A dynamic neural field model of memory, attention and cross-situational word learning

Ajaz A. Bhat (a.bhat@uea.ac.uk)
School of Psychology, University of East Anglia
NR47TJ, Norwich UK

John P. Spencer (j.spencer@uea.ac.uk)
School of Psychology, University of East Anglia
NR47TJ, Norwich UK

Larissa K. Samuelson (l.samuelson@uea.ac.uk)
School of Psychology, University of East Anglia
NR47TJ, Norwich UK

Abstract

Recent empirical studies have affirmed the fundamental role of attention and memory processes in statistical word learning tasks. These processes interact in complex ways to guide spontaneous looking behaviors of learners as well as determine their overall learning performance. On the modelling side, studies have made it clear that computational models must provide process-based rather than only computational accounts of word learning, because these can connect to the empirically observed behaviors at a moment-to-moment timescale. Thus, here we present a neurally-grounded process model of word learning called WOLVES (Word-Object Learning Via Visual Exploration in Space) that integrates visual dynamics and word-object binding across multiple timescales. WOLVES integrates multiple established dynamic neural field models to allow fine-grained indexing of component processes driving the looking-learning loop. We report simulation results for three empirical cross-situational word learning experiments to validate the model.

Keywords: cross-situational word learning; dynamic neural field theory; DFT; attention and memory

Introduction

Word learning is at the core of language acquisition. A central challenge in this domain is referential uncertainty (Quine, 1960): a speaker's word can refer to many possible referents in a given visual scene, and a learner must identify the correct referent intended by the speaker. Though making such inferences seem impossible theoretically, children are quite adept at resolving referential uncertainty. One possible explanation is that children use statistical learning over multiple experiences. That is, while a single naming scenario may be referentially ambiguous, this ambiguity is gradually resolved as a learner tracks the co-occurrence statistics of words and referents across multiple naming events in time. This is commonly called cross-situational word learning and many recent studies, using force-choice tests in adults (Yu & Smith, 2007; Yurovsky, Yu, & Smith, 2013) and preferential looking tasks in infants (Smith & Yu, 2008; Yu & Smith, 2011), have confirmed that learners can extract correct word-object mappings cross-situationally over multiple ambiguous events.

What mechanisms underlie this form of word learning? Computational models have employed different conventions such as rule-based compositionality (Siskind, 1996), probabilistic inferences (Fazly, Alishahi, & Stevenson, 2010), associative matrices (Kachergis, Yu, & Shiffrin, 2012), associative connectionism (McMurray, Horst, & Samuelson, 2012), fast mapping (Trueswell, Medina, Hafri, & Gleitman, 2013) and others to explain the results from various cross-situational word learning experiments. Many of these studies are motivated by a debate over whether learners accumulate graded statistical information of all referents for each word (associative learning; McMurray et al., 2012) or propose and verify only a single referent per word (hypothesis-testing; Trueswell et al., 2013). Recent work suggests, however, that this debate is not well formed because these accounts reside at a 'computational' level and not at the 'algorithmic' level (or below) where they might shed light on the mechanisms underlying word learning (Smith, Suanda, & Yu, 2014). Other papers recommend further that (a) models should incorporate primary psychological processes such as attention and memory that are fundamental to learning words (Smith et al., 2014; Vlach & DeBrock, 2017); and (b) explicitly capture moment-by-moment and trial-by-trial looking and learning behavior of learners in cross-situational word-learning tasks (Yu, Zhong, & Fricker, 2012).

Consistent with these later recommendations, the present report describes a process-based modelling account of cross-situational word learning and validates this neurally-grounded, non-linear statistical learner in different cross-situational word learning tasks. The model uses the framework of Dynamic Field Theory (DFT) (Schöner, Spencer, & The DFT Research Group, 2015) to simulate the moment-to-moment visual dynamics as the learning process unfolds in time. In this article, we provide simulation results of experiments from three recent empirical studies on cross-situational word learning in support of the model.

