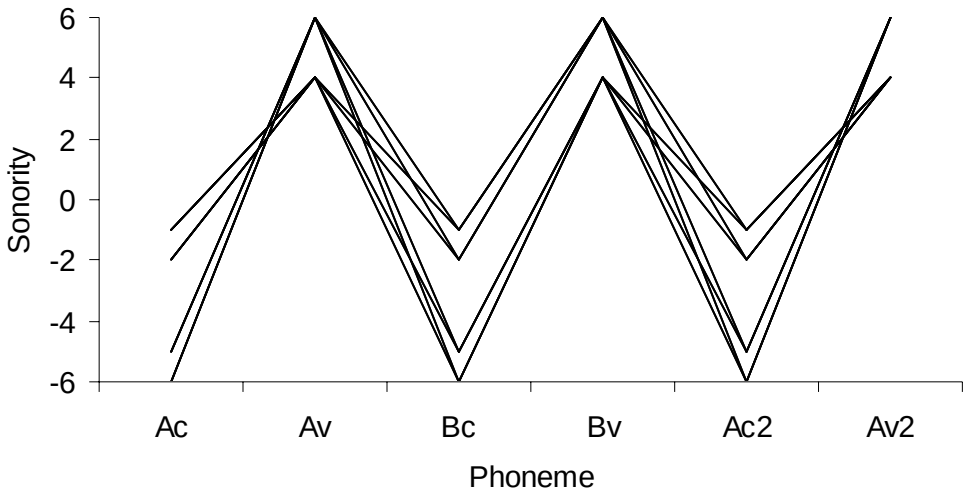

Figure 1. Sonority contours of the training sentences in the ABA condition of Experiment 1 from Shultz and Bale (2001). On the x-axis, *A* refers to the category-A words, *B* to the category-B words. Consonants are identified by *c*, vowels by *v*. The number *2* refers to the second occurrence of a phoneme. Each of the 16 sentences is represented by a line tracing the sonority contour from the consonant phoneme of the A syllable through to the second occurrence of the vowel phoneme of the duplicate A syllable. There is considerable overlap in these contour traces.

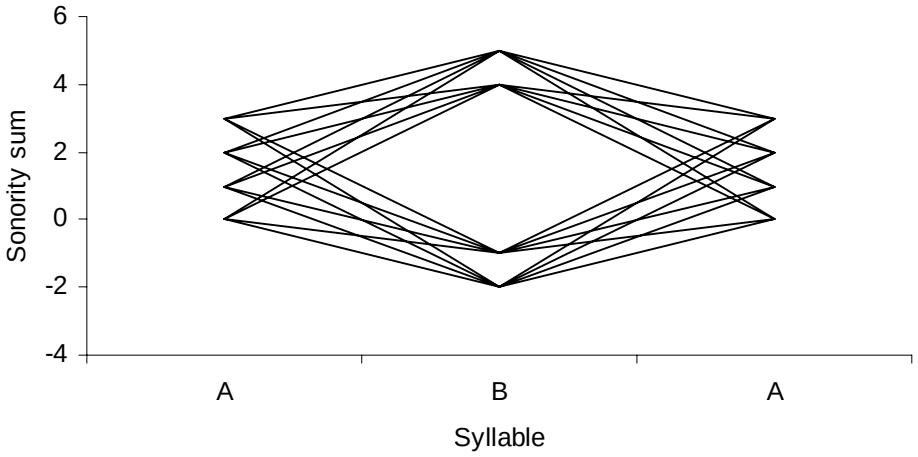

Figure 2. Sonority-sum contours of the training sentences in the ABA condition of Experiment 1 from Shultz and Bale (2001). These sums, of consonant sonority plus vowel sonority, correspond to a network's hidden-unit representations. Each of the 16 sentences is represented by a line tracing the sonority-sum contour from the initial A syllable to the B syllable to the duplicated A syllable.

Sonority-sum contours in these experiments take a variety of shapes in addition to peaks and valleys. These additional contour shapes are illustrated in Figure 3, which plots sonority-sum contours for 3 of the 16 training sentences in the ABB condition of Experiment 2. An *increasing* contour reflects a lower sonority sum for the A word than the B word. A *decreasing* contour reflects a higher sonority sum for the A word than the B word. A *plateau* contour signifies a flat shape in which the sonority sum for the three words is identical.

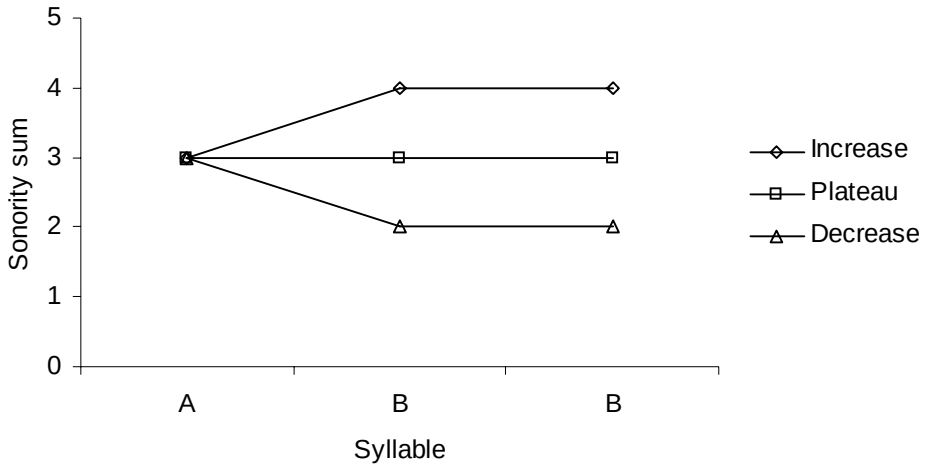

Figure 3. Illustrative sonority-sum contours for 3 of 16 training sentences in the ABB condition of Experiment 2. Each of the three sentences is represented by a line tracing the sonority-sum contour from the initial A syllable to the B syllables.

The frequencies of the various contour shapes of sonority sums in the training patterns of all conditions of the three experiments are listed in Table 2. Note that there are two plateau patterns in the sonority sums for each condition of Experiments 2 and 3. Such plateaus identify sentences in which the A and B categories cannot be distinguished merely on the basis of sonority sums or differences. This might make the training patterns in Experiments 2 and 3 more difficult to learn than those in Experiment 1, as indeed they appeared to be in Marcus et al.'s (1999) infant experiments, as measured by the mean difference in looking time between inconsistent and consistent test sentences: 2.7 for Experiment 1 vs. 1.75 and 2.1 for Experiments 2 and 3, respectively.

Table 2 Number of training sentences of various contours

Experiment	Familiarization condition	Contour of sonority sums				
		Peak	Valley	Increase	Decrease	Plateau
1	ABA	8	8	-	-	-
	ABB	-	-	8	8	-
2	ABA	7	7	-	-	2
	ABB	-	-	7	7	2
3	AAB	-	-	7	7	2
	ABB	-	-	7	7	2

Simulation 1: Cross-experiment differences

This cross-experiment prediction was explicitly tested in a full replication of the Shultz and Bale (2001) simulations. Table 3 presents mean network error to consistent and inconsistent test patterns in this replication. For each experiment, there was again a significant main effect of test-pattern consistency, with more error to the inconsistent patterns than to the consistent patterns. This means that there have now been three published replications of the Shultz and Bale networks capturing the essential main effect of test-pattern consistency.

Table 3 Mean network error to consistent and inconsistent test patterns in Simulation 1

Experiment	Consistent	Inconsistent	$F(1, 14)$	$p <$
1	10.17	13.61	20	.0001
2	14.81	16.01	12	.005
3	14.09	15.60	18	.001

To test the prediction that networks would have relatively more difficulty learning the training sentences in Experiments 2 and 3, the epochs required to learn were subjected to an ANOVA in which experiment was the independent factor. There was a main effect of experiment, $F(2, 45) = 12.05$, $p < .001$. See Figure 4 for a plot of the mean epochs to learn. As expected from the distribution of plateau-shaped training patterns, networks in Experiment 1, with no plateau patterns, learned faster than did networks in either Experiment 2, $t(30) = 6.03$, $p < .001$, or Experiment 3, $t(30) = 2.85$, $p < .01$. Networks in Experiments 2 and 3, each with two plateau training patterns, learned equally fast, $t(30) = 1.58$, *ns*.

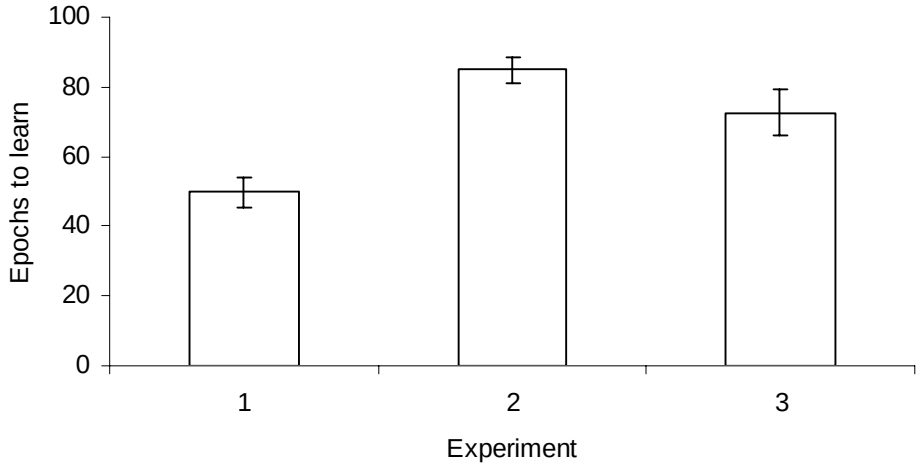

Figure 4. Mean epochs to learn in each experiment of Simulation 1, with standard-error bars.

Importantly, these differences in plateau-induced learning difficulty have implications for the size of the test-pattern consistency effect. A clear way to see this is to analyze the differences in test-pattern error, computed as error to inconsistent test patterns minus error to consistent test patterns. A one-way factorial ANOVA of these difference scores yielded an effect of the experiment factor, $F(2, 45) = 4.38$, $p < .05$. The mean difference scores, plotted in Figure 5, show a larger consistency effect in Experiment 1 than in either

Experiment 2, $t(30) = 2.39$, $p < .05$, or Experiment 3, $t(30) = 2.07$, $p < .05$. In contrast, mean difference scores did not differ between Experiments 2 and 3, $t(30) = 0.575$, *ns*.

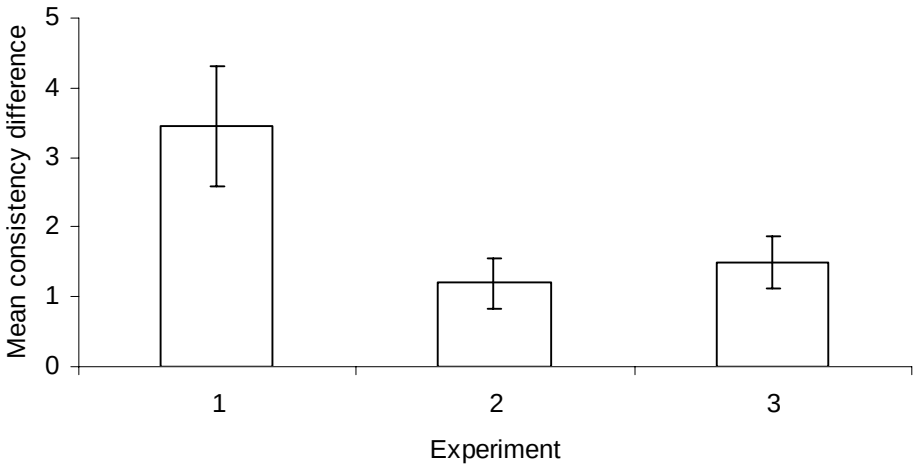

Figure 5. Mean consistency differences in each experiment of Simulation 1, with standard-error bars.

