Infancy, 1–16, 2010

ISSN: 1525-0008 print / 1532-7078 online
DOI: 10.1111/j.1532-7078.2009.00023.x

WILEY-BLACKWELL

Computational Modeling of Statistical Learning: Effects of Transitional Probability Versus Frequency and Links to Word Learning

Daniel Mirman
Moss Rehabilitation Research Institute
Albert Einstein Healthcare Network

Katharine Graf Estes
Department of Psychology
University of California

James S. Magnuson
Department of Psychology
University of Connecticut and Haskins Laboratories

Statistical learning mechanisms play an important role in theories of language acquisition and processing. Recurrent neural network models have provided important insights into how these mechanisms might operate. We examined whether such networks capture two key findings in human statistical learning. In Simulation 1, a simple recurrent network (SRN) performed much like human learners: it was sensitive to both transitional probability and frequency, with frequency dominating early in learning and probability emerging as the dominant cue later in learning. In Simulation 2, an SRN captured links between statistical segmentation and word learning in infants and adults, and suggested that these links arise because phonological representations are more

Correspondence should be sent to Daniel Mirman, Moss Rehabilitation Research Institute, 1200 W. Tabor Road, Philadelphia, PA 19141. E-mail: mirmand@einstein.edu

distinctive for syllables with higher transitional probability. Beyond simply simulating general phenomena, these models provide new insights into underlying mechanisms and generate novel behavioral predictions.

Studies with infants, children, and adults indicate that statistical learning is a mechanism available throughout life, which is capable of acting on many levels of linguistic structure, including phonemes (e.g., Maye, Weiss, & Aslin, 2008), syllables (e.g, Saffran, Newport, Aslin, Tunick, & Barrueco, 1997; Saffran & Wilson, 2003), and words (e.g., Saffran & Wilson, 2003; Thompson & Newport, 2007). Although statistical learning may be particularly useful for learning language, adults and infants can also learn statistical patterns in nonlinguistic sequences, such as tones (e.g., Saffran, Johnson, Aslin, & Newport, 1999) and shapes (e.g., Fiser & Aslin, 2002; Kirkham, Slemmer, & Johnson, 2002), suggesting that this is a general type of learning that can be found in many different domains. The power and breadth of statistical learning has led to the proposal that it is a key underlying mechanism for language acquisition.

Computational models provide a framework for establishing and refining theories of underlying mechanisms. They serve as existence proofs that show what kinds of mechanisms can give rise to observed behaviors and they make novel behavioral predictions for further tests of proposed cognitive mechanisms. The computational model simulations presented here explore two key findings in the statistical learning literature. The first simulation investigates how the model balances sensitivity to frequency versus transitional probability information when these two word segmentation cues conflict. The second simulation investigates the connection between statistical word segmentation and word learning, a connection recently demonstrated in studies with infants (Graf Estes, Evans, Alibali, & Saffran, 2007) and adults (Mirman, Magnuson, Graf Estes, & Dixon, 2008).

As a class, recurrent neural networks are known to exhibit sensitivity to transitional probabilities. Elman (1990) showed that when simple recurrent networks (SRNs) are trained to predict the next syllable in a sequence, they become sensitive to word boundaries based on transitional probabilities in input sequences, an analog of the statistical segmentation performance of infants, children, and adults (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996). Note that in such simulations, the SRN is not trained to perform word segmentation, nor is it explicitly trained to extract frequencies or transitional probabilities among input elements. Rather, sensitivity to informative statistics in the input is an emergent property of how the SRN solves the prediction task. Since their introduction, SRNs have provided computational insights into a wide range of domains where

statistical sequence learning plays a critical role, including word segmentation (e.g., Christiansen, Allen, & Seidenberg, 1998), spoken word recognition (e.g., Magnuson, Tanenhaus, Aslin, & Dahan, 2003), grammatical processing (e.g., Altmann, 2002; MacDonald & Christiansen, 2002), and short-term memory (e.g., Botvinick & Plaut, 2006). Thus, as suggested by Elman's initial simulations (see also Christiansen et al., 1998), SRNs provide a natural account of statistical word segmentation, and have the potential to provide insight into aspects of statistical learning that are not yet well understood. In particular, the balance of frequency and transitional probability information and the connection between statistical learning and word learning have not been addressed from a computational perspective. These topics provide two critical tests of SRNs as models of human language learning and processing and, if the model correctly accounts for the basic behavioral results, provide a concrete framework for developing and refining theories of human language processing.