Dynamic Field Theory

DFT proposes that cognition arises from neurocomputations within dynamic neural fields (DNFs) that simulate the dynamics of neural populations. Expressed using differential equations, these fields are receptive to metric dimensions such as color, shape, space etc. Neurons within a field interact temporally with one another or with the incoming input stimuli to form *peaks* that act as stable states of the field. These peaks are stabilized by excitatory and inhibitory neural interactions within the field. *Self-sustaining* peaks survive even if input is removed due to strong recurrent activations and thus act as a form of working memory to maintain information (Amari, 1977).

A DNF architecture consists of 1- and 2-dimensional fields (see blue rectangles in Figure 1) that interact along shared dimensions through unidirectional or bidirectional projections. Fields can also use a variant of Hebbian learning at a slower timescale that allows neural populations to learn and encode statistical information across trials. This type of memory trace enables specific neural regions in the field to become more strongly activated, increasing the likelihood that a peak will form in that region of the field. This results in a 'pre-shaping' effect, facilitating recognition of familiar inputs.

In the context of word learning, DFT offers two core strengths. First, it has already been used to test predictions regarding component processes such as early visual processing, attention, spatial cognition and working memory, at behavioral and brain levels. Second, DFT scales up allowing integration of multiple models into large-scale systems to explain and predict behavioral data. Exploiting these strengths, this article integrates multiple previously-established DFT models of the component processes involved in word learning; one on word-object mapping (Samuelson, Smith, Perry, & Spencer, 2011) and the others on visual attention and memory (Perone & Spencer, 2013; Schneegans, Lins, & Spencer, 2015) to provide a formal account of cross-situational word learning.

WOLVES: The Model

Word-Object Learning via Visual Exploration in Space (WOLVES): The WOLVES model is composed of multiple 1D and 2D neural fields as shown in Figure 1. Higher dimensional fields are not used due to computing limits. Going from the right side in Figure 1, the model has two 2D visual fields (A, B) that share a visual spatial dimension within a retinal reference frame. Note that we model only a single spatial dimension and it encodes relative rather than absolute spatial locations of objects. When object stimuli are presented to the model, these two fields respond to the detection of color and shape of objects at corresponding locations in the visual field. These fields are coupled to a 1D spatial attention field (C) that receives activation reflecting object locations in the visual field. Each visual field passes activation to three 1D fields along the feature pathway in the model: feature-attention fields (H, I), contrast fields (F, G), and working memory (WM) fields (D, E). Feature-attention fields are reciprocally coupled to visual and contrast fields, while the contrast and WM fields are mutually inhibitory.

Figure 1 depicts a case where two stimuli are in the visual display (see right top corner). The feature-space fields (A, B) have formed peaks of activation for a blue square on the left and a red circle on the right. This input causes peaks to grow at the corresponding locations of the spatial attention field (C) and at the corresponding feature values in the fields along the feature pathway (D-I). The three attention fields (C, H, I) work in a winner-take-all mode such that the first peak to breach the interaction threshold (red line in each field) will suppress all other inputs currently vying for attention. Then, through reciprocal coupling between the visual and attention fields, the model will selectively attend to a single item on the visual field. In Figure 1, the model is currently attending to the left object, with a peak on the left side of the spatial attention field (C), at the blue hue value in the color attention field (H), and at the square shape value in the shape attention field (I).

The contrast fields (F, G) serve as novelty detectors where novelty is defined as any feature that is not currently being actively maintained in WM. For instance, in Figure 1, the model has a robust WM of the circle feature (in E) because it previously attended to the red circle. Thus, the circle feature is not currently novel—there is no contrast peak at the circle value in G; rather, the square feature is novel (G). This novelty peak helps to stabilize attention to this feature via feedback to the shape attention field (I).

Like 1D fields in the feature-pathway, the model has three 1D fields in the spatial pathway—a spatial attention field, (J) a spatial contrast field (K) and a spatial WM field (L). Respectively, these serve to create spatially "bound" representations of the attended object in an allocentric frame, detect changes in object locations and build WMs for the locations of objects, analogous to fields in the feature pathway. This dichotomous information-flow at the visual front-end of the model is guided by the functional and anatomical separation of the mammalian visual system into dorsal ("where") and ventral ("what") streams.

