Generalization in a Model of Infant Sensitivity to Syntactic Variation

Thomas R. Shultz (thomas.shultz@mcgill.ca)
Department of Psychology and School of Computer Science, McGill University, 1205 Penfield Avenue
Montreal, QC H3A 1B1 Canada

Abstract

Computer simulations show that an unstructured neural-network model (Shultz & Bale, 2001) covers the essential features of infant differentiation of simple grammars in an artificial language, and generalizes by both extrapolation and interpolation. Other simulations (Vilcu & Hadley, 2003) claiming to show that this model did not really learn these grammars were flawed by confounding syntactic patterns with other factors and by lack of statistical significance testing. Thus, this model remains a viable account of infant ability to learn and discriminate simple syntactic structures.

One of the enduring debates in cognitive science concerns the proper theoretical account for human cognition. Should cognition be interpreted in terms of symbolic rules or subsymbolic neural networks? It has been argued that infants' ability to distinguish one syntactic pattern from another could only be explained by a symbolic rule-based account (Marcus, Vijayan, Rao, & Vishton, 1999). After being familiarized to sentences in an artificial language having a particular syntactic form (such as ABA), infants preferred to listen to sentences with an inconsistent syntactic form (such as ABB). The claim about the necessity of rule-based processing was promptly contradicted by a number of neural-network modelers, several of whom produced unstructured models that captured the basic finding of more interest in novel than familiar syntactic patterns (Altmann & Dienes, 1999; Elman, 1999; Negishi, 1999; Shultz, 1999; Shultz & Bale, 2001; Sirois, Buckingham, & Shultz, 2000).

However, Vilcu and Hadley (2001, 2003) reported that two of these simulations (Altmann & Dienes, 1999; Elman, 1999) could not be replicated. Vilcu and Hadley (2003) were able to replicate the results of one simulation (Shultz & Bale, 2001). But Vilcu and Hadley (2003) claimed that their extensions of this model failed to generalize, both in terms of interpolation within the training range and extrapolation outside of this range. They concluded that this model did not really learn the grammars.

The present paper contains new simulations establishing that this model (Shultz & Bale, 2001) does indeed learn the simple grammars used in the infant experiments, interpolating and extrapolating successfully.

The Original Simulations

Shultz and Bale (2001) used an encoder version of the cascade-correlation (CC) learning algorithm to simulate the infant data. CC is a constructive algorithm for learning from examples in feed-forward neural networks (Fahlman & Lebiere, 1990). Being a constructive algorithm, CC builds its own network topology as it learns by recruiting new hidden units as needed. New hidden units are recruited one at a time and installed each on a separate layer. The candidate hidden unit that is recruited is the one whose activations correlate most highly with the current error of the network. CC has been used to simulate many aspects of psychological development (Shultz, 2003). For such developmental simulations, there are a number of advantages of constructive learning algorithms over static networks that only adjust connection weights, but do not grow during learning (Shultz, 2005a; Shultz, Mysore, & Quartz, 2005).

Like other encoder networks, the encoder version of CC learns to reproduce its inputs on its output units. Discrepancy between inputs and outputs is considered as error, which CC attempts to reduce. Infants are thought to construct an internal model of stimuli to which they are being exposed, and then differentially attend to more novel stimuli that deviate from their models. Encoder networks are capable of simulating this attention preference. Network error is often used as an index of stimulus novelty in these simulations.

The three-word sentences used in the infant experiments were coded by Shultz and Bale (2001) with a continuous sonority scale, shown in Table 1, based on previous phonological research (Vroomen, van den Bosch, & de Gelder, 1998). Sonority is the quality of vowel likeness and it has both acoustic and articulatory aspects. In Table 1, it can be seen that sonorities ranged from -6 to 6 in steps of 1, with a gap and change of sign between consonants and vowels.

Each word in the three-word sentences used in the infant experiments was coded on two input units for the sonority of the consonant and the sonority of the vowel. For example, the sentence *ga ti ga* was coded on the network inputs as (-5 6 -6 4 -5 6). The consonant /g/ was coded as -5, and the vowel /a/ as 6, yielding (-5 6) for the word *ga*, which was the first and last word in this sentence. The consonant /t/ was coded as -6, and the vowel /i/ was coded as 4, yielding a code of (-6 4) for the *ti* word. Likewise the sentence *ni ni la* was coded on the inputs as (-2 4 -2 4 -1 6).