None of the test sentences in any experiment had plateau patterns, so the diminished discriminability of test sentences is probably due to plateau patterns in the training sentences. More precise coding of individual phonemic features, or a finer sonority scale, could eliminate this problem, but such remedies might also eliminate coverage of the phenomenon that infants showed less discrimination between consistent and inconsistent test patterns in Experiments 2 and 3. This coverage by the networks provides another measure of support for the realistic nature of the Shultz and Bale (2001) sonority coding scheme and for the model itself. A rule-and-variable system that ignored sonority contours (the sort of model apparently favored by Marcus et al. (1999) and Vilcu and Hadley (2005)) would presumably predict no difference between experiments.

Peaks and valleys in sonority contours

As noted earlier, Vilcu and Hadley (2003, 2005) felt that the neural models by Shultz and Bale (2001) and Shultz (1999) learn only numerical contours and not grammatical relations. In two of their experiments, Vilcu and Hadley revealed what they mean by "numerical contours", namely the peaks and valleys formed by ABA sentences. One of these simulations was an extension of the Shultz (1999) model. In that model, the A-category training syllables were arbitrarily coded as 1, 3, 5, and 7, while B-category training syllables were coded as 2, 4, 6, and 8. The test syllables were coded with the interpolated values of 2.5, 3.5, 5.5, and 6.5. Not only did these networks cover the consistency effect found with infants, but they generalized well, both inside and outside of the training range (Shultz, 1999). The coding of training sentences was balanced in the sense that half of the ABA training sentences involved peaks (e.g., 1 3 1) and half involved valleys (e.g., 5 3 5). Thus the Shultz (1999) networks had to do more than merely learn numerical contours, just as the Shultz and Bale (2001) sonority-coded networks did.

Vilcu and Hadley (2003, 2005) distorted this experimental balance in their extension simulations by always selecting a B-category word with a higher numerical code than the A-category words in ABA sentences. Thus, their ABA sentences always formed a peak contour. Then they formed two sets of consistent and inconsistent test sentences, one of which had peak-shaped ABA sentences and the other of which had valley-shaped ABA sentences. With peak-shaped ABA test sentences, there was less error to consistent than to inconsistent sentences, replicating the Shultz (1999) networks and Marcus et al.'s (1999) infants. But with valley-shaped ABA test sentences, there was more error to consistent than to inconsistent sentences. As with the other Vilcu and Hadley simulations, no statistical analysis was provided so it is unknown whether these differences are reliable. If the findings are reliable, they would appear to capitalize on the well-known fact that neural networks can be exquisitely sensitive to the statistics of training patterns. If the ABA training sentences are contrived to all have peaks, then networks will likely discover that feature and then naturally find ABA sentences with valleys to be relatively novel.

Vilcu and Hadley (2003, 2005) reported a similar simulation using the sonority coding scheme of Shultz and Bale (2001). In this simulation the absolute values of both consonants and vowels were greater for the B-category syllables than for the A-category syllables in ABA sentences. Once again, one set of test sentences had ABA sentences with a peak contour, while another had ABA sentences with a valley contour. Vilcu and Hadley reported that error was smaller with peak-contoured ABA test sentences. They argued that this means that grammatical structure does not play a significant role in the model's behavior, but such a conclusion is difficult to prove without a statistical analysis. It could well be that the networks are sensitive to both contour and grammar, as later sections of this paper attest.

We did not bother trying to replicate these two simulations because procedures for constructing the ABB sentences were not described and because the simulations were so deviant from the Marcus et al. (1999) infant studies which did not contain the confound between contour and grammar. These findings could make an interesting prediction for infants, who might in fact show the same sensitivity to sonority contours as these Vilcu and Hadley networks. If so, this would provide additional support for the Shultz and Bale (2001) model.

Generalization

Vilcu and Hadley (2003, 2005) based much of their critique of the grammar-learning capacity of the Shultz and Bale (2001) model on simulation extensions that seemed to show that the model cannot interpolate or extrapolate. Simulations 3-5 investigate this claim in more detail.

Simulation 2: Interpolation

Interpolation refers to the ability to generalize within the range of the training patterns. Vilcu and Hadley (2003, 2005) tested interpolation by introducing a phonemic change to one of the four test patterns in each experiment. The original and new test patterns are shown in Table 4, where the syllables changed by Vilcu and Hadley are identified by a solid underline. With these changes, Vilcu and Hadley reported that networks could no

longer distinguish consistent from inconsistent test patterns, although they do not report any testing of statistical significance.

Table 4 Original and new test patterns[a]

Experiment	Original tests		New tests	
	Sentence	Sonority sums	Sentence	Sonority sums
1	wo fe wo	4 1 4	vo fe vo	2 1 2
	de ko de	0 -1 0	de ko de	0 -1 0
	wo fe fe	4 1 1	vo fe fe	2 1 1
	de ko ko	0 -1 -1	de ko ko	0 -1 -1
2	ba po ba	1 -1 1	ma po ma	4 -1 4
	ko ga ko	-1 1 -1	ko ga ko	-1 1 -1
	ba po po	1 -1 -1	ma po po	4 -1 -1
	ko ga ga	-1 1 1	ko ga ga	-1 1 1
3	ba ba po	1 1 -1	ma ma po	4 4 -1
	ko ko ga	-1 -1 1	ko ko ga	-1 -1 1
	ba po po	1 -1 -1	ma po po	4 -1 -1
	ko ga ga	-1 1 1	ko ga ga	-1 1 1

[a]Original test patterns are those used in the Marcus et al. (1999) infant experiments, the Shultz and Bale (2001) simulations, and the Vilcu and Hadley (2003, 2005) simulation replications. Solid underlines indicate changes to the test patterns by Vilcu and Hadley. Dashed underlines indicate additional changes to test patterns in the present Simulation 2 to eliminate confounding of phoneme and syntactic pattern.

Notice that by changing only one test pattern in each experiment, Vilcu and Hadley confounded phoneme and syntactic pattern. In all previous simulations, researchers followed the Marcus et al. (1999) experimental design in avoiding such confounding by using exactly the same phonemes in both the consistent and inconsistent test sentences. Whenever experimental conditions differ on more than one independent variable (in this case, syntax and phonology) and there are differences in some dependent variable (in this case, network error or infant interest), one cannot be sure that observed variation in the dependent variable is due to one independent variable or the other. This is because variation in the two independent variables is confounded, or correlated. Accurate causal inference requires elimination of such confounds, as in the design of the Marcus et al. (1999) experiments and our simulations (Shultz & Bale, 2001).

Here we eliminate the Vilcu-Hadley confound in each experiment by extending the same phonemic change to a test sentence in the alternate grammar. These additional changes are marked in Table 4 by dashed underlines. Importantly, these additional changes ensure that comparisons across syntactic patterns reflect only syntactic differences, unconfounded with phonemic differences.

Under these more controlled conditions, there are robust differences between consistent and inconsistent test patterns as in the original Shultz and Bale (2001) simulations. In each experiment, run with only eight networks per condition as in the infant experiments, consistent test patterns generated less network error than did inconsistent test patterns. Results regarding the key consistency effect are shown in Table

5. In each case, there was a strong main effect of test-pattern consistency and no other significant effects.

Table 5 Mean network error to consistent and inconsistent test patterns in Simulation 2, a controlled extension of the Vilcu and Hadley simulations involving all of the underlined changes in test patterns marked in Table 4

Experiment	Consistent	Inconsistent	$F(1, 14)$	$p <$
1	7.39	13.17	37	.0001
2	12.49	24.31	123	.0001
3	13.68	24.44	63	.0001

Simulation 3: Interpolation, version 2

Reasoning along similar lines, Vilcu and Hadley (2003, 2005) briefly reported a simulation in which they changed /f/ to /v/ in both the first ABA test sentence and the first ABB test sentence of Experiment 1, thus unconfounding phonemes and syntax. They reported that the networks failed to discriminate consistent test sentences from inconsistent test sentences, but provide no statistical significance test. We tried to replicate this simulation with only eight networks per condition, as in the infant experiments, and found a significant main effect of consistency, $F(1, 15) = 5.52$, $p < .05$, reflecting more error to inconsistent test sentences ($M = 11.69$) than to consistent test sentences ($M = 9.30$). There was no main effect of familiarization pattern and no interaction.

To be more certain, we ran the simulation again with 20 networks per condition to increase statistical power. Here there was a strong main effect of consistency, $F(1, 38) = 87$, $p < .0001$, and an interaction of consistency with familiarization pattern, $F(1, 38) = 9.32$, $p < .005$. Mean network error for the various conditions is plotted in Figure 6. Importantly, there was clearly more error to inconsistent test patterns than to consistent test patterns in each condition, $t(19) = 4.96$, $p < .0001$.

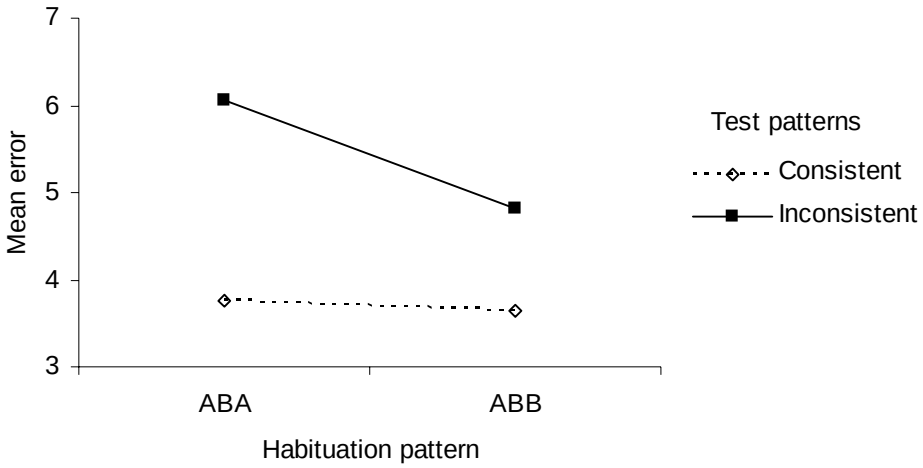

Figure 6. Mean network error to consistent and inconsistent test patterns in Simulation 3, an extension of the Vilcu and Hadley simulation of Experiment 1 involving an /f/ to /v/ change in test sentences, with 20 networks per condition.

So once again, there is strong, replicated evidence that the Shultz and Bale networks can capture the familiarity effect seen with infants when there is no confounding between phonology and syntax.

Simulation 4: Interpolation with confounded changes

Moreover, we could not replicate Vilcu and Hadley's (2003, 2005) finding of a lack of discrimination between consistent and inconsistent test patterns even using their single-pattern changes that confound phoneme with syntactic pattern. Results are presented in Table 6. In Experiment 1, there was the typical main effect of test-pattern consistency and no other significant effects. In Experiments 2 and 3, there were main effects of test-pattern consistency and familiarization pattern and an interaction between them. Importantly, in every condition of these experiments, run with 20 networks per condition to increase statistical power, there was significantly less network error to the consistent test patterns than to the inconsistent test patterns, $p < .001$ by paired-sample t test.