On the left side of the figure, the model has a 1D field—word input (M) and multiple 2D fields—a word-color (N), a word-shape (O), and two scene attention fields (P, Q) with space-feature dimensionality (two additional WM fields are not shown for simplicity). Fields are reciprocally connected such that activation passes along the four shared dimensions: words (N↔M↔O), space (P↔J↔Q), color (N↔P), and shape (O↔Q). Activations related to the attended object in the feature attention fields (H, I) are passed on to the scene attention fields (P, Q) [see the horizontal 'ridges' of activation]. Similarly, activation related to the spatial position of the attended object (J) is passed along the spatial dimension into the scene attention fields [see vertical ridges of activation]. In Figure 1 panels P and Q, these ridges cross, creating a "bound" object representation—a pattern of peaks representing a specific color and shape at a leftward location. Similar coupling across the word dimension (peak in M; vertical ridge in N,

O), binds a word with associated features. The 1D spatial and feature WM fields are coupled to two 2D WM fields (not shown). These fields have the same dimensionality as the scene attention fields (P, Q), but they can maintain multiple peaks.

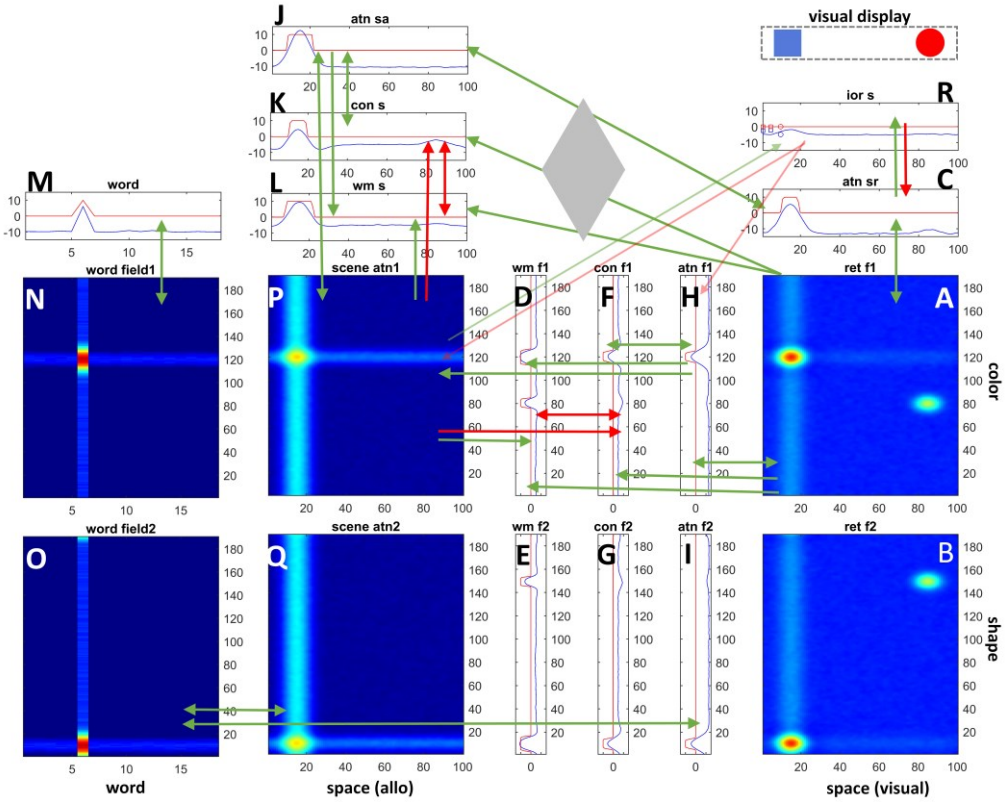

Figure 1: WOLVES model architecture. Rectangles A-R represent component DNFs and arrows represent uni- /bi-directional (green: excitatory, red: inhibitory) connectivity in the model. See text for figure description.

