Online processing is essential for learning: Understanding fast mapping and word learning in a dynamic connectionist architecture

Jessica S. Horst (jessica-horst@uiowa.edu)
Bob McMurray (bob-mcmurray@uiowa.edu)
Larissa K. Samuelson (larissa-samuelson@uiowa.edu)
Department of Psychology, E11 Seashore Hall
Iowa City, IA 52242 USA

Abstract

The acquisition of word meaning is often partially attributed to fast mapping. However, recent research suggests that fast mapping and word learning may represent distinct components of language acquisition. Here we examine their interaction with a Hebbian Normalized Recurrence Network, a connectionist architecture that captures both online processing and long-term statistical learning. After training on a small lexicon, the model performed well above chance on a fast mapping task. Careful analyses of the weight changes, however, suggest that the fast mapping task can be solved with minimal learning. Thus, this model not only captures both long-term learning and online processes, but also provides unique insights regarding the relationship between fast mapping and word learning and that the two should be carefully distinguished.

Fast Mapping and Word Learning

In early childhood, children learn words at very impressive rates; typically saying their first word between 10-14-months-of-age, around 300 words by their second birthday and over 60,000 words by their fifth birthday (Carey, 1978). Young children also solve word learning problems with remarkable ease. When presented with two familiar objects, one novel object and a novel name, a child can correctly pick the novel object as the referent of a novel name (Carey, 1978; Carey & Bartlett, 1978). This ability, known as "fast mapping" is often cited as evidence of children's word learning proficiency. However, recent empirical and computational findings suggest that although fast mapping may be important to word learning, selecting a novel object in response to a novel name is not coextensive with learning the novel name.

In the literature, fast mapping and word learning are often discussed as if they are the same thing (see, for example, Behrend, Scofield, & Kleinknecht, 2001). We argue, however, that fast mapping and word learning represent distinct, though subtly related components of word recognition and learning. First, the child must select the referent in the moment (fast mapping). Second, the child must encode the name-object mapping such that it can be recalled after a delay (complete word learning). The extant literature reinforces this distinction as studies either test children's ability to make the initial mapping, or to remember mappings after a delay, typically with some additional review (see, Horst & Samuelson, submitted). While we have argued that these are potentially logically independent processes (Horst & Samuelson, submitted), the present computational model suggests a subtle integration. Specifically, the model confirms that fast mapping and word learning represent two distinct, but related, time scales and that fast mapping can, in fact, lead to word learning. Thus, we argue that fast mapping and word learning are *related* but not *coextensive*, as previously argued in the literature.

In making the initial mapping, the child is faced with a very specific problem: given a series of objects and a name, the child must determine which object should be associated with that name. In other words, which object is the referent of the name. To solve this problem, the child engages in probabilistic constraint satisfaction. That is, the child determines the most probable and optimal solution to this problem, given present constraints. In the case of making the initial mapping, the constraints include both the present input, which are the objects and name presented, and the child's own developmental history, which includes a vocabulary of known names. Thus, the problem solving required for fast mapping is done in real-time as the child determines the most likely solution (object) to a specific problem (unknown referent). Further, the real-time aspect of fast mapping does not necessarily require learning on the part of the child. However, as we will show, the repetition of these real-time dynamics change the constraints for the next time the novel referent is presented and eventually, with enough repetitions, fast mapping can lead to word learning.

Thus, the process of quickly mapping a novel name to a referent emerges in the moment, while the process of encoding a robust representation of this link unfolds over a longer time scale (for a similar argument see also Capone & McGregor, in press; Carey, 1978). Recent empirical data underscores this hypothesized distinction between fast referent mapping and word learning.