Table 6
Mean network error to consistent and inconsistent test patterns in Simulation 4, an attempted replication of Vilcu and Hadley's simulations involving only the solid underlined changes in test patterns marked in Table 4

Experiment	Consistent	Inconsistent	$F(1, 38)$	$p <$
1	8.56	13.68	67	.0001
2	13.93	20.17	101	.0001
3	13.79	19.59	198	.0001

The results of simulations 2, 3, and 4 indicate that Vilcu and Hadley's claim that the Shultz and Bale networks do not interpolate successfully is incorrect. With properly controlled tests, the interpolation ability of these networks is very strong. And even with the confounding introduced by Vilcu and Hadley, the networks still interpolate well. The lack of statistical significance tests in Vilcu and Hadley's research appears to have obscured real differences between test patterns.

Simulation 5: Extrapolation

To test the Shultz and Bale (2001) model for ability to extrapolate outside of the sonority training range, Vilcu and Hadley (2003, 2005) assigned four consonant values beyond the anchor value of -6, i.e., values of -7, -8, -9, and -10, and combined them with two vowel values beyond the anchor value of 6, i.e., 7 and 8. Vilcu and Hadley reported that networks had more error to consistent test patterns than to inconsistent test patterns. No statistical significance results were presented.

A major problem with testing outside the range of the sonority anchors is that it is unclear what such extreme values might correspond to, either in English or in other languages. As noted in Table 1, the sonority scale starts with voiceless stops /p/, /t/, and /k/ and ends with back low vowels /a/ and /æ/. That is to say the scale covers the entire sonority range of human speech sounds. Arbitrarily picking sonority values outside this range does not map to human speech sounds in any known language on our planet. The relevance of testing network generalization in this way is thus questionable. At best, it could be said that such predictions are not testable with psychological evidence. This is also true of Shultz and Bale's (2001) extrapolation values of -7, -6.5, 6.5, and 7, although

those values are not as far from the realistic boundaries of -6 and 6 as most of Vilcu and Hadley's extrapolation values are.

In making the argument that networks fail to extrapolate beyond the training range of -6 to 6, Vilcu and Hadley (2003, 2005) ignored the Shultz and Bale (2001) results showing that with less extreme deviations beyond the anchors of the training range, networks do successfully extrapolate, with the consistency effect growing significantly larger with more extreme (i.e., +-7) as compared to less extreme (i.e., +-6.5) sonority values. Here we report on a replication of the Shultz and Bale extrapolation results, and also extend the study of extrapolation to the extreme sonority values used by Vilcu and Hadley.

The sonority values used in this study are shown in Tables 7 and 8, along with a reminder of the original anchor values used by Shultz and Bale to simulate the Marcus et al. (1999) experiments. As in Shultz and Bale (2001), there were test values inside the training range (by +-0.5) and values that were outside of this range but close to it (by +-0.5) or far from it (by +-1.0). In addition there were three additional sonority values ranging farther outside of the training range in steps of +-1.0, labeled in Tables 7 and 8 as *farther*, *even farther*, and *farthest*. Sonority values termed *farthest* were as far outside the training range as the most extreme values used by Vilcu and Hadley. Thus we used several sonority gradations, selected them systematically, and systematically applied them to both consistent and inconsistent test sentences. Because Vilcu and Hadley's procedural descriptions of their extrapolation experiments are incomplete, it is unclear whether they applied extreme sonority values to both consistent and inconsistent test sentences. In other words, they may have introduced confounds between syntax and phonology as they had in other simulations.

Table 7 Test patterns for evaluating extrapolation in Simulation 5: Highest vowel paired with lowest consonant and vice versa

	Category A		Category B	
Distance from training range	Consonant	Vowel	Consonant	Vowel
Original anchors	-6.0	6.0	-1.0	4.0
Inside (+-0.5)	-5.5	5.5	-1.5	4.5
Close (+-0.5)	-6.5	6.5	-0.5	3.5
Far (+-1.0)	-7.0	7.0	0.0	3.0
Farther (+-2.0)	-8.0	8.0	1.0	2.0
Even farther (+-3.0)	-9.0	9.0	2.0	1.0
Farthest (+-4.0)	-10.0	10.0	3.0	0.0

In one set of our new simulation experiments, the highest vowel was paired with the lowest consonant, creating a negative correlation between consonant and vowel sonorities in both the A and B categories and keeping the sonority sums for syllables at a constant value of 0.0 in the A category and 3.0 in the B category (see Table 7). In another set of simulations, the vowel columns in Table 7 were switched so as to pair the highest vowel with the highest consonant, creating a negative correlation between consonant and vowel values in category A and a positive correlation in category B. In this set of simulations, as shown in Table 8, the sonority sums of the syllables were allowed to vary with distance

from the training range. Both sets of simulations mimicked Experiment 1 with only eight networks per condition as in the infant study.

Table 8 Test patterns for evaluating extrapolation in Simulation 5: Highest vowel paired with highest consonant and vice versa

	Category A		Category B	
Distance from training range	Consonant	Vowel	Consonant	Vowel
Original anchors	-6.0	4.0	-1.0	6.0
Inside (+-0.5)	-5.5	4.5	-1.5	5.5
Close (+-0.5)	-6.5	3.5	-0.5	6.5
Far (+-1.0)	-7.0	3.0	0.0	7.0
Farther (+-2.0)	-8.0	2.0	1.0	8.0
Even farther (+-3.0)	-9.0	1.0	2.0	9.0
Farthest (+-4.0)	-10.0	0.0	3.0	10.0

In each experiment, the amount of error to these test patterns was subjected to a mixed ANOVA in which familiarization condition served as a between-network factor and consistency and distance served as repeated measures. In both experiments there were significant main effects of consistency and distance and an interaction between them, $p < .0001$. The relevant means are presented in Figures 7 and 8 for the case where sonority sums were constant and where they were allowed to vary, respectively. As in Shultz and Bale (2001), error increased with distance from the training range, error was greater to inconsistent than to consistent test patterns at every distance, and this consistency effect grew larger with increasing distance, $p < .0001$.

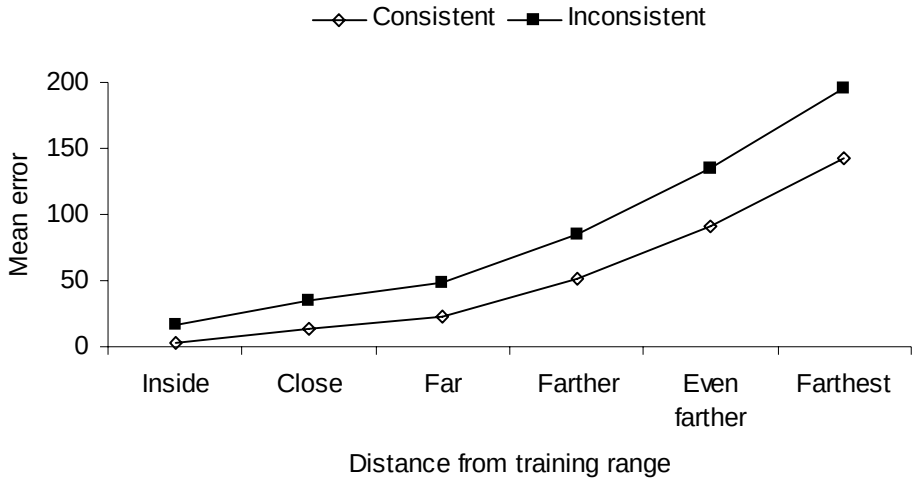

Figure 7. Mean error to consistent and inconsistent test patterns at various distances from the training range in Simulation 5, where sonority sums were constant.

Thus this new evidence confirms that the Shultz and Bale networks generalize well both outside and inside of the training range. The reason that error increases with distance outside the training range is because the network does not recognize the particular novel phonemes and syllables being presented. Importantly, however, even with wildly novel sounds, the networks readily identify the relative syntactic novelty of the sentences. Outside of the training range where there are no human speech sounds, it is difficult to

design realistic tests of the model's predictions, but in-principle evidence of network extrapolation ability is incontrovertible. These results, and those in the extrapolation simulation of Shultz and Bale (2001), underscore network capability of understanding both syntax and sonority.

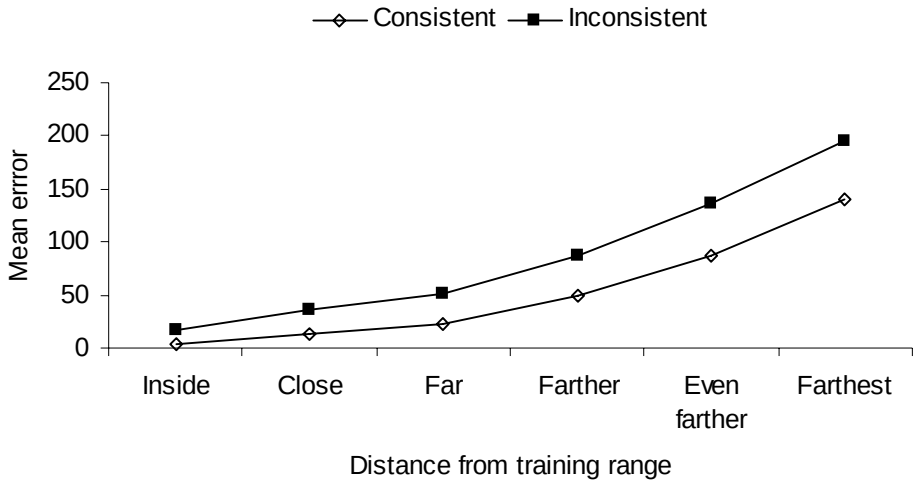

Figure 8. Mean error to consistent and inconsistent test patterns at various distances from the training range in Simulation 5, where sonority sum was allowed to vary.

Knowledge-representation analyses

An important question for this debate is whether a learning system, infant or artificial neural network, would discover and use identity relations in representing sentences. Several results of Shultz and Bale's (2001, Appendix 2) knowledge-representation analyses suggest that their networks do, in fact, learn and use near-identity relations to recognize the grammatical patterns of both new and old sentences. But Vilcu and Hadley (2003, 2005) criticized these analyses arguing that the Shultz and Bale networks fail to learn "abstract grammatical relationships." Here we discuss three types of knowledge-representation analyses of the Shultz and Bale networks, dealing respectively with connection weights, unit activations, and contributions. Contributions are products of connection weights and sending activations.

Simulation 6: Connection weights

Analyses of network connection weights showed that networks learned to encode the duplicate word before they learned to encode the single word in these artificial three-word sentences (Shultz & Bale, 2001). The reason for this is that twice as much error was generated by the duplicate words as by the single word. Importantly, networks learned to decode the representations of the two duplicate words using similar sets of weights entering the output units that represent the duplicate words. This nearly identical pattern of weights entering the output units representing the duplicate words allowed the network to recognize the near identity of these words. For example, Figure 9 shows the connection weights from a newly replicated network in the ABA condition of Experiment 1 at the end of training, when it had two hidden units.