Object binding is based on a winner-take-all attention system that spans J, P, and Q. This selects objects in allocentric space that are the focus of attention, binds the associated features to spatial positions, and supports the consolidation of items in WM. The transformation between visual/retinal (C) and allocentric (J) space is handled by a transformation system (grey diamond). For simplicity we use a one-to-one mapping transformation between the visual (C) and allocentric fields (J), although more rigorous DFT variants of such transformations have been proposed before (Schneegans & Schöner, 2012). Note that all the spatial attention layers—attention in the visual (C) and allocentric spatial (J) frames, and the object attention system (J, P, Q)—are all reciprocally coupled (bi-directional green arrows). This keeps WOLVES oriented in space as it shifts attention between frames of reference. The Inhibition of Return (IOR) field (R) is driven by the formation of peaks in the object attention system: the system can release attention once the features of the current object have been bound.

All WM and contrast fields have memory traces that influence the current activity in these fields according to their histories of neural decisions (or peak formations). These memory traces enable habituation to the locations of objects in P & Q, becoming familiarized with visual features in D & E and importantly, enable trial-to-trial learning of which objects are where and what features go with each word, building a vocabulary of word-object mappings in N & O. The bi-directional coupling between word-feature associations and feature-based attention (see green arrow from O to I; comparable coupling would link N and H) is key to supporting these mappings. It also enables word-feature associations to bias selective attention—a central aspect of the model that drives looking as the system learns word-object mappings.

The full architecture can shift attention back and forth autonomously among a set of objects, learning about their visual features. The initial selection of an object is influenced by salience (via the strength of inputs to the visual field), novelty (via peaks in the contrast fields), and by random fluctuations (i.e., noise). As memory traces of object-word mappings build, words come to drive the model's attention to the object that has been mapped to the current word input.

Results

WOLVES allows in-the-moment representation and measurement of visual attention dynamics as a task unfolds trial by trial. This makes WOLVES well-suited for modelling infant studies where eye-tracking and preferential looking are used to quantify behavior. In the following, we begin by reporting simulations of infant studies on cross-situational word learning (Simulation Study A). Studies B and C apply the model to recent empirical experiments on adults and older children respectively.

Simulation Study A: Smith & Yu (2008) and Yu & Smith (2011)

Smith and Yu published two studies investigating cross-situational learning in 12- and 14-month old infants. Both studies used the same preferential looking task to show that infants seem to learn words by tracking co-occurrences over time. In this task (Smith & Yu, 2008), infants saw two novel objects on a slide that lasted for 4 seconds and heard two

novel words, one after the other as the slide was presented. Across a training period composed of 30 such slides, infants were exposed to six words-object pairs in random combinations with two pairs per slide. Immediately after training, word-object mappings were tested via preferential looking; two objects were presented for 8 seconds along with a single word. Infants' looking to the two objects was recorded using eye-tracking. Greater looking to the object previously paired with the heard word was taken to indicate learning. Each mapping was tested twice.

Yu & Smith (2011), used the same task to explore the relationship between selective attention and learning in infants. The authors categorized infants that looked more at target objects than distractors as Strong learners and the other infants that looked more to distractors as Weak learners. The authors then reported multiple measurements of Strong/Weak infant looking dynamics and learning behavior (see Table 1) to look for relationships.

We simulated the full cross-situational word learning task with WOLVES using Gaussian inputs to represent the color and shape features of objects, presented to the visual field, and word stimuli presented to the word field. Time in the model is scaled such that each simulation timestep equals eight seconds of real experimental time across all studies. The model can autonomously attend to the input stimuli. Over time, as the model attends to the presented objects and words, it forms working memories of visual scenes and accumulates word-object co-occurrence statistics over multiple presentations into memory traces. Later once strong word-object mappings are built, this results in selective attention to objects. These interactive dynamics in the looking-learning loop evolve as the task unfolds. Because we simulate the same preferential looking task presented to infants, we can directly compare the same behaviors across simulations and infants (see Table 1).