Simulation 1 examined whether SRNs predict a crucial detail in human statistical learning: learners' sensitivity to transitional probability versus frequency patterns. Aslin, Saffran, and Newport (1998) reported that following exposure to a speech stream containing only transitional probability cues to word boundaries (high transitional probabilities within words and low probabilities across words), 8-month-old infants discriminated between test items that occurred with equal frequency in the speech stream, but differed in their internal transitional probability. Graf Estes et al. (2007) extended this finding in a word learning task. These results indicate that infants are sensitive to transitional probabilities of syllable sequences, not merely transition frequencies. Transition frequency (i.e., bisyllable frequency) can be high simply due to a high rate of occurrence of the individual elements (e.g., a syllable transition that occurs at the boundary between two very frequent words), but transition probability depends on the relationship between the elements (i.e., within-word transitions have higher probability than between-word transitions even for high frequency words); thus, transitional probabilities are more informative than frequencies about the structure of the language. The first simulation tested whether SRNs exhibit sensitivity to both frequency and transitional probability and how these effects interact throughout learning when placed in competition. Constructing situations in which cues conflict provides a powerful test of computational mechanisms because the results reveal which of the conflicting cues influences processing to a greater extent and how cues might interact (e.g., Johnson & Jusczyk, 2001; Shukla, Nespor, & Mehler, 2007; Thiessen & Saffran, 2003). Sensitivity to transitional probability rather than frequency provides a critical test of the validity of SRNs as plausible models of statistical mechanisms supporting human language learning.

Simulation 2 focused on results from recent behavioral findings showing that learners take advantage of transitional probability information to support word learning. Graf Estes et al. (2007) and Mirman et al. (2008) investigated the connection between statistical word segmentation and learning of new object labels in infants and adults, respectively. In these studies, each participant was first exposed to a nonsegmented syllable stream as in typical statistical segmentation studies, followed by an object label learning task. The infants participated in a habituation-based label-object association task (Stager & Werker, 1997; Werker, Cohen, Lloyd, Casasola, & Stager, 1998), and the adults participated in an artificial lexicon learning procedure (Magnuson et al., 2003). In the infant and adult experiments, one group of participants heard labels that consisted of high probability syllable transitions from the segmentation stream (*word* labels); a second group heard labels that consisted of low probability syllable transitions (*partword* labels, an exposure sequence that straddled a word boundary, with one syllable from one word and one from another); and a third group heard labels composed of syllable sequences that did not occur during the statistical exposure phase (*nonword* labels). Both studies demonstrated a connection between statistical segmentation and word learning. Infants learned the word labels, but not the partword or nonword labels, and adults learned all three types of labels, but they learned the word and nonword labels faster than the partword labels.

These studies are of particular importance because they establish a link between statistical learning and referential word learning (but see Endress & Mehler, 2009 for an alternative view). Given that SRNs have been used to model statistical learning (e.g., Christiansen et al., 1998) and word learning (e.g., Magnuson et al., 2003), a natural question is whether SRNs might also provide insight into this link. In Simulation 2, we tested whether the way an SRN learns transitional probabilities affects learning of word, partword, and nonword labels, and explored the causes of transitional probability effects on label learning. As we shall see, the simulations provide new insights into aspects of statistical learning and generate new predictions.

SIMULATION 1: FREQUENCY VERSUS TRANSITIONAL PROBABILITY

The first simulation addressed the relative influences of frequency and transitional probability on statistical learning. Aslin et al. (1998) showed that infants distinguish between high and low transitional probability syllable sequences, even when they occur with equal frequency in a speech stream. Perruchet and Peereman (2004) found that an SRN's syllable prediction accuracy precisely matched transitional probability and, to a substantially

lesser extent, frequency. Their data suggest that SRNs are more sensitive to transitional probability than frequency, but they did not investigate this issue directly nor whether sensitivity to these cues changes over the course of learning. To investigate how SRNs balance sensitivity to frequency and transitional probability, the simulated language exposure placed transitional probability and frequency cues in conflict: partword sequences (sequences that straddle a word boundary) occurred with greater frequency than the word sequences. Although SRNs are known to be sensitive to both frequency and transitional probability, it is not clear how the model will respond when these cues are in competition, and such conflict cases are particularly informative tests of a system (e.g., Thiessen & Saffran, 2003).

