A Model of Infant Learning of Word Stress

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

LouAnn Gerken (gerken@u.arizona.edu)
Departments of Psychology and Linguistics, University of Arizona
Tucson, AZ 85721-0068 USA

Abstract

Nine-month-old infants can distinguish the word-stress patterns of two artificial languages after a few minutes of exposure to words from one of the languages, apparently by making transitive inferences from known word-stress constraints to unknown constraints. We report on a neural-network simulation of these data using the sibling-descendant cascade-correlation algorithm. The simulations cover the infant data and generate some predictions for further infant research.

Introduction

Learning how to stress the syllables in words is an important part of acquiring a language, easily distinguishing skilled from novice speakers. Word stress is an interesting problem to study because it has been given a fairly complete linguistic description, it can be studied independently of other aspects of language, and it has attracted competing computational models using principles and parameters on the one hand (Dresher & Kaye, 1990) and neural networks on the other (Gupta & Touretzky, 1994).

Remarkably, infants as young as 9 months can learn something about novel word-stress patterns with a few minutes of exposure to artificial words (Gerken, 2004). Such infant research, perhaps coupled with computational modeling, has the potential to uncover some of the fundamental properties of the human language-learning system (Gómez & Gerken, 2000; Shultz, 2003). This paper reports on artificial-neural-network simulations of these infant data. We first review the infant research, then discuss the properties of our computational model, and finally present three simulations.

Evidence with Infants

Gerken (2004) asked whether 9-month-olds could learn an optimality-theoretic stress system that had been used in an adult production study by Guest, Dell, and Cole (2000). The logic of the adult experiment was to expose learners to multisyllabic nonwords whose stress patterns provided evidence for several rankings of stress constraints. In optimality theory (Prince & Smolensky, 1997), different stress-assignment constraints can conflict in their application to a particular word. When two constraints do conflict, only the more highly ranked applies. Importantly, in the Guest et al. (2000) and Gerken (2004) experiments, one ranking was not attested in the initial input, but it could be inferred from attested rankings, based on transitivity. For example, if learners have evidence that constraint A outranks (>>) constraint B and B >> C, they should be able to infer that A >> C. Adults showed evidence of accepting test words reflecting the unattested ranking from their training grammar, while rejecting very similar words that reflected a different grammar (Guest et al., 2000).

To determine if infants would behave similarly, 18 nine-month-olds were familiarized for two minutes to three- and five-syllable words (Gerken, 2004). The stress patterns in these spoken words conformed to either Language 1 (n = 9; Table 1) or Language 2 (n = 9; Table 2). In these tables, syllables in upper case are stressed, whereas those in lower case are unstressed. Seven variants of each familiarization and test word were created using the seven solfège syllables (*do, re, mi, fa, so, la, ti*) and substituting *re* for *do*, *mi* for *re*, etc., in the example words of Tables 1 and 2. No substitutions were made for the syllable *ton*.

Table 1: Example words and constraint rankings for infants familiarized to L1.

Familiarization	Attested ranking
TON ton do RE mi	A >> B
TON do re	B >> C_1
DO re TON	B >> C_1
DO re TON mi fa	B >> C_1
DO re mi FA so	C_1 >> D_1
L1 test	Inferred ranking
do TON re MI fa	A >> D_1

Table 2: Example words and constraint rankings for infants familiarized to L2.

Familiarization	Attested ranking
do RE mi ton TON	A >> B
do re TON	B >> C_2
TON do RE	B >> C_2
do re TON mi FA	B >> C_2
do RE mi fa SO	C_2 >> D_2
L2 test	Inferred ranking
do RE mi TON fa	A >> D_2

The following four word-stress constraints, all of which are typical of constraints in natural languages, were used in the study:

A. Two stressed syllables cannot be adjacent.
B. Heavy syllables (i.e., those ending in a consonant) are stressed.
C. Syllables are stressed if they are second to last (C_1 in L1) or second (C_2 in L2).
D. Alternating syllables are stressed, starting from the left (D_1 in L1) or right (D_2 in L2).

The familiarized words provide evidence for three rankings of stress principles. In L1, A >> B, B >> C_1, and C_1 >> D. For example, the word *TON do re* in L1 attests that constraint B (heavy syllables are stressed) outranks constraint C_1 (syllables are stressed if they are second to last). Likewise, in L2, A >> B, B >> C_2, and C_2 >> D. No direct evidence was provided that A >> D_1 (for L1) or A >> D_2 (for L2). However, these unattested rankings could be inferred using transitivity from the attested rankings.