The original simulation captured the essential features of the infant data including exponential decreases in attention to a repeated syntactic pattern, more interest in sentences inconsistent with the familiar pattern than in sentences consistent with that pattern, occasional familiarity preferences, more recovery to consistent novel sentences than to familiar sentences, and generalization both outside and inside of the range of the training patterns (Shultz & Bale, 2001).

Table 1: Phoneme sonority scale used in the original simulations.

Phoneme category	Examples	Sonority
low vowels	/a/ /æ/	6
mid vowels	/Є/ /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

Note. 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.

Interpolation

Vilcu and Hadley (2003) based their critique of the Shultz and Bale (2001) model on extended simulations that seemed to show that the model cannot actually interpolate or extrapolate. Interpolation is the ability to generalize within the range of the training patterns. Interpolation was tested by introducing a phonemic change to one of the four test patterns in each experiment. The original and modified test patterns are shown in Tables 2 and 3, respectively.

These tables also include sonority sums for these sentences, computed as the sonority of the consonant plus the sonority of the vowel. Knowledge representation analyses had established that CC encoder networks represented these single-syllable words by computing such sums, or equivalently by computing sonority differences (Shultz & Bale, 2001). For example, a network would learn to code the word *wo* as the sonority of the consonant /w/ plus the sonority of the vowel /o/: -1 + 5 = 4.

In Table 3, the syllables changed by Vilcu and Hadley (2003) are identified by a solid underline. Apparently their idea was to trip up the networks with a very small change of a single consonant; from *wo* to *vo* in the third test sentence of Experiment 1, and from *ba* to *ma* in the first test sentence of Experiments 2 and 3. With these changes, Vilcu and Hadley reported that networks could no longer distinguish consistent from inconsistent test patterns, although they did not report any testing of statistical significance.

However, by changing only one test pattern in each experiment, Vilcu and Hadley (2003) confounded phoneme with syntactic pattern. Shultz and Bale (2001) had followed the design of the infant experiments by using the same phonemes in both consistent and inconsistent test sentences. In Experiments 1 and 2, for example, two of the test sentences follow an ABA pattern and two follow an ABB pattern. Depending on condition, one of these syntactic patterns is consistent with those the infant is familiar with, whereas the other pattern is inconsistent. The syntactic patterns in Experiment 3 are slightly different, AAB vs. ABB, but the same principle holds there as well. It is important, whether testing or simulating infants, to use the same phonemes in both patterns so as not to confound phonemes with syntactic patterns. This is because both infants and artificial neural networks can be sensitive to both phonemic content and syntactic structure. When the two are confounded, results cannot be unambiguously interpreted as being due to one or the other factor.

Table 2: Original test patterns.

Experiment	Sentence	Sonority sums
1	wo fe wo	4 1 4
	de ko de	0 -1 0
	wo fe fe	4 1 1
	de ko ko	0 -1 -1
2	ba po ba	1 -1 1
	ko ga ko	-1 1 -1
	ba po po	1 -1 -1
	ko ga ga	-1 1 1
3	ba ba po	1 1 -1
	ko ko ga	-1 -1 1
	ba po po	1 -1 -1
	ko ga ga	-1 1 1

Table 3: Modified test patterns.

Experiment	Sentence	Sonority sums
1	vo fe vo	2 1 2
	de ko de	0 -1 0
	vo fe fe	2 1 1
	de ko ko	0 -1 -1
2	ma po ma	4 -1 4
	ko ga ko	-1 1 -1
	ma po po	4 -1 -1
	ko ga ga	-1 1 1
3	ma ma po	4 4 -1
	ko ko ga	-1 -1 1
	ma po po	4 -1 -1
	ko ga ga	-1 1 1

In the present simulations, I eliminated Vilcu and Hadley's confound by extending the same phonemic change to the other syntactic pattern in each experiment, marked in Table 3 by dashed underlines. In Experiment 1, for example, I used *vo fe vo* as well as *vo fe fe*. In Experiment 2, I tested *ma po po*, as well as *ma po ma*. And in Experiment 3, I included *ma po po* as well as *ma ma po*. These additional changes ensure that comparisons across syntactic patterns reflect only syntactic differences and not phonemic differences. Once the confounding is removed, there are robust differences between consistent and inconsistent test patterns as in the original simulations. In each experiment, with eight networks per condition as with the infant experiments, consistent test patterns showed less error than did inconsistent test patterns ($p < .0001$).