Model architecture and simulation design

All simulations were carried out using Lens software (version 2.63).[1] The model architecture is shown in Figure 1 and followed a standard SRN design. The input fed forward to hidden units, and hidden units fed forward to output units. Context units contained a copy of the hidden unit activations, which served as additional input to the hidden units on the next model time step. The context units comprise the recurrence in the network and allow it to use information from previous time steps to perform its task. An SRN model's ''memory span'' depends on the number of hidden units and the structure of the input. The input and output layers contained 12 units and the hidden and context layers contained 15 units. Each of the input and output units represented a unique ''syllable''[2] (i.e., a localist representation of syllables). A set of four ''words'' was created by randomly concatenating three of the 12 syllables to form each word. The words were then concatenated into a continuous stream. Two of the words occurred with high frequency (400 times), and two occurred with low frequency (200 times). For each of the 3,600 syllables in the stream, the model was trained to activate the next syllable in the sequence. Connection weights were initialized with random values generated from a uniform distribution ranging from −.1 to .1. The network was trained using backpropagation (Rumelhart, Hinton, &

[1]The ''light, efficient neural simulator'' developed by Doug Rohde, see http://tedlab.mit.edu/~dr/Lens/. Script and example files are available from the first author.

[2]Since our goal was to examine the computational mechanisms involved in statistical learning, we chose simple localist input and output representations to maximize model tractability rather than complex representations intended to capture the acoustic or phonological reality of speech input. Note also that the abstract localist input representation can equally be considered a representation of phonemes, syllables, bisyllables, or any other unit over which the transitional probabilities are defined. We refer to the units as ''syllables'' simply because statistical learning studies have typically manipulated transitional probabilities at the syllable level.

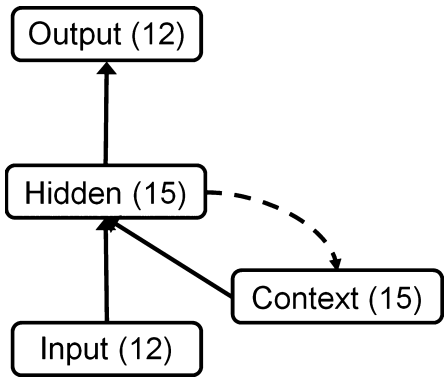

Figure 1 Simulation 1 model architecture. Solid arrows indicate fully connected, trainable weights among units; the dashed arrow indicates copy connections. Numbers of units in each layer are shown in parentheses.

Williams, 1986) to predict the next syllable in the sequence with learning rate set to .03 and momentum set to .9.

The model was tested on predicting the next syllable in the four words and four partwords. Partwords were created in two ways: (1) concatenating the last syllable of one high frequency word and the first two syllables of the other high frequency word ($A_3B_1B_2$); and (2) concatenating the last two syllables of one high frequency word and the first syllable of the other high frequency word ($A_2A_3B_1$). The partwords occurred 229 or 241 times in the stream,[3] which is 17.5% more frequently than the low frequency words. However, the transitional probability for the words was 100% for all within-word syllable transitions; for partwords, it was only 58.75% at the across-word transitions (e.g., for $A_3B_1B_2$ words, the transitional probability from A_3 to B_1). Thus, partwords had a frequency advantage, but low frequency words had a transitional probability advantage, allowing us to compare the relative importance of these factors. Note that the manipulation of transitional probability was substantially stronger than the manipulation of frequency (approximately 1:1.7 versus 1.2:1). This issue is discussed alongside the relevant results below.