During testing, infants heard on different trials words with new stress patterns that were consistent with A >> D_1 or A >> D_2 inferences. Note that L1 and L2 test items had the same stress pattern (second and fourth syllables stressed) and differed only in the location of the heavy syllable TON. Therefore if infants can discriminate the two types of test items, they presumably did so by making inferences across words in their familiarization language.

Infants were tested in a head-turn preference procedure, and their looking times revealed a significant familiarization-language x test-language interaction, $F(1, 16) = 7.78$, $p < .02$. Infants familiarized with L1 listened longer to L2 test words and vice versa, as shown in Figure 1. These data, which have been replicated at least once, $F(1, 16) = 9.97$, $p < .01$ (Gerken, 2004), suggest that infants were able to make inferences across multiple words in their familiarization language, and on that basis could distinguish the stress patterns of the two languages.

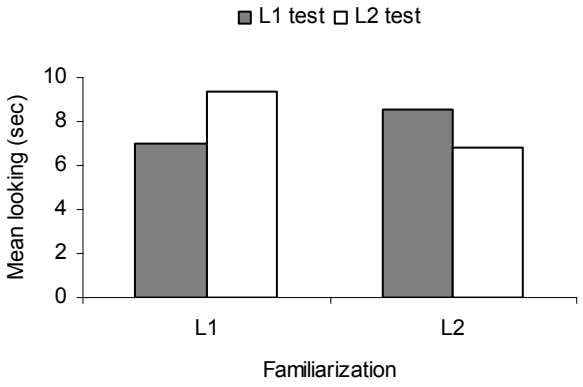

Figure 1: Infant interest in hearing word-stress patterns in two languages.

Properties of Our Computational Model

In familiarization experiments like those of Gerken (2004), it is assumed that infants build categories for repeated stimuli and then subsequently ignore stimuli corresponding to their categories and concentrate instead on stimuli that are relatively novel. Such shifts of attention would be of obvious adaptive value in promoting cognitive and linguistic development. The building of categories for familiar stimuli is typically discussed in terms of recognition memory. If a stimulus is recognized as a member of a familiar category, then it often elicits less attention than a stimulus not recognized as familiar.

Recently it has become possible to simulate such processes with artificial neural networks. One of the most effective techniques employs feed-forward encoder networks (Mareschal & French, 2000; Mareschal, French, & Quinn, 2000; Shultz & Bale, 2001; Shultz & Cohen, 2004). In such networks, stimulus features are represented as real numbers in an input vector, then encoded in representations on a relatively small number of hidden units, and finally decoded on an output vector. Discrepancy between output and input representations is computed as network error. Familiar stimuli produce less error than novel stimuli, and this extra error indicates that the novel stimuli merit further processing and perhaps learning. Such encoder networks are able to generalize, abstract prototypes, and complete input patterns that are missing components. After training, error on familiar stimuli is typically less than that on novel stimuli. Various learning algorithms have been employed for this purpose, including back-propagation (BP) with static networks and cascade-correlation (CC) with constructive networks (Shultz, 2003).

Here we report on our efforts to simulate Gerken's (2004) experiments on word stress with a variant of CC called sibling-descendant cascade-correlation (SDCC; Baluja & Fahlman, 1994). Like CC, SDCC recruits new hidden units as it needs them to learn. But unlike CC, which installs each new hidden unit on a separate layer, SDCC determines whether it is better to install each new recruit on the current highest layer of hidden units (as a sibling) or on its own new layer (as a descendant). Algorithms that construct their own networks, like CC and SDCC, have been found better at simulating a variety of phenomena in psychological development than algorithms like BP that merely learn to adjust connection weights in a static, pre-designed network topology (Shultz, 2005). The mathematics underlying CC and SDCC can be found elsewhere (Shultz, 2003).

Simulation 1: Familiarization x Test-pattern Interaction

The principal aim of the present simulations was to capture Gerken's (2004) familiarization-language x test-language interaction with the tests discussed earlier. To this end, we created seven examples of each of five training-word types, totaling 35 training patterns and seven examples of test words, just as in the infant experiments.

Coding of Words

We coded the words used in Gerken's (2004) experiments on a sonority scale, shown in Table 3. This scale is based on phonological research (Vroomen, van den Bosch, & de Gelder, 1998) and has been used to simulate other infant experiments with artificial-language stimuli (Shultz & Bale, 2001). Sonority is the quality of vowel likeness and it has both acoustic and articulatory aspects. As can be seen in

Table 3, the sonority scale ranges from -6 to 6 in steps of 1, with a gap and change of sign between the consonants and vowels.