Apparently reasoning along similar lines, Vilcu and Hadley (2003) reported a simulation in which they changed /f/ to /b/ in both the first ABA test sentence and the first ABB test sentence of Experiment 1. Their networks failed to

discriminate consistent test sentences from inconsistent test sentences, but again no statistical significance test was provided. I repeated that simulation with eight networks per condition and did find 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 ($M = 9.30$).

Moreover, I could not replicate Vilcu and Hadley's (2003) finding of a lack of discrimination between consistent and inconsistent test patterns even using their single-pattern changes that confound phoneme with syntactic pattern. In each of the three experiments, run with 20 networks per condition to increase statistical power, there was less network error to consistent test patterns than to inconsistent test patterns, $p < .0001$.

These new results contradict Vilcu and Hadley's (2003) claim that the Shultz and Bale (2001) networks do not interpolate successfully. With properly controlled tests, interpolation ability is reliable and strong. Moreover, even with the confounding introduced by Vilcu and Hadley, the networks still interpolate well. The lack of statistical analysis in Vilcu and Hadley's research may have obscured differences between familiar and novel test patterns.

Extrapolation

To test the for extrapolation outside of the sonority training range in the Shultz and Bale model, Vilcu and Hadley (2003) assigned four consonant values beyond the minimum value of -6 (i.e., -7, -8, -9, -10), and combined them with two vowel values beyond the maximum value of 6 (i.e., 7, 8). They found that networks showed more error to consistent test patterns than to inconsistent test patterns, a direction opposite to that of both the infants and the networks. However, once again Vilcu and Hadley did not test the statistical reliability of this difference.

A major problem with testing outside the sonority range is that such extreme values do not correspond to the sounds in human languages. Testing network generalization in this way is thus somewhat irrelevant to simulations of psychology.

Furthermore, in arguing that networks fail to extrapolate beyond the training range, Vilcu and Hadley ignored the Shultz and Bale (2001) results showing that with less extreme deviations beyond the limits 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) sonorities. Here I report on a replication of the Shultz and Bale extrapolation results and extend the study to the more extreme sonority values used by Vilcu and Hadley (2003).

The sonority values I used are shown in Tables 4 and 5, along with a reminder of the original training anchor values used by Shultz and Bale (2001). As in Shultz and Bale (2001), I included 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). There were three additional sonority values ranging farther outside of the training range in steps of +-1.0, labeled in Tables 4 and 5 as *farther*, *even farther*, and *farthest*. The farthest values were as far outside the training range as the most extreme values used by Vilcu and Hadley (2003).

In one set of simulation experiments, portrayed in Table 4, the highest vowel sonority was paired with the lowest consonant sonority, keeping the sonority sums for syllables at a constant value of 0.0 in Pattern A and 3.0 in Pattern B. These simulations were characterized by a negative correlation between consonant and vowel sonority values, which can be seen in Table 4 by ignoring the first, anchor row.

Table 4: Test patterns for evaluating extrapolation in the simulation of Experiment 1: Highest vowel paired with lowest consonant and vice versa.

	Pattern A		Pattern B	
Distance	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 farth. +-3.0	-9.0	9.0	2.0	1.0
Farthest +-4.0	-10.0	10.0	3.0	0.0

In another set of simulations, shown in Table 5, the vowel columns in Table 4 were switched, pairing the highest vowel with the highest consonant. Here the correlation between consonant and vowel sonority values is positive, which can be seen in Table 5 by ignoring the first, anchor row. In this set of simulations, the sonority sums of the syllables were allowed to vary with distance from the training range. Both sets of simulations focused on Experiment 1 and used eight networks per condition as in the infant study. It was unclear whether these two different pairing methods for creating test patterns would produce different results, so it seemed appropriate to run the simulations both ways.

Table 5: Test patterns for evaluating extrapolation in the simulation of Experiment 1: Highest vowel paired with highest consonant and vice versa.