When the test patterns were presented to the model, error was evaluated for predicting the next syllable in the test pattern. That is, on the first time step, the input was the first syllable of a test pattern and the target was the second syllable, then the second syllable was presented, and the target was the third syllable of the test pattern. The model was trained for 200 cycles

[3]The small difference in frequency between partwords had no interesting effects, so the results were collapsed across "partword" frequency.

through the full stream, and tested after every 10 cycles (no weights were changed during the test sessions). After 200 training cycles, the model activated the correct output unit for each syllable (activation $> .98$) and deactivated all other output units (activation $< .01$).

Results and discussion

The model's learning performance was evaluated using cross-entropy error, a measure of the difference between actual output layer activation and target (i.e., correct) output activation. Figure 2 shows that, over the course of statistical learning, error for predicting the second and third syllables decreased for all test stimulus types and the decrease was fastest for high frequency words. Interestingly, early in learning, the higher frequency, lower transitional probability partwords had a strong advantage (lower error) relative to lower frequency, higher transitional probability words. However, this pattern reversed as statistical learning progressed (after approximately 150 training cycles), indicating a difference in the model's learning of frequency and transitional probability information. Figure 3 shows error by stimulus type and position within the word, which reveals the model's sensitivity to transitional probability in greater detail. For both low and high frequency words, error was approximately equal across syllable positions because transitional

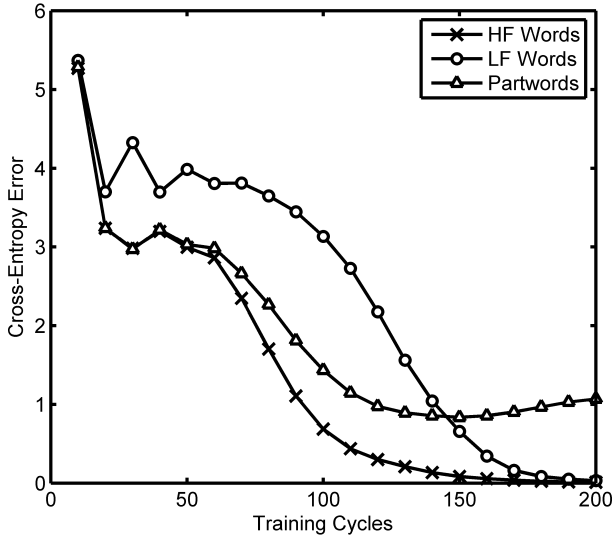

Figure 2 Simulation 1 cross-entropy error for the three conditions over the course of learning.

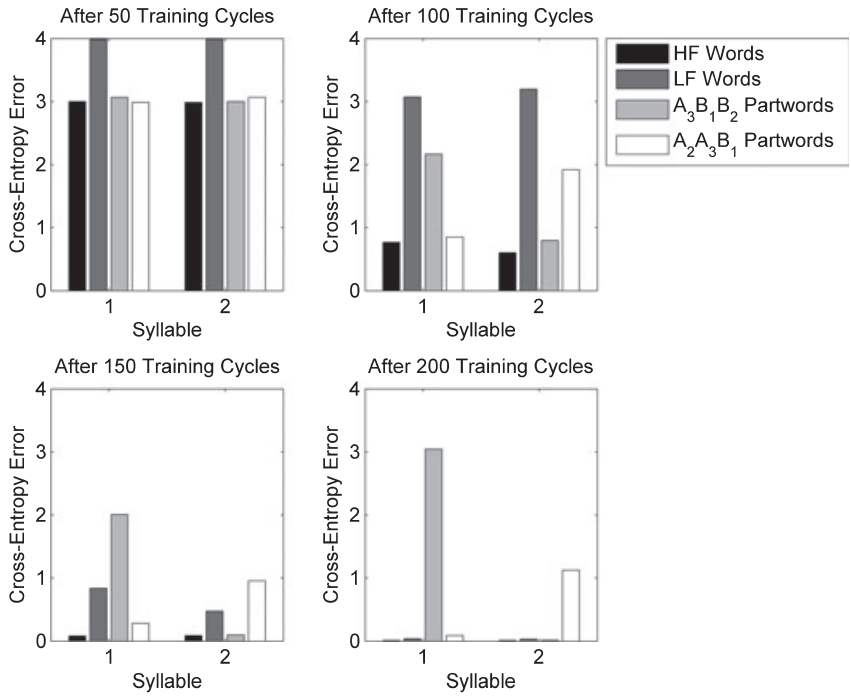

Figure 3 Simulation 1 cross-entropy error by item type and syllable position at select points during training.