Table 3: Phoneme sonority scale.

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, 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.

Because words could have up to five syllables, with up to three phonemes per syllable, 15 units were required to code each word. The stress given to each of the potential five syllables was coded as -0.5 for unstressed and 0.5 for stressed. Twenty-five input units coded each word as follows, with subscripts 1-5 indicating the slots for *cv* or *cvc* syllables and the stress (*s*) each syllable received: cvc_1 cvc_2 cvc_3 cvc_4 cvc_5 s_1 s_2 s_3 s_4 s_5. Five-syllable words required all five syllable slots for both sonority and stress, but three-syllable words were coded only in slots 2-4. Missing final consonants and missing syllables were coded as 0.0. Twenty-five output units had this same structure. Codes for the phonemes of each syllable are shown in Table 4.

Table 4: Sonority codes for syllables.

Syllable	$Consonant_1$	Vowel	$Consonant_2$
do	-5.0	5.0	0.0
re	-1.0	5.0	0.0
mi	-2.0	4.0	0.0
fa	-4.0	6.0	0.0
so	-4.0	5.0	0.0
la	-1.0	6.0	0.0
ti	-6.0	4.0	0.0
ton	-6.0	5.0	-2.0

Procedures and Parameters

We ran nine networks in each familiarization condition in order to match the statistical power of the infant experiments. A training epoch is a single pass through all of the training patterns. SDCC networks alternate between two phases: output phase and input phase. In output phase, weights entering output units are adjusted in order to reduce network error. When error reduction stagnates, SDCC switches to input phase in order to recruit a new hidden unit.

In input phase, weights from inputs and existing hidden units to candidate recruits are adjusted in order to increase the size of a correlation between candidate-unit activation and network error. When those correlations stagnate, the unit with the largest absolute correlation is recruited, and the other candidates are discarded. The candidate pool contains equal numbers (4) of siblings and descendants, each with initially random connection weights from input units and any existing hidden units. As is customary with SDCC, recruitment of descendant candidate units was penalized by multiplying their correlations with error by a factor of 0.8 (Baluja & Fahlman, 1994).

We tested each network's performance every 25 output epochs. Training was stopped after 280 total epochs because pilot simulations suggested that this amount of training provided a good match to the size of the key interaction *F* ratio in Gerken's (2004) infant experiments. The simulation results reported here are quite robust in that they can be replicated over a wide range of maximum training epochs. Infant trials cannot be precisely equated to network epochs because it is unknown how much processing occurs during an infant trial.

Results

After 280 epochs, these networks recruited a mean of 6.2 hidden units on a mean of 1.3 layers. The final topology for a typical network with four hidden units on one layer and two hidden units on the next layer is shown in Figure 2. The arrows indicate full connectivity, with all of the units in one layer being connected to all of the units in the next layer. The bias unit always has an input value of 1 and is connected to all downstream units by trainable connection weights, thus establishing a learnable resting level of activation for each hidden and output unit. As is customary with encoder networks, there are no direct input-to-output connections here. This is to prevent trivial solutions in which a network might learn weights of about 1 from each input unit to its corresponding output unit. Such solutions would memorize the training patterns quickly but would not generalize well to untrained test patterns.

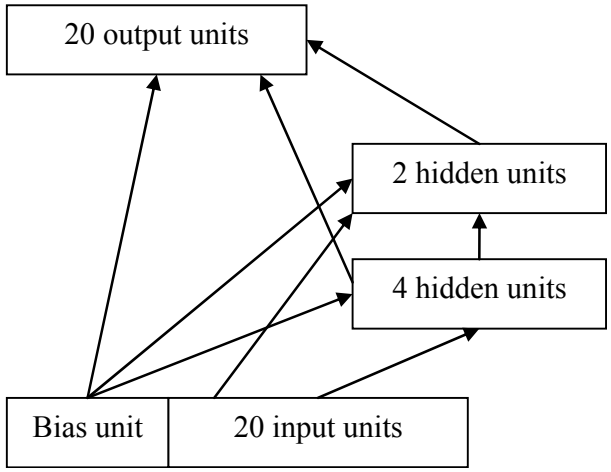

Figure 2: Final topology of an SDCC network.

As is typical of both infant attention to a repeated stimulus category and CC and SDCC simulations of infants in such experiments, network error decreased exponentially over time. Error reduction in a typical network is plotted in Figure 3.