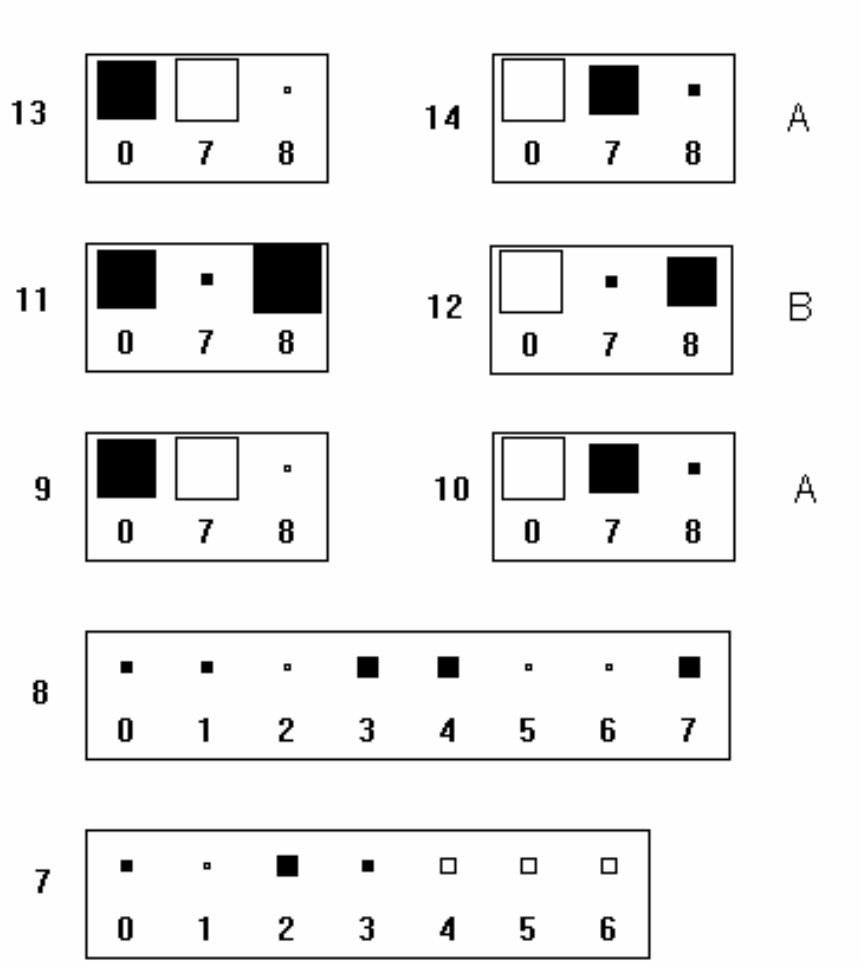

Figure 9. Weights in a network with two hidden units in Simulation 6 of the ABA condition of Experiment 1 after training was completed. Input units are labeled 1-6, hidden units are labeled 7–8, and output units are labeled 9-14. The letters A, B, and A to the right of output-unit pairs indicate the syntactic category of each word. The bias unit is 0, 1 and 9 represent the consonant of the first word, 2 and 10 represent the vowel of the first word, 3 and 11 represent the consonant of the second word, 4 and 12 represent the vowel of the second word, 5 and 13 represent the consonant of the third word, and 6 and 14 represent the vowel of the third word.

In such diagrams, connection weights entering a receiving unit are represented within a rectangular band placed just to the right of the index number of the receiving unit. Inside of each rectangular band, the weights are labeled by the sending unit and are pictured by the color and size of a square. White squares indicate excitatory weights, and black squares indicate inhibitory weights. The size of each weight is represented by the relative size of the square.

In Figure 9, the patterns of output weights for the two A-category words, represented by units 9-10 and 13-14, were highly similar, reflecting the common category of these two words. The relatively large weights to these outputs from unit 7 indicated that this first hidden unit had the task of recognizing the category of the first and third words. The

second hidden unit (unit 8) had the task of recognizing the category of the second word (B), as indicated by its relatively large weights to outputs 11 and 12. At the end of training, when these weights were recorded, the network was accurately recognizing the sentences in the training set. Similar patterns in the relative sizes of connection weights can be found in each network from each experiment. Importantly, they all show that the two duplicate words in a sentence are treated in virtually identical fashion with nearly identical connection weights entering the output units, serviced mainly by the first hidden unit.

Notice that the bias unit (unit 0) has the job of distinguishing vowels from consonants. The bias unit is always on, with an input of 1.0, regardless of what particular input is being presented. The bias unit has trainable connection weights to all downstream units, i.e., non-input units. Bias weights to the outputs encoding consonants are large and negative, consistent with the idea that consonants have negative sonority values. In contrast, bias weights to the outputs encoding vowels are large and positive, consistent with the idea that vowels have positive sonority values.

In their peaks-and-valleys simulation, Vilcu and Hadley (2005) reported that 5 of the 16 networks learned connection weights that deviated from this replicated pattern. Presumably, their other 9 networks did conform to this replicated pattern, but the degree of fit to the prescribed pattern was not quantified, plotted, or statistically analyzed. In any case, the confounding of phoneme differences with syntactic category in their simulations, which deviates from both the infant experiments (Marcus et al., 1999) and the replicated simulations (Shultz & Bale, 2001 and the present simulations), renders Vilcu and Hadley's connection-weight deviations of questionable relevance to issues of data coverage and generalization.

Simulation 7: Hidden-unit activations

Another analysis examined hidden-unit activation patterns in response to different input patterns (Shultz & Bale, 2001). Relations were plotted between hidden-unit activation and sonority sums of the A and B categories in the training patterns. These plots showed that the first hidden unit learned to encode the sonority sum of the duplicate word, and the second hidden unit learned to encode the sonority sum of the single word. Once again, it was clear that the two duplicated words were being treated in nearly identical fashion.

Analyses of connection weights in simulation 6 clearly indicate the impact of hidden units on network outputs. What they do not illuminate is how the hidden units integrate input coming from the bias unit, the input units, and the earlier hidden units. Indeed, connection weights entering hidden units had complex and variable patterns. There were suggestions, from connection-weight diagrams that the first hidden unit represented sonority variation in the duplicated word and that the second hidden unit represented sonority variation in the single word, but that is about all we could discover from examining weights alone.

To more directly investigate how hidden units integrate their inputs, we examined the activation patterns that the hidden units exhibited in response to different input patterns. Such activation patterns essentially summarize the knowledge representations of the hidden units. We ran a few networks in each condition of each experiment in a full

replication of the Shultz and Bale (2001) networks, while recording hidden-unit activations at the end of training. For each network, we plotted the relation between hidden-unit activation and sonority sums of the A and B categories in the 16 training sentences. Figure 10 provides such a plot for a network in the ABA condition of Experiment 1. It shows a negative relation between activation of the first hidden unit and the sum of sonority values (consonant plus vowel) for the category-A words. Some of the points of this plot at the activation extremes of -0.5 and 0.5 are highly overlapping, reflecting very consistent performance. Figure 11 shows a negative relation between activation of the second hidden unit of this same network and the sum of sonority values for the category-B words.

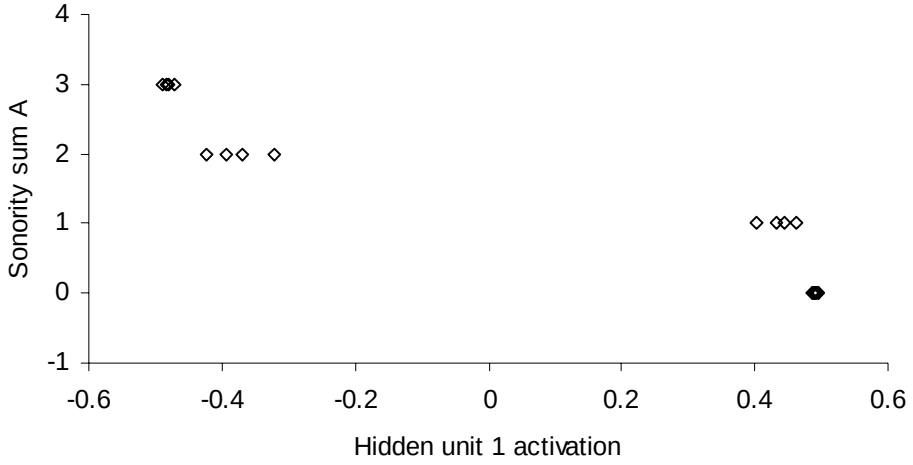

Figure 10. Relation between sonority sums and activation of the first hidden unit in a network at the end of training in Simulation 7.

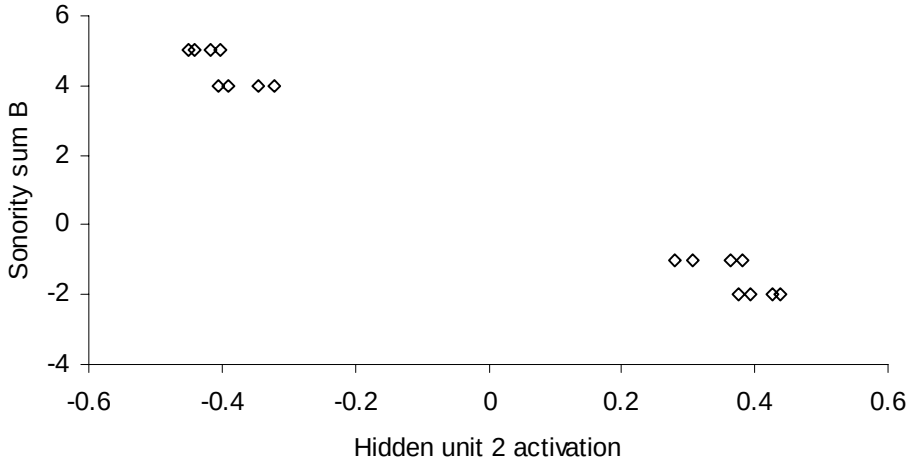

Figure 11. Relation between sonority sums and activation of the second hidden unit in the same network as in Figure 10 at the end of training in Simulation 7.

As with the other network analysis techniques, one could equally well substitute sonority differences for sonority sums in such plots. The results and conclusions would

be essentially the same. Plots of sonority variation in either consonants or vowels alone also produce similar results.

It can be concluded from these results that the first hidden unit represents sonority variation in the duplicate-word category, whereas the second hidden unit represents sonority variation in the single-word category. Again, this is a natural result of networks focusing on the largest current source of error. Because duplicate words initially generate about twice as much total error as do single words, networks deal first with the duplicate-word category.

Thus, the highly variable pattern of connection weights entering a hidden unit implements a function representing sonority variation in the word category that is generating the currently largest source of error. Each network implements such functions in a somewhat distinct way, but the functions themselves appear critical to a successful solution of this grammar-recognition problem. Sonority variation within each word category must be accurately represented by the hidden units because this variation needs to be reproduced on the output units. Importantly once again, these networks achieve a compact representation on the hidden units such that both of the duplicate words in a sentence are represented on the same, first hidden unit, thus allowing a network to realize the near equivalence of these duplicate words. All of the networks we have analyzed show this pattern, although sometimes the relation between hidden-unit activation and sonority sums is positive rather than negative.

Hidden-unit activations correlate with vowel and consonant sonorities

Vilcu and Hadley (2005) undertook similar analyses of hidden-unit activations and claimed instead that these activations correlated only with consonant, but not with vowel, sonorities. However, they did not report the values of any such correlations or evaluate their statistical significance. Moreover, their plots of vowel sonorities against hidden-unit activations for one network (their Figure 8) suggest negative correlations between the two variables rather than no correlation.