Supporting Empirical Data

Horst & Samuelson (submitted) found that twenty-four-month-old children were able to fast map as many as eight new words in a single session, but did not retain these words over a five minute delay. On each fast mapping trial, children were presented with two familiar objects and one novel object. On half of the trials they were asked to get a familiar object (e.g., "can you get the car?"), on the other half of the trials were asked, (e.g., "can you get the blicket?"). Overall, children were excellent at selecting both the familiar and novel referents (see Figure 1A). Five minutes later children were tested for retention of the fast-

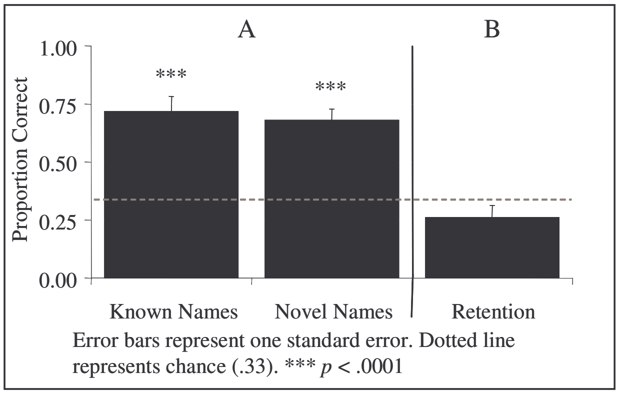


Figure 1: Results from Horst & Samuelson (submitted) Experiment 1.

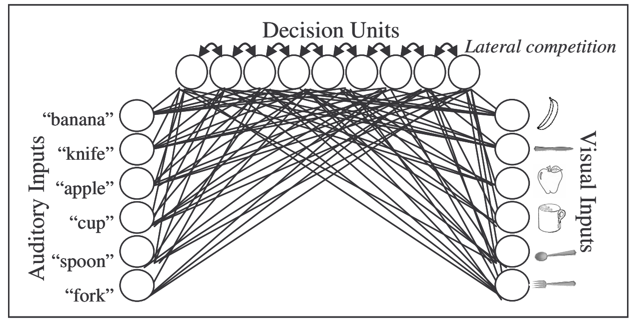


Figure 2: The Hebbian Normalized Recurrence Network

mapped names. Children were presented with two previously seen novel objects and one novel target. As Figure 1B shows, they were unable to determine the referents of the previously fast mapped names.

Follow-up experiments revealed that children's difficulty in retaining the name-object mappings was not due to the number of names presented in the session or to the number of fast mapping trials. Specifically, in a second experiment all but one of the novel name fast mapping trials was replaced with a filler trial (e.g., "can you get the one you like the best?"). In a third experiment the number of trials was reduced such that each child was presented with only three fast mapping trials: two familiar name trials and a single novel name trial. Both of these experiments replicated the general effect of Experiment 1: excellent referent selection but no retention—and will these children only retained four names. This finding is consistent with the literature, which indicates that children are able to retain name-object mappings if they are reviewed prior to the retention trials (see for example, Goodman, McDonough, & Brown, 1998). Thus, only when the objects were explicitly named and singled out by the experimenter do children show any retention and the retention they show is quite limited,

These findings suggest that although children are excellent at selecting the referent in a fast mapping task, they do not learn the name-object mappings in the moment. Children are able to select the novel object in response to the novel name, but are not able to encode the name, the object and their link strongly enough to survive a delay. This evidence supports the distinction of fast mapping and word learning as two distinct time scales in language acquisition.

These results suggest a number of conflicting interpretations. First, it is entirely possible that fast mapping and word learning are completely independent and unrelated. However, the alternative, more subtle interpretation, is that perhaps word leaning is a slow incremental process (one too slow to be seen on a single trial), but online fast mapping processes enable it to look much quicker. To examine this latter hypothesis, we simulated these data with a Hebbian Normalized Recurrent Network (HNRN, McMurray & Spivey, 1999).

The Hebbian Normalized Recurrent Network

The Hebbian Normalized Recurrent Network (HNRN) is based on a simple interactive architecture (Normalized Recurrence) in which multiple sources of probabilistic inputs are integrated and compete (in real-time) to arrive at an optimal integration. This has been shown to solve a large number of graded constraint satisfaction problems (c.f. Spivey & Dale, 2004) including high-dimensional categorization and visual search. Thus, the Normalized Recurrence architecture is ideal for capturing the fast mapping task, in which a child is presented with inputs (a novel name and several objects) and must also select a referent given a variety of graded constraints.