	Language A		Language B	
Distance	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 farth. +-3.0	-9.0	1.0	2.0	9.0
Farthest +-4.0	-10.0	0.0	3.0	10.0

In each of the two simulations, test error 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 as well as an interaction between them, $p < .0001$. The relevant means are presented in Figures 1 and 2 for constant and varying sonority sums, respectively. Note that extrapolation

is involved at all distances from the training range except for the condition labeled *inside*. As in the Shultz and Bale (2001) simulations, error increased with distance from the training range, error was greater to inconsistent than to consistent test patterns at each distance, and the consistency effect was larger with increasing distance.

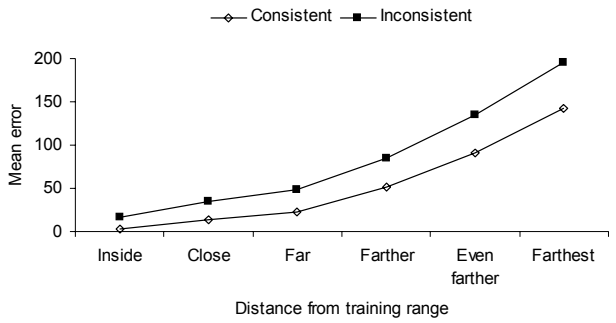

Figure 1: Mean error to consistent and inconsistent test patterns at various distances from the training range, where sonority sums were constant.

These results indicate that the Shultz and Bale networks interpolate and extrapolate very well. Error increases with distance outside the training range because the networks do not recognize the particular novel phonemes and syllables being presented. But even with very novel sounds, the networks are sensitive to the relative syntactic novelty of the sentences. As noted, outside of the range of human speech sounds, it is difficult to design realistic tests of the model's predictions, but the present results provide in-principle evidence of network extrapolation ability.

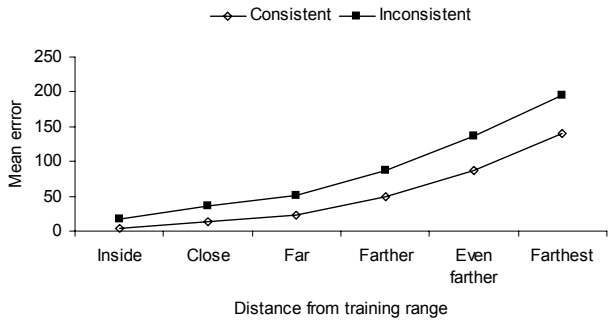

Figure 2: Mean error to consistent and inconsistent test patterns at various distances from the training range, where sonority sum was allowed to vary.

Discussion

Vilcu and Hadley's (2003) critique of neural-network models of infant learning of artificial grammars is important because it addresses a debate that has dominated cognitive science for the last 20 years – whether human cognition is better explained by symbolic rules or subsymbolic connections. They focused on the Shultz and Bale (2001) model because that is the simulation they could replicate that covered the Marcus et al. infant findings. Because Vilcu and Hadley's extensions of the Shultz and Bale model failed to generalize to novel sentences, in terms of both interpolation and extrapolation, they concluded that this model does not really learn the grammars.

However, results presented here showed that, by not confounding phoneme and syntactic pattern as in Vilcu and Hadley's experiments, there was robust interpolation and extrapolation, in the form of reliable differences between consistent and inconsistent test patterns. Moreover, even with Vilcu and Hadley's confounds left in, these effects were still reliable by conventional statistical tests. The inadvertent experimental confounds and lack of statistical significance tests in Vilcu and Hadley's research appeared to obscure reliable differences between test patterns, thus leading to underestimation of network ability to learn these simple grammars. If generalization by interpolation and extrapolation is the *sine qua non* for grammar learning, then these networks did learn these simple grammars.

To be fair and complete, Vilcu and Hadley (2003) raised another argument against these network models besides the alleged generalization difficulties. They also argued that the Shultz and Bale (2001) networks only learned the numerical contours of the artificial sentences, and not the syntactic relations involving the duplicated words. Vilcu and Hadley supported this argument with a simulation in which sonority contours of the familiar ABA sentences always formed a peak, whereas sonority contours of the test sentences could form either a peak or a valley. When sonority contours of test sentences formed a peak, then there was the usual novelty preference; but when sonority contours of the test sentences formed a valley, then there appeared to be a familiarity preference. Again, there were no tests of statistical significance.