To investigate this issue more systematically, we reanalyzed some networks from the initial Shultz and Bale (2001) simulations, computing Pearson product-moment correlations between hidden-unit activations and word-category sonorities of consonants and vowels and their sums and differences. Results for two representative networks from the ABA-familiarization condition of Experiment 1 are presented in Tables 9 and 10. The network portrayed in Table 9 finished with two hidden units. The network in Table 10 finished with three hidden units. Correlations significant at $p < .01$, $df = 15$ are indicated in these Tables in bolded font. The rest of the correlations in Tables 9 and 10 are not statistically significant.

The patterns of correlations in both networks reveal that activation of the first hidden unit correlated strongly with sonority of the A-word category, whereas activation of the second hidden unit correlated strongly with sonority of the B-word category. These correlations were about equally strong for vowel sonority as for consonant sonority, and for the sum and difference of vowel and consonant sonority. The third hidden unit in the network shown in Table 10 had activations that correlated somewhat with both word categories, but more strongly with the A category.

Table 9 Pearson product-moment correlations between hidden-unit activations and word-category sonorities of consonants and vowels and their sums and differences for a representative network with two hidden units

Sonority variable	Hidden 1	Hidden 2
Consonant A	**-0.97**	0.01
Vowel A	**0.94**	-0.01
Sum A	**-0.95**	0.01
Difference A	**0.97**	-0.01
Consonant B	-0.19	**0.98**
Vowel B	-0.18	**0.99**
Sum B	-0.19	**0.99**
Difference B	0.18	**-0.91**

Table 10 Pearson product-moment correlations between hidden-unit activations and word-category sonorities of consonants and vowels and their sums and differences for a representative network with three hidden units

Sonority variable	Hidden 1	Hidden 2	Hidden 3
Consonant A	**0.93**	0.00	**-0.81**
Vowel A	**-0.91**	0.00	**0.79**
Sum A	**0.90**	0.00	**-0.79**
Difference A	**-0.92**	0.00	**0.81**
Consonant B	-0.29	**-0.98**	-0.41
Vowel B	-0.28	**-1.00**	-0.41
Sum B	-0.29	**-0.99**	-0.41
Difference B	0.29	**0.92**	0.39

It is unclear why Vilcu and Hadley (2005) concluded that vowel sonority was unimportant in this context. Our analysis shows that vowel sonority correlates as highly with hidden-unit activation as does consonant sonority or any linear combination of vowel and consonant sonority. This is despite the fact that vowel sonority has a more restricted range (4 to 6) than does consonant sonority (-1 to -6) as shown in Table 1. A negative correlation is just as important as a positive correlation of the same size. Visual examination of bivariate plots can be a good technique for exploring correlations, but this must be done carefully. Actually computing the numerical values of correlations and examining their statistical reliability, as here, is even more definitive.

Again, a more general conclusion is that the first-recruited hidden unit represents sonority variation in the duplicate-word category, whereas the second-recruited hidden unit represents sonority variation in the single-word category. Such representations form the basis for discriminating category-A words from category-B words and discriminating grammatical patterns like ABA from patterns like ABB. Both vowels and consonants figure importantly in these representations of sonority.

Simulation 8: Contribution analysis

A third type of knowledge-representation analysis incorporates both unit activations and connection weights. Network contributions are products of sending-unit activations and connection weights going into the network's output units (Sanger, 1989). Because net input to a unit is the sum of such products, there is a sense in which contributions

represent all of the influence on the output units. These considerations make contribution analysis an important tool because occasionally the effects of connection weights can be swamped by large activations, or conversely the effects of unit activations can be swamped by large connection weights.

Large contribution matrices (of input patterns x contributions) are typically simplified by subjecting them to a Principal Component Analysis (PCA) (Sanger, 1989). PCA is a data-reduction method for detecting the major independent features of variation in a dataset by taking advantage of the correlations between variables (Jolliffe, 1986). Here the variables are network contributions. PCA often provides a revealing picture of knowledge representations in cross-connected CC networks (Shultz, Oshima-Takane, & Takane, 1995). Typically we apply PCA to the covariance matrix of contributions, employ a varimax rotation in order to better interpret the principal components, and retain only eigenvalues greater than the mean eigenvalue. We record contributions from a few networks in each condition of each experiment at the close of each output phase, by which time networks have fully adjusted to each newly recruited hidden unit.

Figures 12 and 13 show PCA results for a single representative network run in a new replication of the original Shultz and Bale simulations. This particular network was familiarized with ABA sentences in Experiment 1. Figure 12 shows the network's knowledge structure after adapting to a single hidden unit; Figure 13 shows the knowledge structure at the end of training after adapting to a second hidden unit.

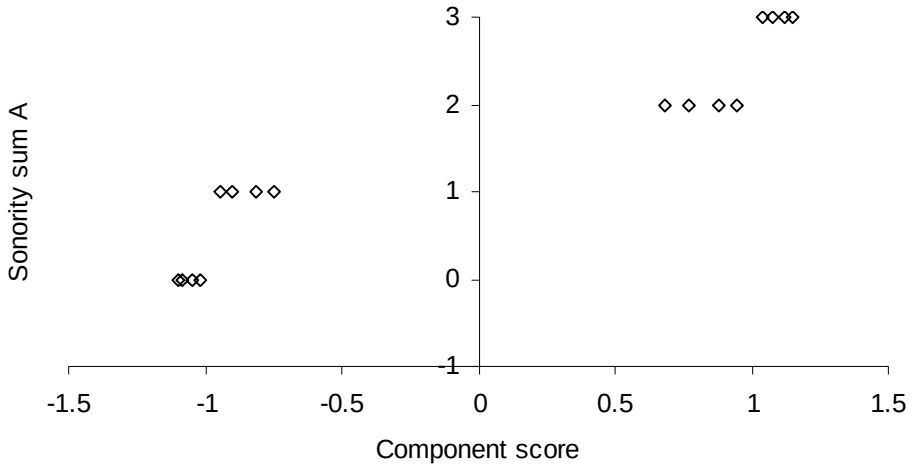

Figure 12. Mean sonority sums in A-category words as a function of component scores for the 16 ABA training sentences in a network with a single hidden unit at the end of the second output phase in Simulation 8. The single component represents sonority variation of the duplicated (A) word.

As shown in Figure 12, this network with one hidden unit had a single-component solution emphasizing sonority variation in the category-A words. The figure plots sonority sums (consonant plus vowel) of the A-category words against component scores on each of the 16 training sentences. The steady increase in these sonority sums with increases in component scores indicate that this single component is sensitive to variation in sonority of the A-category words. Remembering the analysis of connection weights,

this makes sense because this network was familiarized to ABA sentences, where category-A words would receive more initial attention because they are twice as frequent. This single component accounts for 100% of the variance in network contributions. At this point, the network does not possess enough computational power to also represent sonority variation in the B-category words.

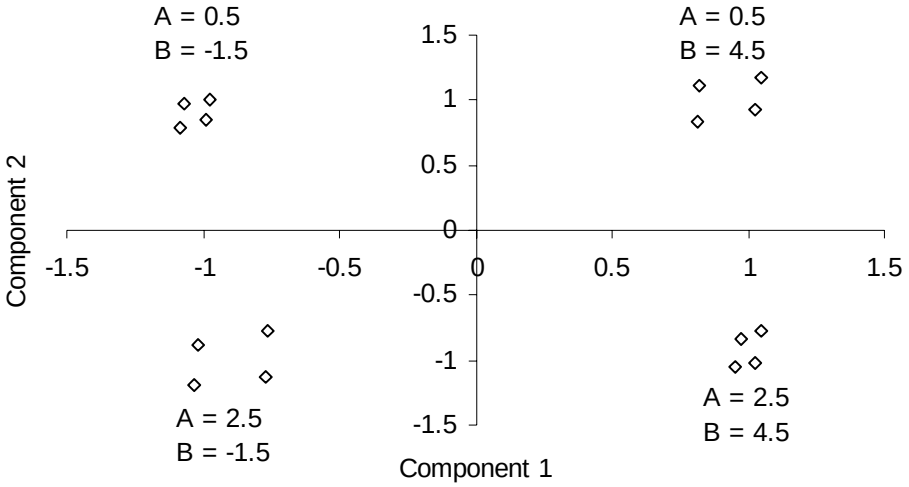

Figure 13. Component scores on the 16 ABA training sentences for the same network in Figure 12 with two hidden units after training was completed, along with mean sonority sums for the A- and B-category words in each sentence. Sonority variation of A-category words is represented by Component 2, explaining 29.3% of the variance. Sonority variation of B-category words is represented by Component 1, explaining 70.2% of the variance.

At the end of training, after adapting to the second hidden unit, the PCA revealed two principal components as shown in Figure 13. Together the two components accounted for 100% of the variance in contributions. Figure 13 plots Component 2 scores against Component 1 scores for each of the 16 training sentences. Notice that the 16 training sentences cluster into four groups of four sentences each. The nature of these clusters can be understood by noting the mean sonority sums for the A and B categories. Component 1, with a large loading from the second hidden unit, reflects sonority variation in the B-category words, while Component 2, with a large loading from the first hidden unit, reflects sonority variation in the A-category words. Both Figures 12 and 13 again provide strong evidence that the networks learned to treat the duplicate words nearly identically. Networks do not need separate representations for the duplicate A-category word – a single representation is sufficient.

Component scores correlate with vowel and consonant sonorities

These PCAs of network knowledge representations precisely replicate those reported for the original simulation (Shultz & Bale, 2001). Vilcu and Hadley (2005) criticized the original PCA analyses on the same basis that they criticized the analysis of hidden-unit activations, namely by claiming that the components reflect variation in consonant, but not vowel, sonority. In this case they presented plots of sonority sums (their Figures 2 and 4) and consonants (their Figures 3 and 5) against component scores for two networks.

But unlike their critique of the hidden-unit-activation analysis, they did not provide any plot of the relation between vowel sonority and component scores. As in their critique of the hidden-unit-activation analysis, they did not actually report the values of the correlations that they described.

To investigate this issue more systematically, we reanalyzed data from several networks in the original simulations (Shultz & Bale, 2001). For each network we computed Pearson product-moment correlations between component scores and word-category sonorities of consonants and vowels and their sums and differences. Results for two representative networks, both from the ABA-familiarization condition of Experiment 1, are presented in Tables 11 and 12. In both of these Tables, the *Only component* column lists correlation coefficients at the end of output-phase two when the network had adjusted to the first hidden unit; the columns labeled *Component 1* and *Component 2* list correlation coefficients at the end of training when the network had adjusted to two hidden units. As before, those correlations that are statistically reliable, $p < .01$, $df = 15$, are highlighted in bold typeface. The pattern of correlation coefficients confirms that component scores from the PCAs correlate highly with vowel, as well as consonant, sonorities, and with linear combinations of the two, namely sums and differences. These results contradict Vilcu and Hadley's contention that vowel sonorities have no impact on network contributions. Vowel sonority, despite having a relatively restricted range, has at least as much impact as does consonant sonority.