McMurray and Spivey (1999) added a form of unsupervised Hebbian learning to the HNRN in order to incorporate sensitivity to statistical regularities in the constraint satisfaction. This provides the sort of slow learning mechanism that may allow for long term word learning. This learning is fundamentally associative in nature is a realistic mechanism for children's early vocabulary acquisition. Smith (2000) and Samuelson (2002) have explicitly demonstrated benefits for associative learning in word learning. Moreover, while this associative scheme has been criticized as insufficient given the large number of visual competitors in the child's environment, McMurray, Horst, Toscano and Samuelson (in press) present simulations that suggest that HNRN can learn words even with 90% of the lexicon visually copresent.

Thus, HNRN has the potential to capture both the short time-scales of fast-mapping behaviors and the long time-scales of word learning. It presents an ideal architecture in which to explore these two components of acquisition.

The HNRN Word Learner

Our model of word learning consists of two input layers: an auditory (word) layer and a visual (object) layer (Figure 2). Each input layer was created with 15 localized input units, that is, each unit in the auditory input layer represents one name and each unit in the visual input layer represents one object. The network also includes one layer consisting of 90 decision/integration units. While this is more decision units than would be ultimately needed for the task, it ensures that initially (when weights were random) the model will be generally likely to choose different decision units for different inputs as a function of the degrees of freedom available.

Activation from these two input layers is sent concurrently over a series of learnable weights to the layer of decision/integration units. Here, activation accumulates such that the activation of a decision unit is the sum of its previous activation (d_x), the weighted (w_{xz}) activation of the auditory input (a_z) units and the weighted (u_{xz}) activation of the visual input units (v_z, Equation 1).

$$d_x = d_x + \sum_{z \in a} w_{xz} \cdot a_z + \sum_{z \in v} u_{xz} \cdot v_z \tag{1}$$

The activation of decision units is squared and normalized to sum to one. This implements a form of competition that is equivalent to lateral inhibition, in which each unit inhibits all other units in that layer as a function its proportion of the total activation.

Activation is then fed back to the input layers using a similar accumulator scheme. Here, however, activation from the decision layer is multiplied by the input array (preventing input nodes with no bottom-up support from being activated solely as the result of feedback, Equation 2). That is, the activation of an auditory unit is the sum of its previous activation (a_y) and the product of its previous activation and the weighted activation (w_{zy}) of the decision units (d_z). Likewise, the activation of a visual unit is the sum of its previous activation (v_y) and the product of its previous activation and the weighted activation (u_{zy}) of the decision units (d_z).

$$\begin{aligned} a_y &= a_y + a_y \cdot \sum_{z \in d} d_z w_{zy} \\ v_y &= v_y + v_y \cdot \sum_{z \in d} d_z u_{zy} \end{aligned} \tag{2}$$

The activation of inputs is then normalized so that the activations of the units sum to 1.

Activation continues to cycle from the input layers to the decision layer and back until the activation at the decision layer settles (the derivative approaches 1e-10).

Typically, the competition amongst decision units (implemented by squared-normalization) ensures that a single decision unit will be active when the model settles. Crucially, on each cycle (during both training and testing) weights are changed using a modified Hebbian learning rule (Equation 3). Here, η represents the learning rate and is typically very small (~5e-005). This rule ensures that the model will behave in one of three ways. First, when both an input unit (e.g., auditory unit) and a decision unit are active the connecting weights will increase in strength. Second, when an input unit is active while the corresponding decision unit is inactive, or when the decision unit is active and the input unit is inactive, the weights will slightly decrease. Finally, when neither the input unit nor the decision units are active, there will be no change to those connection weights. Importantly, this latter fact preserves plasticity in the weights for new names and objects (for a similar learning rule, see Grossberg, 1988).