Table 1: Infant and Model looking and learning results

	Smith & Yu (2008)	**Yu & Smith (2011)**	**WOLVES Model**
Mean # of words learned (out of 6)	4	3.5	3.6
Mean looking time to objects at test (8s trial)	5.85 sec	5.92	5.1
Mean looking time at Target vs Distractor	3.4 vs 2.45 sec	3.25 vs 2.67	2.9 vs 2.3
Mean duration looking at objects training (4s trial)	3.14 sec	Strong 2.96 Weak 3.07	2.6
Mean # of fixations per trial	NA	Strong 2.75 Weak 3.82	2.6
Mean fixation duration (in sec)	NA	Strong 1.69 Weak 1.72	1.6

Overall, the model behavior is in close proximity to that of the reported infant behavior in the task. Like the empirical studies, the model also gives rise to Strong and Weak learners (in some cases (e.g., Table 1) even for parametrically same model instances). Most measures of looking and learning from the empirical studies and the model correlate well. A key insight from our model results was that the individual differences in the behavior of different model learners (or infants) did not necessitate different parametric instantiations (or cognitive conditions) but simply resulted from the differences in the order of stimuli presentations, internal noise of the learning system, and the non-linear interactions between these factors.

Simulation Study B: Yurovsky, Yu & Smith (2013) Experiment 1.

In this study, the authors explored competitive mechanisms involved in cross-situational word learning. The hypothesis was that on a single presentation trial, learning multiple referents for one word should be difficult because of the local competition between word-object mappings. To examine this, adult participants were exposed to six *single* words (words mapped correctly to a single referent in a trial) and six *double* words (words mapped to two different referents in a trial) over a learning session of 27 randomized trials. At test, participants heard one of the twelve words and ranked four presented objects (by clicking on them) in the order of their likelihood of being the referent. Participants were credited with knowing the correct referent for a *single* word, if it was their first guess ('single' bars in Figure 2). Participants were credited with knowing a double word if they selected *either* of the correct referents as first guess ('either' bars in Figure 2). If participants selected *both* the referents as first and second guesses, they were credited with knowing both referents ('both' bars in Figure 2).

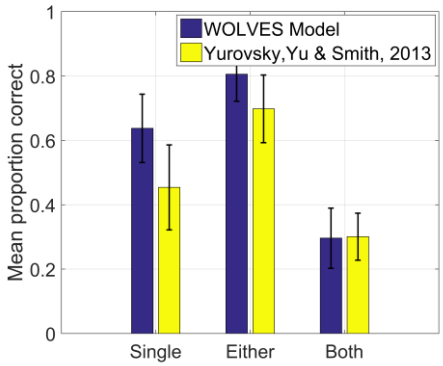

Figure 2: Adult and model accuracy at test for each word type. The model results closely match the empirical data.

We simulated this experiment with WOLVES using the same procedure as in the empirical study but took our test measurements from the model's preferential looking behavior during the first 1000 millisecond time window of word presentation. The model was credited with knowing the correct referent for a single word if it looked more to the

target than to the three distractors. The model was credited with knowing the correct referent for a double word if it looked at either of the correct referents more than the other three objects ('either' bars). If looking time to both the referents was more than looking time to the two distractors, the model was credited with knowing both the correct referents ('both' bars). The model learned at rates comparable to adult performance, within an RMSE of 0.1228 in overall correct response proportions (Figure 2).

Yurovsky et al., (2013) concluded that competition is involved in every trial with a *double* word because referents inhibit one another's mapping to the word. They suggest learners divide their attention between the two referents on *double* word trials. WOLVES confirms that multiple referents put strain on limited attentional resources and that sometimes only one of the referents is well attended. However, even if both referents get attention, the memory trace of a word associated with object in early trials will inhibit formation of mappings between the word and a second referent by directing selective attention to the previously mapped object. In this way, the model reflects how selective attention and memory interact online to give rise to the observed behavior of adults and indicates that the explanation for such results may not require use of additional processes like competition or mutual exclusivity.