probability (and frequency) was equal across syllable positions. For partwords, error was higher at the low probability transition than at the high probability transition. That is, for $A_3B_1B_2$-type partwords, error was higher on the first syllable (the A_3 to B_1 transition) than the second syllable (the B_1 to B_2 transition), and for $A_2A_3B_1$-type partwords, error was lower on the first syllable (the A_2 to A_3 transition) than the second syllable (the A_3 to B_1 transition). The model's sensitivity to frequency is also evident in Figure 3. After 50 and 100 training cycles, the error is higher for low frequency words than for partwords because the partwords have slightly higher frequency.

This pattern reveals that SRNs are sensitive to both frequency and transitional probability. Frequency is initially easier to learn, but probability is ultimately more powerful. As mentioned above, the manipulation of transitional probability was substantially stronger than the manipulation of frequency. Despite this disadvantage, the frequency difference dominated network performance early in learning, and it was eventually overtaken by transitional probability. A stronger frequency manipulation would likely delay this cross-over, but there is no clear way this cross-over pattern could

have arisen as a consequence of the asymmetry in manipulation strength. Note that this pattern emerged because the connection weights in an SRN are structured by the statistical properties of the input and the informational demands of the task (in this case, predicting the next input). There was no explicit tracking of frequency or transitional probabilities and no stipulated rule for balancing these sources of information. The model's ultimate reliance on transitional probability matches human data (e.g., Aslin et al., 1998) and the model predicts that human learners should show greater sensitivity to the frequency of a transition than its probability early in training. Interestingly, Toro and Trobalon (2005) reported that rodents seem to learn frequencies, but not transitional probabilities.[4] This simulation suggests that Toro and Trobalon may have stopped training during the frequency-dominant stage of learning. Rodents may be able to learn transitional probabilities if given more exposure. Conversely, humans may exhibit greater sensitivity to transition frequency than transitional probability early in learning.

SIMULATION 2: THE LINK BETWEEN STATISTICAL SEGMENTATION AND WORD LEARNING

Recent studies have shown that both infants and adults are better at learning novel object labels when the labels consist of high probability syllable transitions (the "words" in the exposure phase) rather than low probability syllable transitions ("partwords"; Graf Estes et al., 2007; Mirman et al., 2008). Following a statistical exposure phase, infants learned statistically defined word labels more readily than nonword or partword labels. Adults learned word labels more quickly than partword labels and learning was equally fast for word and nonword labels. To our knowledge, no formal account of the effect of statistical learning on word learning has been proposed. We conducted simulations of an SRN model to test the effects of statistical word segmentation on object label learning and to develop a formal account of these effects.

Model architecture and simulation design

Figure 4 shows the modified SRN architecture of the model. Each of the 20 input units and the 10 output units represented a unique syllable. Ten of the

[4]Nonhuman primates have been shown to learn transition frequencies (Hauser, Newport, & Aslin, 2001), but the materials used in that experiment did not manipulate transition probability independently of transition frequency.

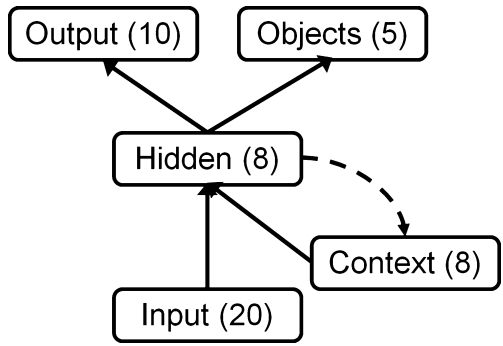

Figure 4 Simulation 2 model architecture. Solid arrows indicate fully connected, trainable weights among units; the dashed arrow indicates copy connections. Numbers of units in each layer are shown in parentheses.