$$\begin{aligned} \Delta w_{xy} &= \eta \cdot \left(a_y d_x - a_y w_{yx} - d_x w_{yx}\right) \\ \Delta u_{xy} &= \eta \cdot \left(v_y d_x - v_y u_{yx} - d_x u_{yx}\right) \end{aligned} \tag{3}$$

Simulation 1

The first simulation of the HNRN (described above) was trained on a small lexicon and then tested in a fast mapping task. The goal of this simulation was twofold. First, we sought to determine if the architecture of the model could exhibit fast mapping behavior. Second, and more importantly, we sought to determine the extent of learning that occurred during fast mapping. We reasoned that it would be possible to solve the problem of fast mapping with minimal learning, but that if fast mapping and word learning are related time scales of language acquisition that some learning should occur on each fast mapping trial.

Vocabulary Acquisition Phase

To simulate fast mapping we first needed a vocabulary of known names and objects. Thus, we trained 20 simulations on an initial vocabulary of five words for 5000 epochs and then presented the fast mapping and retention trials described below. Before the vocabulary acquisition phase began, the connection strength of all input units to the decision units was set to random values between 0 and .2.

On each cycle during this acquisition phase, one of the training words was randomly selected and its activation was set to 1. Next, the object that corresponded to that word was selected along with a variable number (on average 3) of visual competitors and their activation was normalized to 1. The activation from both layers spread to the decision layer and back to the inputs. Activation continued to cycle in this way until the model settled. Learning occurred throughout this cycling allowing the network to learn which name referred to which object.

Each of the 20 networks differed in a) the initial random weight matrix, b) the particular order of words, and c) the visual competitors for a given word on any trial. This created sizeable differences in performance.

Fast Mapping Trials

After vocabulary acquisition, the networks were presented with a three-alternative fast mapping task analogous to the Horst and Samuelson (submitted) task. On each fast mapping trial, one auditory unit (from the novel set) and exactly three object units were active: two trained object units and one object unit that was never activated during the vocabulary acquisition phase (i.e., a novel object). The networks were presented with five novel fast mapping trials, which were randomly intermixed with five fast mapping trials in which the referent was a known object, as an additional control. After each trial, the networks' success was evaluated by determining whether the most active visual unit corresponded to the most active auditory unit.

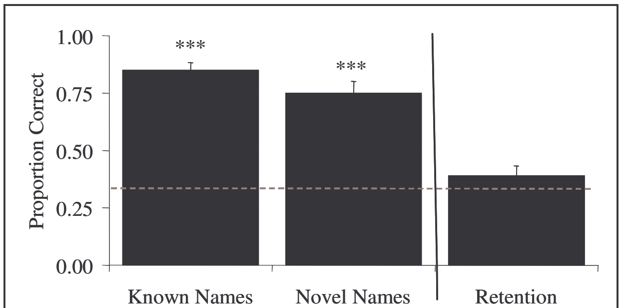
Importantly—and unlike many other connectionist simulations—learning continued throughout testing. Allowing the networks to continue learning during the test phase more accurately reflects the situation for children, who do not know to "stop learning." More importantly, the goal of the simulation was to determine how much learning (if any) could occur on the fast mapping trials. This was determined (post learning) by an analysis of the weight matrix.

Retention Trials

After the fast mapping trials, the networks were presented with five retention trials, analogous to those used in the Horst and Samuelson (submitted) task. On each retention trial, one name unit and exactly three object units were active: two object units that were activated previously during the fast mapping trials but not during the acquisition phase and one object unit that was never activated during the fast mapping trials or the acquisition phase (i.e., a novel foil object). Again, if the network settled on the object unit that corresponded to the activated name, this trial was scored as a correct response. Learning remained on during the retention trials.