Simulation Study C: Suanda, Mugwanya & Namy (2014) Experiment 1.

This experiment investigated how the diversity of learning contexts (i.e., the degree of word-referent cross-correlations) affected performance in 6-year old children. The hypothesis was that if children employ a cross-situational learning strategy, then they should be less successful in learning the correct mappings in lower contextual diversity conditions (i.e., when there are many strong word-referent cross-correlations) than in situations with higher contextual diversity. To explore this, children were randomly assigned to three conditions: In a High Contextual Diversity (HCD) condition, the other word-object pairings seen were different in each of the four training trials (i.e. no cross-correlations between mappings) for an object-word pair. In the Moderate CD condition, word–picture pairings co-occurred with one word–picture pairing on two trials and two other word–picture pairings on the other two trials, resulting in less diversity across trials. Finally, in the Low CD condition, word–picture pairings co-occurred with one word–picture pairing three times and another word–picture pairing once, so the diversity across trials was low (many cross-correlations). First, the children were familiarized with the task for a separate set of words and objects. Then as per the CD condition assigned, children saw two pictures and heard two words corresponding to the two pictures per trial. Eight word-object pairs were presented over a session of 16 training trials. Finally, on each force-choice test trial children were presented with a word and asked which of four displayed pictures the word referred to. Children's performance decreased with decreasing contextual diversity (i.e. increasing cross-correlations made learning difficult, see Figure 3).

We simulated the same experimental procedure in WOLVES and measured preferential looking behavior at test. If in the first 1000 milliseconds following the word presentation at test, the model looked more to the target than all the distractors, it was credited with knowing the correct mapping. As Figure 3 shows, the simulation results follow the same downward trend in mean proportion correct response across conditions as the empirical work with children with values within an RMSE of 0.1813.

Suanda, Mugwanya, & Namy (2014) reported that the precise reason for why contextual diversity helps was unclear but suggested possibilities: (a) increasing variability of learning instances allows for more decontextualized representations; (b) variability allows for a greater number of potential cues at memory retrieval time; and (c) variability initially creates 'desirable difficulties' in learning that boosts the strength of learning in the long run. WOLVES provides clarity by highlighting the real-time interactions between memory trace formation and the selective attention these memories capture. In the HCD condition, memories for correct mappings are reinforced after every exposure to a highly diverse context while incorrect mappings are not. This allows the correct mappings to selectively guide the model to attend to the correct referent after a word is presented, leading to stronger correct mappings and therefore better learning. In the LCD condition, memories for both incorrect and correct mappings are reinforced on every exposure to the less diverse context. Thus, word-driven selective attention gets directed to incorrect referents more often, leading to less learning overall.

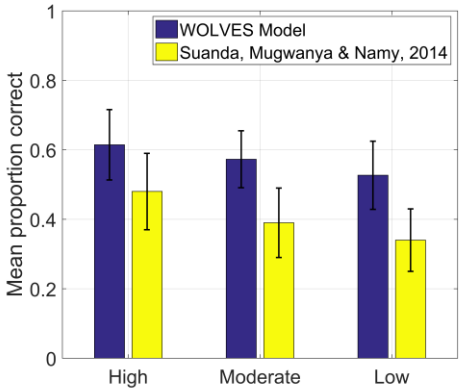

Figure 3: Response accuracy of the model and children show a comparable descending pattern across levels of contextual diversity.

Discussion

In this article, we described the first neurally-grounded model of word learning that incorporates visual dynamics in the word-object learning processes and tested it in the context of cross-situational learning scenarios. Across three different cross-situational word-learning simulations, the

model successfully learned novel words and showed looking and learning behavior akin to that reported in empirical studies on infants, children and adults. Our simulations indicate that mechanisms including memory trace consolidation, speed of visual processing, strength of working memories, top-down and bottom up attentional influences operate in non-linearly complex ways to give rise to a diverse range of learning trajectories, although we are still exploring the model behavior to come up with a unified theory of the word learning phenomena

In contrast to existing computational models that focus on corpus analyses and hence avoid direct modelling and comparison to empirical results, the proposed model allows moment-to-moment behavioral modelling, analysis and prediction that we are pursuing in our ongoing work. For instance, WOLVES allows us to examine individual differences and to make inferences about individual learning trajectories. Similarly, modeling studies from different age groups will be important in drawing the developmental trajectory of learning and looking for the component processes that drive this development.