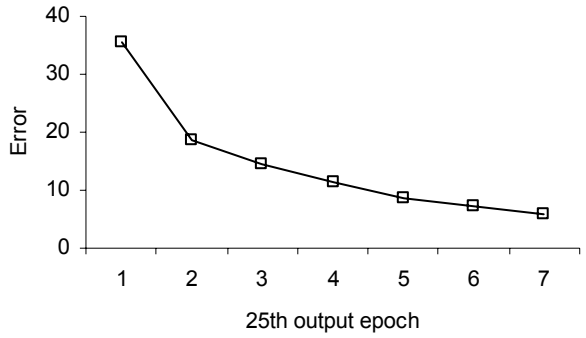

Figure 3: Error reduction in a representative network.

Network error after 280 epochs was subjected to a mixed ANOVA in which familiarization language served as a between-network factor and test language served as a within-network factor. The key interaction between the two was significant, $F(1, 16) = 15$, $p < .001$. The associated means are plotted in Figure 4. As with infants, there was more interest in the test language with novel stress patterns, signaled here by network error.

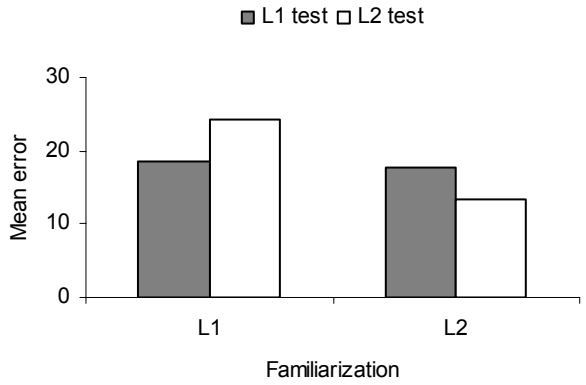

Figure 4: Network interest in word-stress patterns in two languages.

Simulation 2: Effects of Constraint Deletion

If the results of the infant experiments and Simulation 1 are due to transitive inferences about constraints, then it should be possible to produce different results depending on which constraints are deleted from the familiarization words. For example, deletion of one of the three B >> C familiarization word types should disrupt transitive inferences less than deletion of the only C >> D familiarization word type. The former deletions should not matter much because there are still two other word types present that attest to the B >> C constraint ranking. In contrast, the latter C >> D deletions might matter a lot because they would break the transitive inference chain required to make the A >> D inference.

To test this idea, we ran a simulation in which we deleted the third of the B >> C word types from each language for 18 networks in one condition, and the C >> D word type from each language for 18 networks in another condition. See Tables 1 and 2 for the particular word types deleted.

Network error was subjected to a mixed ANOVA in which deletion condition and familiarization language served as between-network factors and test language served as a within-network factor. With a significant deletion x familiarization-language x test-language interaction, $F(1, 32) = 17$, $p < .001$, we then analyzed the familiarization-language x test-language interaction for each deletion condition. The familiarization-language x test-language interaction was significant for the C >> D deletions, $F(1, 16) = 6.7$, $p < .03$, as well as the B >> C deletions, $F(1, 16) = 27$, $p < .001$. Interaction means are plotted in Figure 5 for the B >> C deletions, and in Figure 6 for the C >> D deletions. It is noteworthy that the variance associated with the familiarization-language x test-language interaction was 36 times larger for the B >> C deletions than the corresponding variance for the C >> D deletions, showing that the C >> D deletions were more disruptive to this key interaction than were the B >> C deletions, as expected.

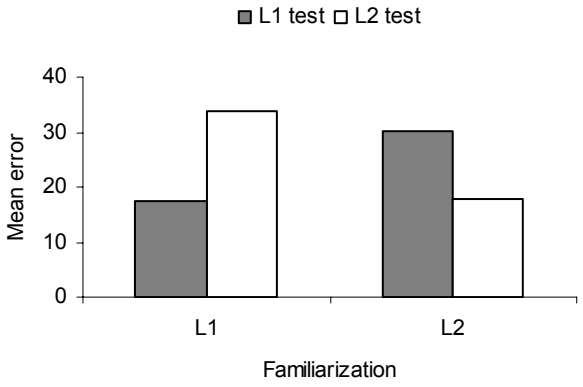

Figure 5: Network interest in word-stress patterns in two languages when a B >> C word type is deleted.