Results

Results are depicted in Figure 3. Overall, the networks were exceptionally accurate on the novel name fast mapping trials M_{novel} = .75, SD = .22, $t(19) = 8.40$, $p < .0001$, two-tailed and the known name fast mapping trials, M_{known} = .85, SD = .14, $t(19) = 16.32$, $p < .0001$, two-tailed. However, despite settling on the correct object unit during the fast mapping trials, the networks did not settle on the correct object units during the retention trials, $M_{retention}$ = .39, SD = .19, $t(19) = 1.42$, *ns*. Thus, the networks showed the same pattern of results as the children in the empirical studies.



Error bars represent one standard error. Dotted line represents chance (.33). *** $p < .0001$

Figure 3: Results from Simulation 1

To gain a better understanding of the processes underlying fast mapping and word learning, specifically the differences in learning, the changes in the weight matrices during acquisition and testing were analyzed. That is, we assessed the amount of weight change (learning) for connections between decision units and known or novel names (the localist flavor of the input arrays allowed the weight matrix to be portioned out in this way quite simply). For each portion of each weight matrix the root mean squared (RMS) difference of the weights at two points in time was calculated.

We calculated the difference between the initial state and the state at the end of the acquisition phase ($RMS_{acquisition}$), that is, the weight change while the system was learning the known names. Next we calculated the difference between the end of the acquisition phase and the end of the fast mapping trials ($RMS_{testing}$), that is, the weight change while the system was fast mapping the novel names. If fast mapping represented true word learning, then the quantity of weight change during fast mapping should be similar to that seen in the acquisition phase.

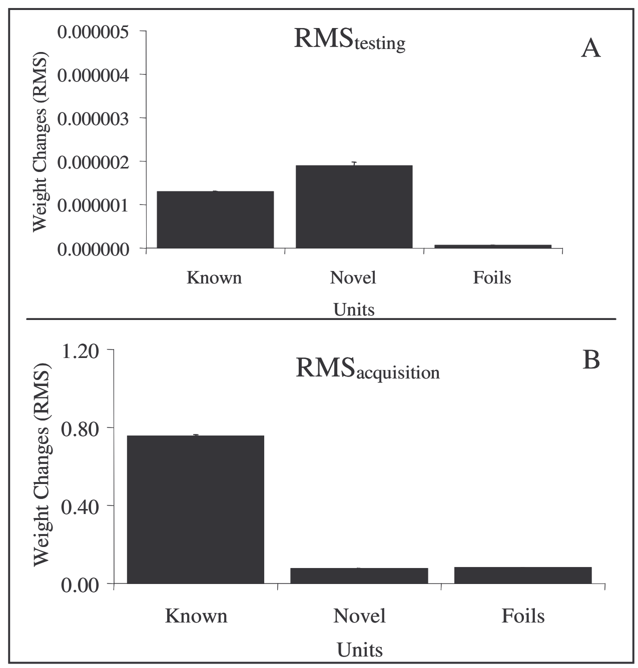


Figure 4: Weights Changes to the HNRN

The results of these analyses are depicted in Figure 4. The RMS deviation was averaged across both the auditory and visual units for each portion of the matrix (known items, novel items and foil items, which were held out). Clearly, learning did occur during the fast mapping trials (Novel $RMS_{testing}$ = 1.95e-6, SD = 2.24e-7, Figure 4A). However, although the lion's share of the weight changes during the fast mapping trials affected the connections between the novel units and decision units, this change remained only a fraction of the change that occurred for the known names during the vocabulary acquisition phase (Known $RMS_{acquisition}$ = .76, SD = .03, Figure 4B). Note the large difference in scale used in the two panels.

Discussion

Overall, the HNRN captured the empirical results and simulated both the moment-by-moment time scale of fast mapping and the more gradual time scale of word learning. In addition, the analyses of the weight matrices provide insight into the processes underlying both fast mapping and word learning. Specifically, these results suggest that a single fast mapping experience in and of itself does not constitute complete word learning. Moreover, the ability to select the same object in a novel context, as in the retention trials, is not gained over the course of a single trial. However, learning does occur on each fast mapping trial, although this learning is insufficient to create a robust enough representation of the name-object link to withstand further testing.