20 possible input syllables were presented during the exposure phase. The exposure phase was modeled on the statistical exposure phase of the behavioral experiments (i.e., Graf Estes et al., 2007; Mirman et al., 2008): 100 repetitions of five two-syllable "words" were concatenated into a pseudorandom continuous syllable sequence (with the constraint that no word occurred twice in succession). For each syllable in the input sequence, the target output pattern was the next syllable in the sequence. The model was trained for 75 runs through this 1,000-syllable sequence with learning rate set to .05 and momentum set to .9. After 75 runs, the model activated the correct output syllable unit for each word more than any other output unit.

After the exposure phase, the model was trained on an analog of the label learning task. The model was trained (using backpropagation with the same learning rate and momentum parameters) to activate a unique "object" output unit for each of five different two-syllable input patterns. The labels were either words (100% probability transitions), partwords (25% probability transitions), or nonwords (0% transitions). We tested two types of nonwords: *Novel-sequence nonwords* were composed of syllables from the exposure phase presented in novel pairings; *Novel-syllable nonwords* were composed of the 10 input syllables that did not occur during the exposure phase. The behavioral experiments used nonword syllable sequences that did not occur in the exposure stream, but because they contained familiar native-language syllables they were not entirely novel either. In order to examine the effects of syllable familiarity on label learning, we compared learning of labels composed of completely novel syllables (novel-syllable nonwords) versus labels composed of familiar syllables presented in novel sequences (novel-sequence nonwords).

During the label learning phase, the model was trained for 100 presentations of each of the five labels. For maximum balance between the label learning conditions, the same postexposure weights were used at the start of each label learning phase (i.e., the model started with exactly the same weights at the beginning of the label learning phase for words, partwords, etc.). The full two-phase simulation was repeated 10 times with different random initial weights to verify that the results did not depend on idiosyncratic initial conditions.

Results and discussion

Figure 5 shows the mean object label learning (cross-entropy error) curves for each label condition. After a brief initial period, the error for word labels was clearly lower than the error for partword labels. That is, the model was better at mapping two-syllable input sequences to unique objects when the sequences had high probabilities in the preceding exposure phase compared to when they had low probabilities. This word advantage over partwords is consistent with the behavioral results observed for infants (Graf Estes et al., 2007) and adults (Mirman et al., 2008). Novel-syllable nonword labels were learned more slowly than word, partword or novel-sequence nonword labels. By contrast, novel-sequence nonword labels were initially learned nearly as fast as word labels, up to an intermediate point in training as shown in the right panel of Figure 5, after which the learning of novel-sequence nonwords more closely matched the learning of partword labels.

We examined the learned hidden representations in order to understand why the model performed better on the novel-sequence nonwords compared

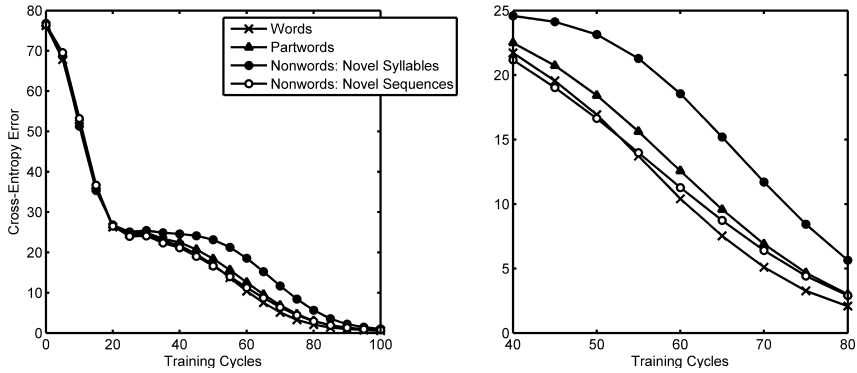

Figure 5 Simulation 2 label learning results. Left panel shows results from the full course of learning, right panel is an enlarged portion at an intermediate time range to highlight differences between the nonword conditions.

to both the novel-syllable nonwords and the partwords (at least initially). Specifically, we examined the distance between hidden layer representations of the labels for each label condition using cosine distance between hidden layer activation patterns. The analysis indicated that the difference in learning the two types of nonwords arose because exposure to the input representations allowed the model to build up distinct hidden unit representations for each syllable. Compared to novel syllables, syllables presented during the exposure phase had more distinct hidden layer representations, thus they were easier to map to objects (that is, to act as labels) even though they occurred in novel combinations when acting as labels. This is analogous to a simple familiarity effect, exposure to a pattern facilitated learning for that pattern in a subsequent task.