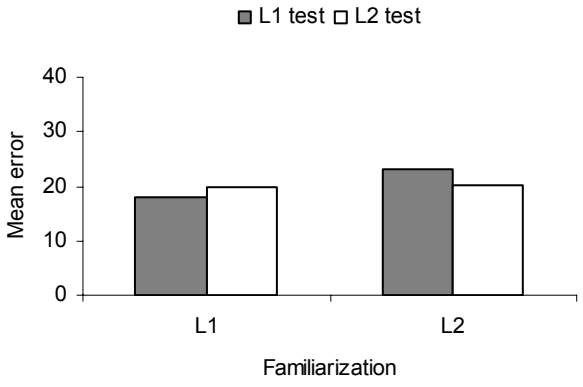

Figure 6: Network interest in word-stress patterns in two languages when the C >> D word type is deleted.

Thus, omitting a unique link in the transitive-inference chain disrupts the key familiarization-language x test-language interaction more than omitting a redundant link does.

Simulation 3: Position of a Heavy Syllable

A reviewer of the paper reporting the infant experiments (Gerken, 2004) suggested an alternative explanation for successful generalization to the test sentences. This reviewer noted that the heavy syllable TON occurs earlier in L1 familiarization words than in L2 familiarization words. These average positions happen to correlate with the positions of TON in the test sentences. TON is in second position in the test sentence for L1 and fourth position in the test sentence for L2. Perhaps infants generalized to the test sentences, not by performing transitive inference on constraint rankings, but rather by using the relative positions of the heavy syllable.

We tested this idea in neural networks by equating the position of TON in the two languages. We omitted the first B >> C familiarization word type from both languages and doubled the frequency of the second B >> C familiarization word types. As the reader can verify from Tables 1 and 2, this yields a mean serial position of 3.0 for TON in the familiarization words of each language (assuming that three-syllable words are coded on the middle-three banks of input and output units). With these changes to the familiarization languages, nine networks were run in each familiarization condition as in Simulation 1.

Network error was analyzed as in Simulation 1. The familiarization-language x test-language interaction was still present, $F(1, 16) = 38$, $p < .001$. Associated means are plotted in Figure 7. As far as networks are concerned, the serial position of the heavy syllable TON is irrelevant to the expected interaction between familiarization language and test language. Coupled with the results of Simulation 2, this suggests that, at least for networks, differential generalization to the familiar and novel languages arises, not from the position of the heavy syllable, but rather from transitive inferences across known constraints.

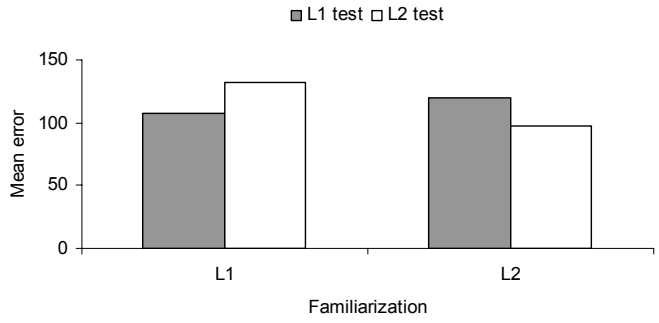

Figure 7. Network interest in word-stress patterns in two languages with equated serial positions of the heavy syllable.

Discussion

The infant results simulated here are quite remarkable when one considers the nature of the test words. The two test words have entirely different stress patterns than do the familiarization words in either L1 or L2. Moreover, the two test words exhibit the same stress pattern, differing only in the location of the heavy syllable TON. Therefore, the obtained familiarization x test interaction suggests that infants are able to generalize beyond stress patterns encountered during familiarization to the abstract system underlying these patterns. That a simple neural-network model, after being familiarized to words in the same fashion as the infants were, can also generalize in this manner is perhaps equally remarkable. How do they (infants or networks) do that, particularly on the basis of rather limited exposure to a few example words? Poverty-of-the-stimulus arguments might well be employed to support some sort of innate knowledge of word-stress rules (Dresher & Kay, 1990). We are still quite far from understanding how infants execute this skill, but it is somewhat easier to examine computational models to determine how they perform.