Clearly, then, learning does occur during fast mapping, though it is minimal compared to the amount of learning necessary for names to become "known words." This suggests that word learning is a slow, incremental process, and fast mapping is too quick to constitute complete word learning. However, because learning does occur during fast mapping, it is possible that fast mapping, when repeated, can lead to complete word learning. We tested this possibility in a second simulation. Specifically, we provided additional training on the novel items after the retention task and then retested the networks on the fast mapping and retention task. This allowed us to confirm that a novel name can become a known name with sufficient training.

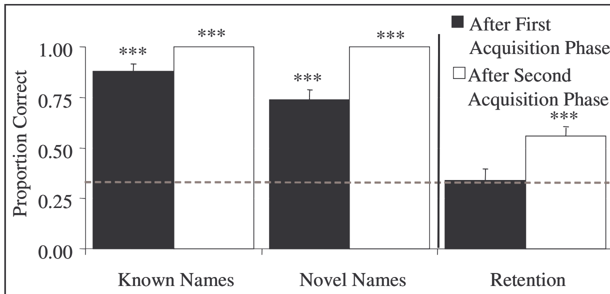
Simulation 2

We created 20 additional models and ran them through training and testing from Simulation 1. In this simulation, however, after the last retention trial the networks engaged in a second acquisition phase in which the five units that had served as the novel names were trained for 3000 epochs. Following this second learning phase, the networks were presented with a second set of fast mapping and retention trials as before.

Results

The results are depicted in Figure 5. After the first learning phase, the networks performed as in Simulation 1. That is, they correctly settled on the referent on both the novel name $M_{novel} = .74$, $SD = .21$, $t(19) = 8.89$, $p < .0001$, two-tailed, and the known name trials, $M_{known} = .88$, $SD = .15$, $t(19) = 16.31$, $p < .0001$, two-tailed. And, as observed previously, the networks did not retain the name-object mappings, $M_{retention} = .34$, $SD = .24$, $t(19) = .18$, *ns*. In contrast, after the second acquisition phase the networks accurately settled on the correct referent in both the fast mapping and retention trials (all p's < 0001). As can be clearly seen, the networks significantly improved in accuracy on between the first and second sets of retention trials, $t(19) = 3.69$, $p < .01$, two-tailed.

Again we examined the weight changes at different points in time: between the initial state and after the vocabulary acquisition phase, after the vocabulary acquisition phase to after the fast mapping trials and after the fast mapping trials to after the second acquisition phase.



Error bars represent one standard error. Dotted line represents chance (.33). *** $p < .0001$

Figure 5: Results from Simulation 2

We found the same pattern of results as in Simulation 1 for weight changes after the acquisition phase and after the fast mapping trials (Novel $RMS_{testing} = 2.0e\text{-}6$, $SD = .3e\text{-}8$; Novel $RMS_{acquisition} = .08$, $SD = .006$). In addition, the weights continued to change during the second acquisition phase (Novel $RMS_{acquisition}2 = .13$, $SD = .005$). Clearly, the amount of learning that occurred for the novel items during the second acquisition phase was still less than that of the known items during the first acquisition phase (Known $RMS_{acquisition} = .73$, $SD = .04$). However, this is not surprising given that some learning had occurred during the fast mapping trials, and the networks were trained for 200 fewer epochs during the second acquisition phase. Importantly, these data replicate the findings that the network can fast map novel names and that minimal learning occurs on each fast mapping trial. These data also show that the network is able to continue learning until novel names become known names.

Discussion

The goal of Simulation 2 was to test whether repeated fast mapping can lead to complete word learning. Because minimal learning did occur during fast mapping in Simulation 1, we reasoned that more complete word learning could arise from many, many fast mapping trials. Indeed, Simulation 2 confirmed that, with sufficient exposure, the network can come to treat novel names similarly to known names. After the second acquisition phase, there was no statistical difference in the networks' ability to select the referent in response to names acquired in the first or second acquisition phase. This indicates that with enough fast mapping complete word learning can occur.