Table 11 Pearson product-moment correlations between component scores and word-category sonorities of consonants and vowels and their sums and differences for a representative network

Sonority variable	Only component	Component 1	Component 2
Consonant A	**0.98**	-0.02	**0.98**
Vowel A	**-0.99**	0.02	**-0.99**
Sum A	**0.92**	-0.01	**0.92**
Difference A	**-0.99**	0.02	**-0.99**
Consonant B	0.10	**0.98**	0.01
Vowel B	0.11	**1.00**	0.02
Sum B	0.10	**1.00**	0.01
Difference B	-0.09	**-0.92**	-0.01

Table 12 Pearson product-moment correlations between component scores and word-category sonorities of consonants and vowels and their sums and differences for a second representative network

Sonority variable	Only component	Component 1	Component 2
Consonant A	**-0.98**	0.05	**-0.98**
Vowel A	**0.99**	-0.05	**0.99**
Sum A	**-0.92**	0.05	**-0.92**
Difference A	**0.99**	-0.05	**0.99**
Consonant B	0.11	**0.98**	0.06
Vowel B	0.10	**1.00**	0.05
Sum B	0.11	**0.99**	0.06
Difference B	-0.12	**-0.91**	-0.08

The reason that we (Shultz & Bale, 2001) argued that networks were integrating vowel and consonant sonorities into a syllable-level representation is that only one PCA component was required to summarize sonority variation for each word category (A or B). If networks had to represent each syllable's consonant and vowel separately, then PCA results would show separate components for consonants and for vowels, which has never happened in our analyses.

As noted earlier, networks develop even more efficient representations than that by including the duplicate word (e.g., A in ABA familiarization, B in ABB familiarization) on the same component. Again, this underscores that networks discover a near-identity relation between the duplicate words.

Summary of knowledge-representation analyses

Results of these network-knowledge-representation analyses can be summarized as follows:

1. These networks learn to encode syllables as a linear combination of consonant and vowel sonority, i.e., as either the sum or difference of the sonorities of the consonant and vowel.
2. Despite the relatively limited range of vowel sonority, vowels are equally as important as consonants in network knowledge representations.
3. Because these networks always try to reduce as much error as possible, the first-recruited hidden unit focuses on the two duplicate words and the second-recruited hidden unit focuses on the single word in each three-word sentence.
4. Bias weights in these networks learn to encode the distinction between consonants and vowels.
5. These networks learn to decode duplicate words with very similar sets of weights to the output units that represent the duplicate words.

These results and conclusions replicate and extend those presented in Shultz and Bale (2001). Taken together these different analytical techniques provide replicated and consistent evidence that the Shultz and Bale networks learn to distinguish the Marcus et al. (1999) grammars by recognizing the near identity of duplicate words in simple three-word sentences. This involves discovering a somewhat abstract relationship, something that is different than mere contour matching.

Simulation 9: The role of sonority contours in syllabification

It may be that Vilcu and Hadley (2003, 2005) considered contours in an over-compartmentalized way, by pitting contour following against syntactic discrimination, as if by learning about sonority contours a system cannot also learn about syntax. Sonority contours are generally important in language acquisition because they are one of the keys to identifying syllables and words. Infants (and presumably network simulations) listen to speech-sound sequences and the information gained can be useful for a variety of language issues, including syllable and word identification and acquisition of syntactic rules. These language functions are eventually dealt with by one integrated computational system and cannot always be easily separated into abstract linguistic partitions. Indeed,

identifying syllables and words is arguably of central importance in acquiring sentential syntax.

To see why this is true, it is useful to review the standard linguistic interpretation of syllables. A syllable is the smallest pronounceable unit in a language. More formally, as portrayed in Figure 14, a syllable is a unit of sound composed of an optional onset (normally a consonant or consonant cluster) followed by a required rime (Kaye, 1989). A rime, in turn, is defined as having a required nucleus (normally a vowel) followed by an optional coda (normally a consonant or consonant cluster).

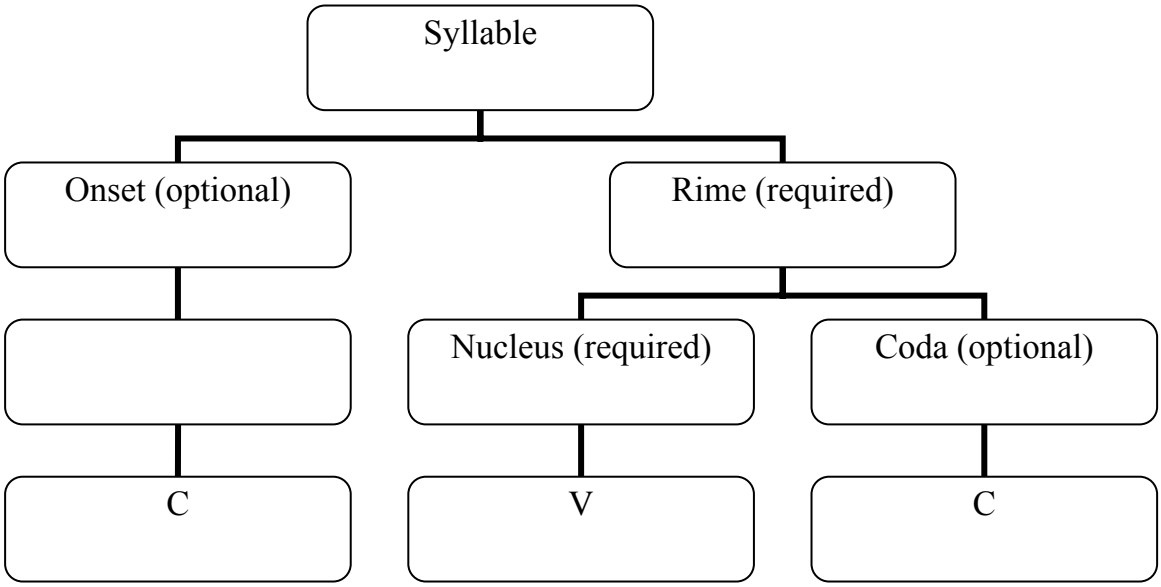

Figure 14. General scheme for a syllable.

As an example, the syllabic structure of the word *simulate* is shown in Figure 15. Here the first two syllables possess an onset and a rime containing an onset but no coda. The third syllable also has an onset and a rime, but this rime contains both a nucleus and a coda.

Syllables are known to affect psychological processing of speech input in adults (Mehler, Dommergues, Frauenfelder, & Segui, 1981; Vroomen & de Gelder, 1997). But for infants, the syllables of their particular language are not provided for free – rather they must be learned. Because the speech stream that infants hear is often continuous, this learning can be quite difficult, but there is evidence that infants as young as eight months are able to learn about syllables from the statistics available in continuous speech (Saffran, Aslin, & Newport, 1996).

There is an emerging view that some of the important clues to syllable detection are provided by the sonority contours of human speech (Vroomen et al., 1998). The *sonority principle* states that, within a syllable, sonority starts low at the onset, increases to a peak at the nucleus, and decreases into the coda (Selkirk, 1984). This notion of sonority is compatible with the way it is used in our work because we based our sonority scale on this linguistic research (Shultz & Bale, 2001).

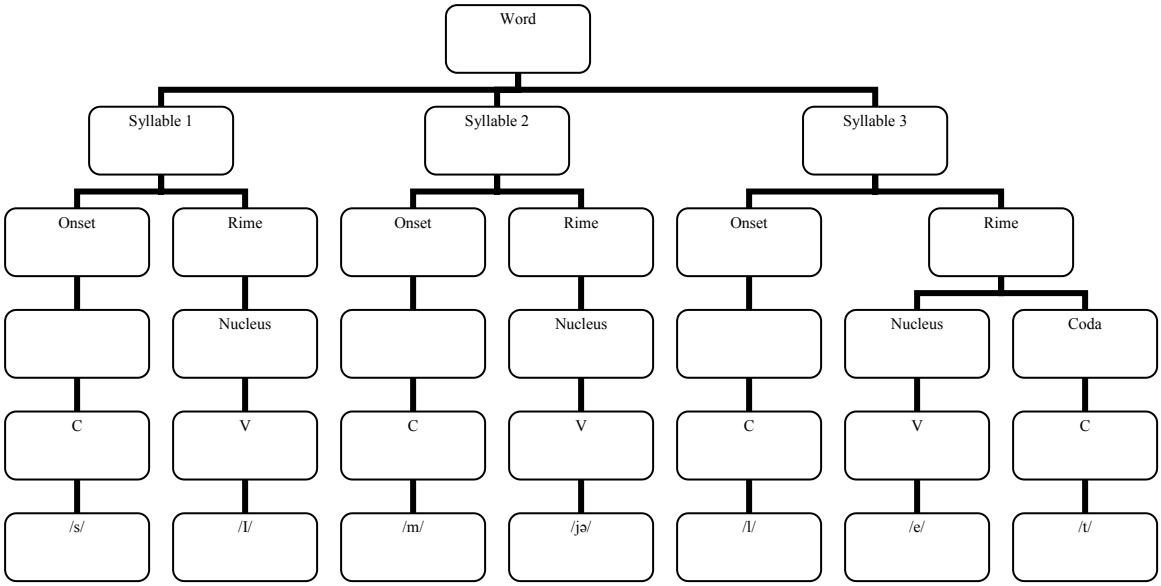

Figure 15. Syllable structure of the word *simulate*.

To see whether networks like ours could learn syllable boundaries from sonority contours, we ran an exploratory simulation using the three most common phonemic sequences defining syllables: CV, CVC, and V, where C refers to a consonant and V to a vowel. All six orders of these common sequences were used to construct three-syllable words, as shown in Table 13, where periods indicate syllable boundaries. We created 100 token words of each of these six sequence types with random assignments of integer values: C from -1 to -6 in steps of 1 and V from 4 to 6, also in steps of 1, as in the sonority scale in Table 1. There were six input units to code this sonority information, and six output units to code a binary decision about the location of syllable endings at each of the six positions: -0.5 for *no*, 0.5 for *yes*. This target information would be akin to infants hearing pauses between syllables, syllables at the beginning or end of single words, or isolated monosyllabic words.

Table 13 Syllable Strings Used in Simulation 9[a]

CV.CVC.V	CV.V.CVC	CVC.CV.V
CVC.V.CV	V.CV.CVC	V.CVC.CV

[a]Target boundaries are marked with a period.

Such information is not always available in language input, but it is likely to be often available and particularly useful in helping to identify syllables based on sonority contours. Random assignment of sonority values reflects the fact that we were not simulating any particular words in this exploratory simulation, but rather three-syllable words in general. Importantly though, all of the syllable boundaries in these words conformed to the three common syllable types of CV, CVC, or V. Note that these are not encoder networks as in simulations 1-8, but rather *pattern-associator* networks that learn a function mapping input patterns to (different) output patterns.

Instead of standard CC, we used an interesting variant called sibling-descendant cascade-correlation (SDCC) that dynamically decides whether to install each new hidden unit on the current highest layer (as a sibling) or on its own separate layer as a descendant

(Baluja & Fahlman, 1994). Correlations of network error with activations of descendant candidates were penalized by being multiplied by 0.8, a value that has been found to reduce network depth while maintaining good generalization. We have recently come to prefer SDCC over standard CC because it creates a greater variety of network topologies of a more biologically realistic depth (Shultz, 2006). Otherwise though, the performance and functionality of SDCC is quite similar to that of standard CC.

Twenty SDCC networks were trained until all output units on all 100 training patterns were within 0.4 of their target values or a maximum of 1000 epochs was reached. That value of 0.4 is the standard score-threshold used to signal successful training with binary output units (Shultz, 2003). Each network was trained on a randomly-selected 100 training patterns and tested on another randomly-selected 100 test patterns, none of which were used in training.

The hidden-unit structures created by these 20 networks are shown in Table 14. For example, a structure of 3-3-1 refers to three hidden units on each of the first two layers and one hidden unit on the third layer. As is typical of SDCC, the network topologies here range from entirely flat to quite deep.