Although this appears to suggest that networks are sensitive to input contours and not syntax, it ignores the fact that, in both the infant experiments and the Shultz and Bale (2001) simulations, sonority contours were balanced within each language rather than confounded with syntax. The contours of these familiar sentences were not simply sonority peaks or sonority valleys as Vilcu and Hadley (2003) suggested, but rather a complex combination of sonority-sum contours containing peaks, valleys, and plateaus in Experiments 1 and 2, and increases, decreases, and plateaus in Experiment 3 (Shultz & Bale, 2005b).

In the ABA familiarization condition of Experiment 1, eight of the training sentences formed a sonority-contour peak and the other eight formed a sonority-contour valley (Figure 3). In the ABB condition of that experiment, eight of the familiar sentences showed an increasing sonority contour and the other eight showed a decreasing sonority contour (Figure 4). The same was true of Experiment 2 except that two of the ABA (Figure 5) and two of the ABB (Figure 6) familiar sentences showed a completely flat sonority profile. Both the AAB (Figure 7) and ABB (Figure 6) familiar sentences of Experiment 3 had a similar mix of sonority contours: seven had an increasing contour, seven a decreasing contour, and two a flat profile. For simplicity, Figures 3-7 all show schematic sonority profiles.

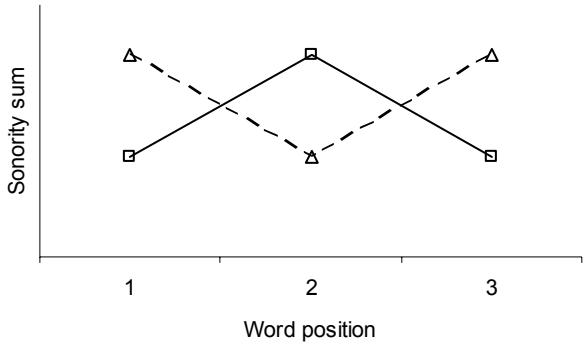

Figure 3: Sonority profiles in the training patterns of the ABA familiarization condition of Experiment 1.

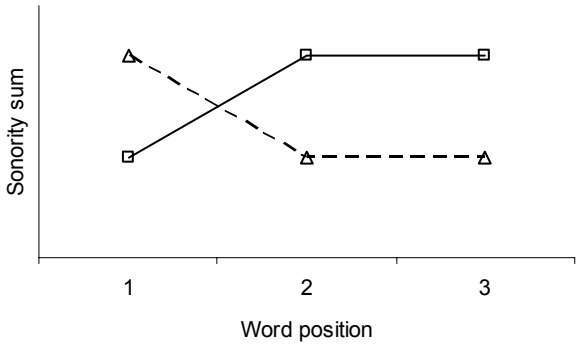

Figure 4: Sonority profiles in the training patterns of the ABB familiarization condition of Experiment 1.

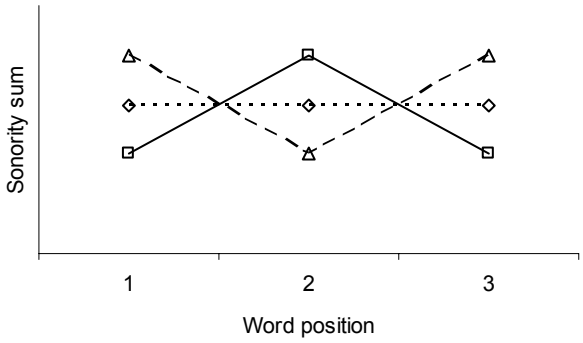

Figure 5: Sonority profiles in the training patterns of the ABA familiarization condition of Experiment 2.

To deal with this complex mix of contours, the networks (and presumably the infants) discovered near-identity relations to differentiate the syntactic patterns of old vs. new sentences. In none of these experiments was it sufficient to learn a single sonority profile as suggested by Vilcu and Hadley.

It is unknown how infants would perform in an experiment with Vilcu and Hadley's confounds between syntactic pattern and sonority contour in familiar sentences, but infants might well be sensitive to sonority contours. Sonority contours might help the infant identify syllable boundaries which might, in turn, facilitate word identification. If so, this would be a difficult pattern of results for a purely syntactic, symbolic model to account for.