Conclusions

The goal of these simulations was to shed light on the processes that govern fast mapping and word learning by capturing both time scales and investigating the learning processes that support them. Overall, the Hebbian Normalized Recurrent Network yielded the same pattern of results as found in the empirical studies. Further, the HNRN simulated both the moment-by-moment time scale of fast mapping and the more gradual time scale of word learning while showing that the two are subtly and importantly interrelated. The analyses of the weight matrices underscore this relationship. These analyses suggest that while learning does occur on each fast mapping trial, this learning is insufficient to create a robust enough representation of the name-object link to withstand further testing. That is, fast mapping in and of itself is not compete word learning and the ability to select the correct referent during a fast mapping trial does not promise the ability to select the same object in a novel context, as that of the retention trials.

Simulation 2 deepens our understanding of these processes by demonstrating that when provided with a review of the name-object mappings, that is, when provided with sufficient exposure, representations can be encoded robustly enough to withstand testing in a novel context. Put another way, a known name is a novel name that has been fast-mapped many, many times.

Taken together then, these simulations along with the supporting empirical results indicate that fast mapping and word learning represent two distinct, but related, components in vocabulary acquisition, and point to a promising future direction for a more complete understanding of these processes.

Acknowledgments

This research was supported by NICHD grant 5R01HD045713 to L.K.S.. We would like to thank Joseph Toscano for programming assistance.

References

Behrend, D. A., Scofield, J., & Kleinknecht, E. E. (2001). Beyond fast mapping: Young children's extensions of novel words and novel facts. *Developmental Psychology, 37*(5), 698-705.

Capone, N. C., & McGregor, K. K. (in press). The effect of semantic representation on toddler's word retrieval. *Journal of Speech, Language, and Hearing Research*.

Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan & A. Miller (Eds.), *Linguistic Theory and Psychological Reality* (pp. 264-293). Cambridge, MA: MIT Press.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Proceedings of the Stanford Child Language Conference, 15*(17-29).

Goodman, J. C., McDonough, L., & Brown, N. B. (1998). The role of semantic context and memory in the acquisition of novel nouns. *Child Development, 69*(5), 1330-1344.

Grossberg, S. (1988). Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. In J. A. Anderson & E. Rosenfeld (Eds.), *Neurocomputing: Foundations of research.* (pp. 245-258). Cambridge, MA: The MIT Press.

Horst, J. S., & Samuelson, L. K. (submitted). Fast mapping but not fast word learning: Referent selection without retention.

McMurray, B., Horst, J. S., Toscano, J., & Samuelson, L. K. (in press). Connectionist Learning and Dynamic Processing: Symbiotic Developmental Mechanisms. In J. P. Spencer, M. Thomas & J. McClelland (Eds.), *Towards a New Grand Theory of Development? Connectionism and Dynamic Systems Theory Reconsidered*: Oxford University Press.

McMurray, B., & Spivey, M. (1999). The categorical perception of consonants: the interaction of learning and processing. *Proceedings of the Chicago Linguistics Society, 34*(2), 205-220.

Samuelson, L. K. (2002). Statistical regularities in vocabulary guide language acquisition in connectionist models and 15-20-month-olds. *Developmental Psychology, 38*(6), 1016-1037.

Smith, L. B. (2000). Learning How to Learn Words: An Associative Crane. In R. M. Golinkoff, K. Hirsh-Pasek, L. Bloom, L. B. Smith, A. L. Woodward, N. Akhtar, M. Tomasello & G. Hollich (Eds.), *Becoming a Word Learner: A Debate on Lexical Acquisition* (pp. 51-80). New York: Oxford University Press.

Spivey, M. J., & Dale, R. (2004). On the continuity of mind: Toward a dynamical account of cognition. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 45, pp. 87-142). San Diego, CA: Elsevier Academic Press.