Table 14 Hidden-unit structures developed by 20 SDCC networks in Simulation 9

3 3 1	5	3 1 3	5 1 3	1 1 4
2 3	5 1	7	3 3	3 2
3 1 3 1	4 1	1 3 3	2 3	1 2 1
1 7	1 1 3	5 1 1	1 2 2 2	1 1 4

Mean results for these systematically-trained networks are presented in the penultimate row of Table 15, where it can be seen that these networks were virtually perfect on the training patterns and generalized very well to the test patterns. To make sure that these networks had really learned about syllable boundaries, a control set of 20 networks was trained on randomly-selected input values in the sonority range of -6 to 6. These input patterns were each paired with sets of six binary output values, a randomly-selected three of which indicated a syllable boundary (0.5), the other three indicating no syllable boundary (-0.5). For these control networks, then, there were no systematic relations between sonority contours and syllable boundaries, but the actual input and output values used in training were in the same range used with the systematically-trained networks. These control networks were additionally tested on 100 randomly selected, but systematic test patterns of exactly the same type that the systematically-trained networks were tested on. The mean performance of these control networks is presented in the last row of Table 15. None of the control networks mastered the training patterns within the limit of 1000 epochs, but some learning of the random training patterns did occur, as indicated by their mean success rate of 45%. However, because there was no generalization to the systematic test patterns (2% success), it is clear that this learning involved memorization of the training patterns rather than function abstraction as in the systematically-trained networks. This is one case where a statistical test of the differences between conditions is not required because the differences are so clear.

These results show that systematically-trained networks learned to find syllable boundaries in streams of speech based on realistic sonority contours. The abysmal performance of the randomly-trained networks indicates that the performance of the

systematically-trained networks was based on genuine function abstraction and not on sheer memorization of the training patterns. This is not a definitive simulation of any psychological experiment but rather a *proof of concept* to show that our networks can do this kind of thing. Similar simulations with words in real or artificial languages, rather than randomly-constructed words as here, could be used to generate predictions for new psychological research.

Table 15 Mean performance of 20 SDCC networks in Simulation 9 (Standard deviations in parentheses)

Training	Epochs	Proportion correct	
		Train	Test
Systematic	868 (122)	.99 (.01)	.73 (.06)
Random	1000 (0)	.45 (.06)	.02 (.01)

Another demonstration of this sort would be to examine whether output sonority peaks at the position of the nucleus (represented by a vowel) and decreases on either side of the nucleus in encoder networks like those used in simulations 1-8. This has already been found with simple recurrent networks learning to predict the next phoneme in continuous speech (Vroomen et al., 1998). Such a hat-shaped sonority profile, with sonority falling off on either side of a vowel-defined nucleus, essentially reflects the sonority principle. It may not actually be necessary to test this hypothesis with our networks because previous and present network analyses suggested that our networks employ the bias (always on) unit for this task, creating positive weights to vowels and negative weights to consonants. The Marcus et al. (1999) words are all composed of CV syllables, which with our coding scheme creates a sawtooth-shaped sonority contour across the sentence (see Figure 1). Simulating this pattern would be unnecessary to show and quite easy to achieve. With more complex varieties of syllables, as in the present syllable-detection simulation, bias weights alone won't suffice, and hidden-units become critical, making simulations more interesting.

A general point is that there is no need to expect perfect syntactic generalization in either infants or networks from sonority contours alone, as Vilcu and Hadley (2003, 2005) apparently require. This is partly because sonority contours are used for multiple purposes in language acquisition: detecting syllable boundaries (and then perhaps using these to detect word boundaries), discriminating syntactic forms, and maybe other functions that are not necessarily incompatible.

Discussion

Vilcu and Hadley's (2003, 2005) critique of neural-network models of infant learning of artificial grammars is important because it goes to the heart of a debate that has dominated cognitive science for the last 20 years – whether human cognition is better explained by symbolic rules or by subsymbolic connections. Their papers focused on the Shultz and Bale (2001) model because that is the one that could cover the Marcus et al. (1999) infant data according to Vilcu and Hadley's replication methods. Vilcu and Hadley focused in particular on whether the Shultz and Bale model actually learns the artificial grammars to which the infants were exposed. Because their extensions of the Shultz and Bale model failed to generalize to novel sentences, in terms of both interpolation and extrapolation, Vilcu and Hadley concluded that this model does not

really learn the grammars. Instead Vilcu and Hadley argued that the Shultz and Bale model only learns the numerical sonority contours of the artificial sentences, and not grammatical relationships involving the near identity of duplicated words.

Results presented here show that Vilcu and Hadley seriously underestimated the ability of the Shultz and Bale model to discover the somewhat abstract near-identity relations that are key to distinguishing the Marcus et al. grammars and to generalizing both outside and inside of the training range. Our results confirm that the numerical contours actually used by the networks to discriminate syntactic patterns are not raw sonority values (input to the network) as Vilcu and Hadley had suggested, but rather sonority sums (or equivalently sonority differences). Furthermore, the contours of these sonority sums learned by the networks and represented on hidden units are not simply peaks (or valleys, or other simple patterns), but rather a complex combination of contours involving peaks, valleys, increases, decreases, and plateaus in various experimental conditions. Faced with this complex mix of contours, the networks discovered and used near-identity relations to recognize the grammatical patterns of both new and old sentences. This analysis of the actual sonority-sum contours enabled simulation of apparent differences between the Marcus et al. (1999) experiments. Experiments 2 and 3, containing flat sonority-sum profiles which made it difficult to distinguish monosyllabic words, and thus the syntactic categories of these words, were more difficult to learn than Experiment 1 which contained only more discriminating profiles.

The networks learned to decode the representations of the two duplicate words in a sentence by using similar sets of weights entering the output units that represent the duplicate words. This nearly identical pattern of weights entering the output units representing the duplicate words allowed the network to recognize the near-identity of these words.

The relatively large connection weights to the duplicated-word outputs from the first hidden unit indicated that this hidden unit had the task of recognizing the category of these duplicate words. The second hidden unit had the complementary task of recognizing the category of the single word, as indicated by its relatively large weights to outputs representing that single word.

Analyses of hidden-unit activations showed that the first hidden unit learned to encode the sonority sum of the duplicated words, and the second hidden unit learned to encode the sonority sum of the single word. Again, it is clear that the two duplicated words were being treated in identical fashion.

Networks learned to take the raw sonority values that constituted network inputs and re-represent them as sums (or equivalently as differences) on the hidden units. A significant advantage of neural networks over symbolic rule-based learners is the ability of networks to build such novel representations. A leading symbolic rule-learning algorithm was unable to build such representations and unable to simulate the infant data considered here (Shultz, 2001).

PCAs of network contributions revealed two components, one representing sonority variation in the duplicate-word category and the other representing sonority variation in the single-word category. Once again this is evidence that the networks learned to treat the duplicate words identically. Analyses of both hidden-unit activations and network

contributions confirmed that sonority variation in both vowels and consonants was important in network representations.

Under well controlled conditions, without confounding phoneme and syntactic pattern as in Vilcu and Hadley's experiments, there were robust differences between consistent and inconsistent test patterns as in the original Shultz and Bale simulations and the infant data. Moreover, even with Vilcu and Hadley's confounds left in, these effects were still robust as assessed by conventional statistical tests. The introduction of experimental confounds and lack of statistical significance tests in Vilcu and Hadley's research appeared to obscure real differences between test patterns and thus underestimated their networks' ability to distinguish these simple grammars. It is commonplace in experimental work to eliminate confounds in conditions in order to make accurate causal inferences about the effects of conditions on dependent variables. If the goal is to determine whether infants have learned the syntax of a language they are exposed to, and this will be determined by their interest in novel test sentences with familiar or alternate syntax, then these test sentences should vary only in syntax, not in phonology or any other characteristics.

Vilcu and Hadley's tests of network extrapolation beyond the training range used sonority values that do not map onto human speech sounds in any known language, thus invalidating their simulations as realistic tests of grammar-learning ability. Even with sonority values as extreme as those introduced by Vilcu and Hadley, the networks did successfully extrapolate beyond their training range. We demonstrated this with experimental designs that systematically varied sonority values without confounding them with syntactic differences. If generalization by interpolation and extrapolation is the criterion for grammar learning, then these networks indeed learned to distinguish these simple grammars.

Rather than accepting Vilcu and Hadley's view that tracking sonority contours is inconsistent with syntactic discrimination, we supported the emerging view that sonority contours provide important clues to the identification of syllables in speech streams. Our exploratory simulation showed that SDCC networks can detect syllables using sonority-contour cues. Syllable identification, in turn, may help to identify words, an essential step in acquiring syntax.

It is important to be realistic about what is happening in the infant experiments. There are 7-month-olds listening to recorded speech in a nonsense language. They have no idea the experiment concerns syntax or even that they are in an experiment. The speech sounds they hear have a variety of phonemes and sonorities. Measures taken from these infants are not explicit judgments about syntactic correctness, as might be done with adult linguists or occasionally with ordinary adults. Instead, the experimenters measured mere looking time towards an audio speaker that was emitting words with one of two different syntactic patterns. The infants showed a slight, but significant preference for listening to sentences from the more novel of the two languages, suggesting some sort of ability to distinguish one syntactic pattern from another. There is no compelling reason to assume these infants were focused only or even mainly on syntax. They may equally well have focused on sonority contours in the syllables they were hearing. Indeed, such processing of phonological content would have been important to their ability to distinguish syntactic patterns. In other words, like adult listeners, these infants might be expected to

process spoken language in a variety of different linguistic aspects. Thus, it would not be surprising to find evidence that the infants were processing phonology as well as syntax. If so, a computational model that includes and remains sensitive to phonological variation, as the Shultz and Bale model does, will likely fare better than a purely syntactic, symbolic model of the sort favored by some others (Marcus et al., 1999; Vilcu & Hadley, 2005). As infant researchers test predictions of some of the computational models of these initial data, a more complete description and explanation will doubtless emerge.

Vilcu and Hadley (2005) reported that Marcus (2001) does not regard the arbitrarily-coded model of Shultz (1999) as a counterexample to Marcus' own claims about the need for rules with variables. Marcus' (2001) argument is based on the supposed presence of identifiable *variables* in the input and output layers of the network, an argument that could also, according to Vilcu and Hadley, be applied to the sonority-coded model of Shultz and Bale (2001). This argument ignores two important considerations. One is that all computational models that learn from examples, whether connectionist or symbolic, require some systematic coding of the inputs and outputs. If inputs and outputs cannot be described in terms of their essential features, no mapping from inputs to outputs can be learned. The other consideration is that to characterize such input and output coding in networks as *variable binding* ignores fundamental differences between symbolic and neural systems in terms of how knowledge is represented and processed. Even if some researchers are uncertain about how neurons or connectionist units might encode variables, the functioning of symbolic and neural computational systems has been well understood in the computational literature for some time. The most popular type of symbolic system is implemented in what are called *production systems*. Here long-term knowledge is represented in symbolic if-then rules, which are formed out of propositions often containing variables. These rules are selected and used (fired) when their conditions match the contents of a working-memory buffer containing propositions describing the current state of the problem being worked on. During that matching process, considerable computational machinery makes sure that the binding of values to variables is consistent across rule clauses and instances. Without consistent variable binding, performance of production systems falls apart. Unstructured neural networks (such as CC and SDCC), in contrast, represent long-term knowledge in connection weights, and process information by passing activation from unit to unit without regard for whether input values are bound to variables in a consistent fashion. Importantly, in unstructured neural networks there are no variables, no propositions, no if-then rules, no matching of rule conditions, no rule firing, and no separate working-memory buffer. Any assignment of values to input nodes are lost as soon as activation is passed forward from the input layer to other layers of the network. These fundamental differences between symbolic and neural systems are well known to modelers and non-modelers on both sides of the symbols vs. connections debate (e.g., Anderson, 1993; Pinker, 1997; Shultz, 2003). Not only do symbolic and neural models operate differently, they often make different predictions for the same phenomena (Shultz, 2003). A single example that forms the basis for the current controversy is that symbolic rules predict perfect generalization to all instances whose descriptive values can be consistently bound to rule variables (Vilcu & Hadley, 2005), whereas neural networks predict contextualized generalization, depending in part on the content of the items (see, for example, Simulations 1 and 5). To argue that ordinary,

unstructured neural networks such as CC are binding values to variables is to offer a bizarre and incorrect interpretation of what is transpiring in these very different computational systems.