Our ongoing work, seeks a comprehensive modeling account of cross-situational word learning by investigating the role of partial knowledge (Kachergis et al., 2012), effects of varying referential uncertainty and referent frequency (Yu & Smith, 2007), role of different memory subsystems (Vlach & DeBrock, 2017) and the effect of novelty and selective attention on learning in cross-situational scenarios (Yu et al., 2012). This larger set of simulations provides the basis to explore the parametric space of the model and the influence of these parameters on the model's learning, attention and memory.

Acknowledgments

This work was funded by grant no. R01HD045713 from the NICHD awarded to LKS.

References

Amari, S. I. (1977). Dynamics of Pattern Formation in Lateral-Inhibition Type Neural Fields. *Biological Cybernetics*, *27*(2), 77–87.

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A Probabilistic Computational Model of Cross-Situational Word Learning. *Cognitive Science*, *34*(6), 1017–1063.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2012). An associative model of adaptive inference for learning word-referent mappings. *Psychonomic Bulletin & Review*, *19*(2), 317–24.

McMurray, B., Horst, J. S., & Samuelson, L. K. L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*(4), 831–877.

Perone, S., & Spencer, J. P. (2013). Autonomy in action: linking the act of looking to memory formation in infancy via dynamic neural fields. *Cognitive Science*, *37*(1), 1–60.

Quine, W. V. O. (1960). *Word and object*. Cambridge: MIT PRess.

Samuelson, L. K., Smith, L. B., Perry, L. K., & Spencer, J. P. (2011). Grounding word learning in space. *PLoS ONE*, *6*(12), e28095.

Schneegans, S., Lins, J., & Spencer, J. P. (n.d.). Integration and selection in multi-dimensional dynamic fields. In G. Schöner, J. P. Spencer, & The DFT Research Group (Eds.), *Dynamic Thinking: A Primer on Dynamic Field Theory*. New York, NY: Oxford University Press.

Schneegans, S., & Schöner, G. (2012). A neural mechanism for coordinate transformation predicts pre-saccadic remapping. *Biological Cybernetics*, *106*(2), 89–109.

Schöner, G., Spencer, J. P., & The DFT Research Group. (2015). *Dynamic Thinking: A Primer on Dynamic Field Theory*. New York, NY: Oxford University Press.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1–2), 39–91.

Smith, L. B., Suanda, S., & Yu, C. (2014). The unrealized promise of infant statistical word-referent learning. *Trends in Cognitive Sciences*, *18*(5), 251–258.

Smith, L. B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3),1558–68.

Suanda, S. H., Mugwanya, N., & Namy, L. L. (2014). Cross-situational statistical word learning in young children. *Journal of Experimental Child Psychology*, *126*, 395–411.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*(1),126–56.

Vlach, H. A., & DeBrock, C. A. (2017). Remember dax? Relations between children's cross-situational word learning, memory, and language abilities. *Journal of Memory and Language*, *93*, 217–230.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics: Research article. *Psychological Science*, *18*(5), 414–420.

Yu, C., & Smith, L. B. (2011). What you learn is what you see: Using eye movements to study infant cross-situational word learning. *Developmental Science*, *14*(2), 165–180.

Yu, C., Zhong, Y., & Fricker, D. (2012). Selective attention in cross-situational statistical learning: evidence from eye tracking. *Frontiers in Psychology*, *3*(June), 148.

Yurovsky, D., Yu, C., & Smith, L. B. (2013). Competitive processes in cross-situational word learning. *Cognitive Science*, *37*(5), 891–921.