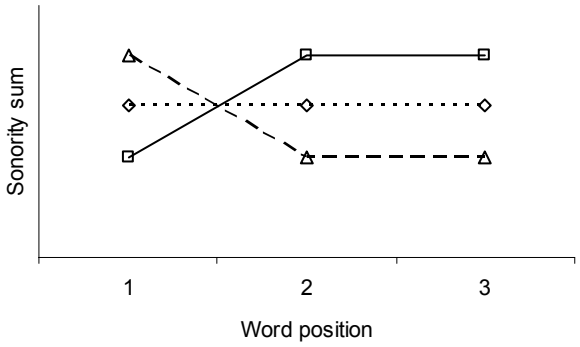

Figure 6: Sonority profiles in the training patterns of the ABB familiarization conditions of Experiment 2 and 3.

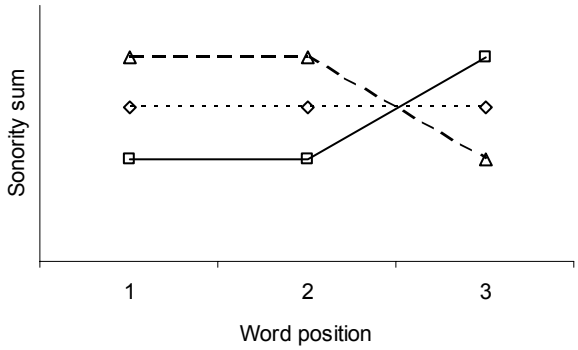

Figure 7: Sonority profiles in the training patterns of the AAB familiarization condition of Experiment 3.

Networks learned to decode representations of the two duplicate words in a sentence by using highly similar sets of weights entering the output units that represent the duplicate words (Shultz & Bale, 2001, 2005b). This virtually identical pattern of weights entering the output units representing the duplicate words allowed the network to recognize the near identity of the duplicate words.

The relatively large connection weights to the duplicated-word outputs from the first hidden unit showed that this hidden unit recognized the category (A or B) of these duplicate words. The second hidden unit performed the complementary job 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. This means that the duplicated words were being treated in similar fashion.

Additional analysis of network knowledge representations used principle-component analyses of network

contributions. Network contributions are products of sending-unit activations and connection weights entering the output units. They effectively summarize all of the information used by the network to generate its outputs (Shultz, Oshima-Takane, & Takane, 1995). This analysis revealed two components, one representing sonority variation in the duplicate-word category and the other representing sonority variation in the single-word category. This provides additional evidence that the networks learned to treat the duplicate words in a nearly identical fashion.

All of this is not to say that these rather simple networks could acquire the full grammar of a human language. It is certain that some aspects of human syntactic acquisition would require different and more powerful models. But the ability of the Shultz and Bale (2001) 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 algorithm, C4.5 (Shultz, 2001).

An interesting feature of this controversy is that it can be surprisingly difficult to replicate computer simulations. Vilcu and Hadley (2003) were unable to replicate the results of two other connectionist simulations of the infant data. Also, the present paper reveals that I could not replicate the results of some of the Vilcu and Hadley simulations. It is commonplace that human or animal results cannot always be replicated, but the notion that replication can be a problem with computer simulations seems novel. The mathematical and computational precision of these models have led many to assume that replication of results would not be a problem. The numerous non-replications uncovered in this relatively small literature suggest that researchers should perhaps replicate simulations routinely. In this context, it should be remembered that several other unstructured network simulations of the infant data have not been shown to be difficult to replicate (Negishi, 1999; Shultz, 1999; Sirois, Buckingham, & Shultz, 2000).

Another important lesson of this exercise is that, even with computer simulations, it is important to use statistical tests to evaluate the significance and reliability of results. Such tests are particularly critical with neural network models, because of their stochastic properties. It is not always sufficient to rely on visual comparisons of means.

Acknowledgments

Thanks to the Natural Sciences and Engineering Research Council of Canada for financial support and to Alan Bale for helpful discussions.

References

Altmann, G. T. M., & Dienes, Z. (1999). Rule learning by seven-month-old infants and neural networks. *Science, 284*, 875.

Elman, J. L. (1999). *Generalization, rules, and neural networks: A simulation of Marcus et al.* [Online]. Retrieved April 27, 1999 from the World Wide Web: http//www.crl.ucsd.edu/~elman/Papers/MVRVsim.html

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.

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

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

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. (2005a, in press). 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*. Oxford: Oxford University Press.

Shultz, T. R., & Bale, A. C. (2005b). *Neural networks discover a near-identity relation to distinguish simple syntactic forms*. Submitted for publication.

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. (2005, 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.

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.