The current successful coverage of these data by the Shultz and Bale (2001) model does not imply that these rather simple networks could acquire the full grammar of a human language. Much more needs to be understood about such abilities and how they emerge in children before more general models can be built. It is likely that some aspects of human language acquisition would require different and more powerful models. But the ability of these networks to master the simple artificial grammars used by Marcus et al. (1999) with infants is well established. Indeed, these unstructured neural networks can learn these grammars more effectively and generalize better than a leading symbolic rule-learning method provided with the same training and test patterns (Shultz, 2001). That algorithm, called C4.5 (Quinlan, 1993), is arguably the most successful existing technique for learning symbolic rules from examples, and yet it failed to learn the rules specified by Marcus et al. (1999), completely failed to generalize to test patterns, and covered none of the phenomena observed with infants.

In their footnote 4, Vilcu and Hadley (2005) assert without proof or argument that the learning rules used in the CC models by Shultz and Bale (2001) are difficult to justify biologically. Of interest to this issue is a recent demonstration that these learning rules are mathematically equivalent to slightly extended versions of the so-called *Hebb* rule that is widely considered to be biologically plausible (Rivest & Shultz, 2005). Other new evidence for the constructive learning implemented in CC and SDCC networks has been summarized elsewhere (Shultz, Mysore, & Quartz, 2006). The many brain and neuronal features implemented in CC and other connectionist learning algorithms are detailed by Shultz (2003). This is not to say that CC mimics actual neural circuits, just that it is roughly compatible with current understanding of the principles of brain-style computation. In contrast, production systems and other symbolic computational methods are considered even by their proponents as purely functional models having no biological plausibility. Symbolic algorithms may work on some problems, although not apparently in the present case, but it is unknown how they might be implemented in real brains.

Even successful scientific models typically enjoy only a brief life before they are overturned by new evidence or replaced by better models that capture more phenomena in a more principled way. Indeed, precisely formulated, working models that generate testable predictions often speed the way towards better understanding of natural phenomena (Shultz, 2003). However, the evidence presented here suggests that the life of the Shultz and Bale (2001) model in covering the Marcus et al. (1999) infant data is not yet over. Moreover, this model looks fairly healthy in a field of connectionist models, some of whose results cannot be replicated, and one symbolic rule-based model that cannot learn the correct rules or cover any of the infant phenomena.

There are a number of more general lessons that the present controversy provides. One is that it can be surprisingly difficult to replicate computer simulations. Vilcu and Hadley (2001, 2003, 2005) reported that they were unable to replicate the basic results of two of the connectionist simulations of the Marcus et al. infant data. Notice also that we could not replicate the results of some of the Vilcu and Hadley simulations. Although it is not surprising that human or animal behavioral results do not always replicate, one might

have thought that this would not be a problem with computer simulations because of their mathematical and computational precision. Apparently simulations are not immune from replication problems. Perhaps researchers should start replicating simulations just as they already do with behavioral studies. Modelers could routinely run substantial numbers of networks in each condition and report statistical analyses of all of the networks they run, not just the successful ones. In this context, it should not be forgotten that there are several other unstructured network simulations of the Marcus et al. data that have not been shown to be difficult to replicate (Christiansen & Curtin, 1999; Negishi, 1999; Sirois, Buckingham, & Shultz, 2000). These simulations as well as the Marcus et al. (1999) infant data do not have either a published replication or a replication failure at this point.

Another important lesson of this exercise is to avoid introducing inadvertent confounds that obscure interpretation of results. In this case, changing phonemes in one syntactic type of test sentence, but not in the other syntactic type, confounds syntax and sound in a way that can invalidate evidence of syntactic discrimination. With such confounds, it cannot be determined whether a difference or lack of difference is due to variation in syntactic type or in speech sounds. Conventional experimental design takes care to eliminate such confounds (Campbell & Stanley, 1963), as was done in the Marcus et al. (2001) infant experiments and in the simulations reported here.

Yet another lesson is that, even with computer simulations, it is important to use statistical tests to evaluate the significance and reliability of results. Particularly with neural network models, with their stochastic properties and individual differences, such tests are critical. It is not always sufficient to rely on visual comparisons of means and visual plots or verbal descriptions of correlations. Important correlation coefficients should be computed and tested for statistical significance, and it should be remembered that a significant negative correlation is not the same as an absence of correlation.

In reviewing this work, Vilcu and Hadley requested that we remind readers on the substantial sample sizes they employed. Vilcu and Hadley's (2005) sample sizes were quite large. But of course large samples are not an adequate substitute for tests of statistical significance. Regardless of sample size, we cannot know whether mean differences are different from chance variation without tests of significance. ANOVAs, such as we used here, take account of variance within conditions (also known as error variance), as well as means and sample sizes, to determine whether mean differences are significantly larger than those differences that could be expected by chance alone (e.g., Winer, 1962). Major segments of the field of statistics address proper techniques for doing just that, and should not be ignored by experimenters.

Rather than simply striving for large sample sizes, we feel it is better to match sample sizes in simulations to those in the psychology experiment being simulated in order to equate statistical power as we do here. It is well known that increasing sample size increases the likelihood of statistical significance. Or to put it another way, the smaller the sample size, the more difficult it is to reach statistical significance. From a simulation point of view, it is more interesting to reach about the same level of statistical significance using the same sample size as in the experiment being simulated. In this way, statistical analysis can become part of the simulation.

Acknowledgments

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada to the first author.

References

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum

Baluja, S., & Fahlman, S. E. (1994). Reducing network depth in the cascade-correlation learning architecture. Technical Report CMU-CS-94-209, School of Computer Science, Carnegie Mellon University

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally

Christiansen, M. H., & Curtin, S. L. (1999). The power of statistical learning: No need for algebraic rules. *Proceedings of the twenty-first annual conference of the Cognitive Science Society* (pp. 114-119). Mahwah, NJ: Erlbaum

Fahlman, S. E., & Lebiere, C. (1990). The cascade-correlation learning architecture. In D. S. Touretzky (Ed.), *Advances in neural information processing systems 2* (pp. 524-532). Los Altos, CA: Morgan Kaufmann

Gómez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences, 4*, 178-186

Jolliffe, I. T. (1986). *Principle component analysis*. Berlin: Springer Verlag

Kaye, J. (1989). *Phonology: A cognitive view*. Hillsdale, NJ: Erlbaum

Marcus, G. F. (2001). *The algebraic mind*. Cambridge, MA: MIT Press

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science, 283*, 77-80

Mehler, J., Dommergues, J., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior, 20*, 298-305

Negishi, M. (1999). Do infants learn grammar with algebra or statistics? *Science, 284*, 433

Pinker, S. (1997). *How the mind works*. New York: Norton

Pinker, S. (1999). Out of the minds of babes. *Science, 283*, 40-41

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann

Rivest, F., & Shultz, T. R. (2005). Learning with both adequate computational power and biological realism. *Proceedings of the 2005 Canadian artificial intelligence conference: Workshop on correlation learning* (pp. 15-23). Victoria, BC: University of Victoria

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old-infants. *Science, 274*, 1926-1928

Sanger, D. (1989). Contribution analysis: A technique for assigning responsibilities to hidden units in connectionist networks. *Connection Science, 1*, 115-138

Selkirk, E. O. (1984). On the major class features and syllable theory. In M. Aronoff & R.T. Oehrle (Eds.), *Language sound structure* (pp. 107-136). Cambridge, MA: MIT Press

Shultz, T. R. (1999). Rule learning by habituation can be simulated in neural networks. *Proceedings of the twenty-first annual conference of the Cognitive Science Society* (pp. 665-670). Mahwah, NJ: Erlbaum

Shultz, T. R. (2001). Assessing generalization in connectionist and rule-based models under the learning constraint. *Proceedings of the twenty-third annual conference of the Cognitive Science Society* (pp. 922-927). Mahwah, NJ: Erlbaum

Shultz, T. R. (2003). *Computational developmental psychology*. Cambridge, MA: MIT Press

Shultz, T. R. (2006). Constructive learning in the modeling of psychological development. In Y. Munakata & M. H. Johnson (Eds.), *Processes of change in brain and cognitive development: Attention and performance XXI* (pp. 61-86). Oxford: Oxford University Press

Shultz, T. R., & Bale, A. C. (2000). Infant familiarization to artificial sentences: Rule-like behavior without explicit rules and variables. *Proceedings of the twenty-second annual conference of the Cognitive Science Society* (pp. 459-463). Mahwah, NJ: Erlbaum.

Shultz, T. R., & Bale, A. C. (2001). Neural network simulation of infant familiarization to artificial sentences: Rule-like behavior without explicit rules and variables. *Infancy*, *2*, 501-536

Shultz, T. R., Mysore, S. P., & Quartz, S. R. (2006, in press). Why let networks grow? In D. Mareschal, S. Sirois, & G. Westermann (Eds.), *Constructing cognition: Perspectives and prospects*. Oxford: Oxford University Press

Shultz, T. R., Oshima-Takane, Y., & Takane, Y. (1995). Analysis of unstandardized contributions in cross connected networks. In D. Touretzky, G. Tesauro, & T. K. Leen (Eds.), *Advances in neural information processing systems 7* (pp. 601-608). Cambridge, MA: MIT Press

Sirois, S., Buckingham, D., & Shultz, T. R. (2000). Artificial grammar learning by infants: An auto-associator perspective. *Developmental Science, 4*, 442-456

Vilcu, M., & Hadley, R. F. (2001). Generalization in simple recurrent networks. *Proceedings of the twenty-third annual conference of the Cognitive Science Society* (pp. 1072-1077). Mahwah, NJ: Erlbaum

Vilcu, M., & Hadley, R. F. (2003). Two apparent "counterexamples" to Marcus: A closer look. *Proceedings of the twenty-fifth annual conference of the Cognitive Science Society* (pp. 1188-1193). Mahwah, NJ: Erlbaum

Vilcu, M., & Hadley, R. F. (2005). Two apparent "Counterexamples" to Marcus: A closer look. *Minds and Machines*, *15*, 359-382

Vroomen, J., & de Gelder, B. (1997). Activation of embedded words in spoken word recognition. *Journal of Experimental Psychology, 23*, 710-720

Vroomen, J., van den Bosch, A., & de Gelder, B. (1998). A connectionist model for bootstrap learning of syllabic structure. *Language and Cognitive Processes, 13*, 193-220

Winer, B. J. (1962). *Statistical principles in experimental design*. New York: McGraw-Hill