Examining the learned hidden representations for the novel-sequence nonwords versus the partwords further explained the difference in performance. During the exposure phase, each word-initial syllable had 100% transitional probability to a unique following syllable; that is, each word-initial syllable input pattern had a fully consistent target output pattern (since the model was trained to perform syllable prediction). By contrast, word-final syllables were not as predictive: each could be followed by four different word-initial syllables (i.e., the onsets of the other four words), each at 25% transitional probability, so the word-final input syllables had a $1 \rightarrow 4$ input $\rightarrow$ target relationship. In addition, each of these 25% probability (word-initial) target syllables occurred after four different word-final input syllables; that is, there was a $4 \rightarrow 1$ input $\rightarrow$ target relationship for each target pattern. As a result, the learned hidden unit representations for the first syllables were highly distinct, but the representations for second syllables were much more similar. By design, partwords all started with a second syllable. By contrast, some novel-sequence nonwords started with a first syllable; these nonwords were learned more quickly and were responsible for the nonword labels' advantage over the partword labels. Note that the critical issue here is predictive power, not syllable position. First syllables had distinct representations because they had highly predictive one-to-one relationships with the syllables that followed them. By contrast, second syllables had less predictive many-to-many relationships with the syllables that followed them. In natural speech, final syllables generally have low power to predict the following syllable; indeed, the very idea of word segmentation by transitional probabilities rests on this pattern. The key finding here is that high transitional probability leads to representations that are more distinct, which facilitates learning of label-object pairs. That is, the model provides a formal account of *why* infants and adults are better at learning labels with high transitional probability, because syllables involved in higher transitional probability sequences have more distinct phonological representations.

The simulation results produced the insight that transitional probability improves label learning by making underlying representations more distinct, which makes novel predictions and may help to explain the difference in nonwords label learning between adults and infants. Adults learned partword labels more slowly than word labels and nonword labels as quickly as word labels (Mirman et al., 2008), but, infants only learned word labels and failed to learn partword and nonwords labels (Graf Estes et al., 2007). Nonwords form an intermediate baseline case between partwords and nonwords; they possess neither the benefit of high transitional probabilities nor the cost of low transitional probabilities. The difference between adult and infant learning of nonword labels may be a result of the level of this baseline. The label learning task was relatively easy for adults (all conditions reached 90% accuracy within 20 training trials per word), thus the baseline case may be near a ceiling-level label learning rate and the clearest effect of transitional probability would be the cost of low transitional probability for partwords. By contrast, the label learning task was relatively difficult for infants (unlike studies using a similar paradigm [e.g., Werker et al., 1998], Graf Estes et al., 2007 found that without statistical exposure, infants were unable to learn the novel bisyllabic object labels); thus, the nonword baseline case may be near floor-level learning performance and the clearest effect of transitional probability would be the benefit of high transitional probability for words. If, as the simulations suggest, the distinctiveness of underlying representations is the critical factor, then adults may exhibit better learning of word labels than nonword labels if phonologically similar syllables are used (so that the phonological representations are less distinct) or if the labels are composed of novel elements (such as unfamiliar nonnative language syllables or novel nonspeech sounds) so that learners can not take advantage of prior phonological knowledge. For infants, greater exposure to the segmentation language, and the opportunity to develop distinct representations, should lead infants to exhibit better learning of nonword labels than partword labels.

In sum, Simulation 2 showed that an SRN can account for reported links between statistical learning and word learning. The simulation also indicated that the distinctiveness of learned hidden representations provides a plausible (and testable) basis for these links. Note that the SRN was designed to predict the next syllable in a sequence, not to transfer statistical segmentation knowledge to word learning, nor to show a difference between nonwords types. Rather, both of these effects are emergent properties of the nature of learning and representation in such models. Furthermore, examination of the model's learned hidden representations provided novel insights into why human listeners exhibit these behavioral patterns in statistical learning tasks.