Although we have described the inferences on which this generalization is based in terms of transitive reasoning, it is noteworthy that SDCC networks do not perform any explicit logical reasoning on symbolic propositions. A simulation of transitive inference in older children and adults shows that CC networks, even without hidden units, can simulate such inferences in an entirely neural manner (Shultz & Vogel, 2004). An important computational trick used by those networks is to learn connection-weight strengths that effectively represent the linear order of the training stimuli. Such weights can then be used to make accurate comparisons of unattested stimulus pairs. In several ways, the word stimuli and training regime used in the present simulations are more complicated, but the ability of the learning algorithm to construct an essentially neural solution to a transitivity problem is comparable. We plan to explore the exact nature of this solution to word-stress assignments in a longer paper in which we perform detailed analyses of network knowledge representations.

As far as we are aware, no other computational models have yet been applied to these infant data. Of the two most prominent models in the area of word-stress learning, it is interesting that one is a symbolic model based on principles-and-parameters theory (Dresher & Kaye, 1990) and the other is a connectionist model using static BP networks (Gupta & Touretzky, 1994). One thing that these two, very different models share is that each learning algorithm is presented with stress-pattern information abstracted away from the actual phonemes in a word. As might be expected, such pre-processing of the input simplifies the learning task immensely. Speech sounds can be ignored in these models, enabling the learning algorithm to focus only on the pre-abstracted stress patterns, which are identical for all words of the same number of syllables in the language being learned. Although these models have a number of interesting features, we believe that our model is more realistic in view of the fact that both numerically-coded speech sounds and syllabic-stress information are included. This might enable our model to deal with anticipated effects

of phonological content on the learnability of stress patterns. Humans, whether infants or adults, are never presented with abstract stress patterns, only with streams of speech in which word syllables are stressed according to the stress syntax of the language.

We plan to implement and compare alternate models to our SDCC model in future work. The principles-and-parameters model (Dresher & Kaye, 1990) will be particularly interesting to study in this context because many of its predictions would be quite different from those that our model would make.

One current prediction of our model concerns the differential effects on test-word performance of deleting certain word-stress constraints from the familiarization words. Simulation 2 showed that deleting one of the three familiarization word types attesting to the B >> C constraint disrupted performance on the test words less than did deleting the familiarization word type containing the only evidence of the C >> D constraint. Although it is likely that a model using explicit transitive inference would predict something similar, its prediction would be more extreme than the one made by our model. This is because our model, even without the important C >> D familiarization information, still retained some capacity to make a correct transitive inference regarding the A >> D constraint. An explicit-reasoning model would likely generate no A >> D inference at all without some C >> D evidence. Like other neural-network models, ours predicts a graceful degradation of performance in the face of important evidential gaps, in contrast to the more brittle performance of an explicit-reasoning model.

Another prediction of our model ruled out an alternate hypothesis concerning the relative position of a heavy syllable in both familiarization words and test words. Even when mean serial positions of the heavy syllable were equated across the two languages, networks were still relatively more interested in the novel language, as revealed by a difference in mean network error. Predictions such as these should be interesting to test with infants.

Acknowledgments

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada to T. R. Shultz and by NIH grant R01HD42170 to R. Gómez, L. A. Gerken, and E. Plante. Thanks to Gabrielle Pagé, J.-P. Thivierge, and Frederic Dandurand for helpful comments on an earlier draft.

References

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.

Dresher, B., & Kaye, J. (1990). A computational learning model for metrical phonology. *Cognition, 34*, 137-195.

Gerken, L. A. (2004). Nine-month-olds extract structural principles required for natural language. *Cognition, 93*, B89-B96.

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

Gupta, P., & Touretzky, D. S. (1994). Connectionist models and linguistic theory: Investigations of stress systems in language. *Cognitive Science, 18*, 1-50.

Guest, D. J., Dell, G. S., & Cole, J. S. (2000). Violable constraints in language production: Testing the transitivity assumption of Optimal Theory. *Journal of Memory & Language, 42*, 272-299.

Mareschal, D., & French, R. M. (2000). Mechanisms of categorization in infancy. *Infancy, 1*, 59-76.

Mareschal, D., French, R. M., & Quinn, P. (2000). A connectionist account of asymmetric category learning in infancy. *Developmental Psychology, 36*, 635-645.

Prince, A., & Smolensky, P. (1997). Optimality: From neural networks to universal grammar. *Science, 275* (5306), 1604-1610.

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

Shultz, T. R. (2005, 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. (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., & Cohen, L. B. (2004). Modeling age differences in infant category learning. *Infancy, 5*, 153-171.

Shultz, T. R., & Vogel, A. (2004). A connectionist model of the development of transitivity. *Proceedings of the Twenty-sixth Annual Conference of the Cognitive Science Society* (pp. 1243-1248). 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.