CONCLUDING REMARKS

Simulations using SRNs addressed two important issues in statistical learning. First, an SRN exhibited sensitivity to transitional probability beyond transition frequency as observed in human subjects, providing a crucial test of the validity of SRNs as candidate models for human statistical learning. The SRN also showed a learning time course difference between sensitivity to frequency and sensitivity to transitional probability. Frequency dominated performance early in training, and transitional probability dominance emerged later in training. That is, transitional probability was the stronger cue, but more training was required to use it. This leads to the novel prediction that with less exposure, humans may exhibit sensitivity to frequency rather than transitional probability.

Second, an SRN model showed that, similar to human listeners, statistically learned transitional probabilities affected word learning. Examination of learned hidden representations revealed that links between statistical learning and object label learning are due to increased syllable familiarity and distinctiveness in phonological representations. This account generates the novel prediction that the impact of exposure statistics on word learning should interact with the familiarity and phonological similarity of the syllables. Thus, SRNs provide a promising test bed for further examining the link between statistical segmentation and referential word learning.

These simulations reveal new findings regarding the timecourse of statistical learning processes and provide insight into the basis for the link between statistical learning and word learning. They also illustrate the utility of computational modeling for testing and refining theories of language acquisition. Computational models demonstrate to what extent a set of theoretical principles can account for observed behavioral data, elucidate what aspects of those principles are critical for the account, and make novel predictions.

ACKNOWLEDGMENTS

This research grew from discussions at the Workshop on Current Issues in Language Acquisition: Artificial Languages and Statistical Learning (held in June, 2007, in Calgary, Alberta, Canada) and we are grateful to Suzanne Curtin and Dan Hufnagle for organizing the Workshop. This research was supported by NICHD NRSA F32HD052364 to DM, NIDCD grant R01DC005765 to JSM, and by NICHD grant HD01994 to Haskins Laboratories.

REFERENCES

Altmann, G. T. M. (2002). Learning and development in neural networks—The importance of prior experience. *Cognition*, *85*, B43–B55.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*, 321–324.

Botvinick, M. M., & Plaut, D. C. (2006). Short-term memory for serial order: A recurrent neural network model. *Psychological Review*, *113*, 201–233.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, *13*, 221–268.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Endress, A. D., & Mehler, J. (2009). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language*, *60*, 351–367.

Fiser, J. Z., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 458–467.

Graf Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science*, *18*, 254–260.

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition*, *78*, B53–B64.

Johnson, E., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, *44*, 548–567.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence of a domain general learning mechanism. *Cognition*, *83*, B35–B42.

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review*, *109*, 35–54.

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., & Dahan, D. (2003). The time course of spoken word learning and recognition: Studies with artificial lexicons. *Journal of Experimental Psychology: General*, *132*, 202–227.

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science*, *11*, 122–134.

Mirman, D., Magnuson, J. S., Graf Estes, K. G., & Dixon, J. A. (2008). The link between statistical segmentation and word learning in adults. *Cognition*, *108*, 271–280.

Perruchet, P., & Peereman, R. (2004). The exploitation of distributional information in syllable processing. *Journal of Neurolinguistics*, *17*, 97–119.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing* (Vol. 1, pp. 318–362). Cambridge, MA: MIT Press.

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

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*, 27–52.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–621.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science*, *8*, 101–105.

Saffran, J. R., & Wilson, D. P. (2003). From syllables to syntax: Multilevel statistical learning by 12-month-old infants. *Infancy*, *4*, 273–284.

Shukla, M., Nespor, M., & Mehler, J. (2007). An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology*, *54*, 1–32.

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*, 381–382.

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology*, *39*, 706–716.

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, *3*, 1–42.

Toro, J. M., & Trobalon, J. B. (2005). Statistical computations over a speech stream in a rodent. *Perception & Psychophysics*, *67*, 867–875.

Werker, J. F., Cohen, L. B., Lloyd, V. L., Casasola, M., & Stager, C. L. (1998). Acquisition of word-object associations by 14-month-old infants. *Developmental Psychology*, *34*, 1289–1309.