Word learning emerges from the interaction of online referent selection and slow associative learning

Bob McMurray,
Dept. of Psychology and Delta Center University of Iowa

Jessica S. Horst, and
School of Psychology University of Sussex

Larissa K. Samuelson
Dept. of Psychology and Delta Center University of Iowa

Abstract

Classic approaches to word learning emphasize the problem of referential ambiguity: in any naming situation the referent of a novel word must be selected from many possible objects, properties, actions, etc. To solve this problem, researchers have posited numerous constraints, and inference strategies, but assume that determining the referent of a novel word is isomorphic to learning. We present an alternative model in which referent selection is an online process that is independent of long-term learning. This two timescale approach creates significant power in the developing system. We illustrate this with a dynamic associative model in which referent selection is simulated as dynamic competition between competing referents, and learning is simulated using associative (Hebbian) learning. This model can account for a range of findings including the delay in expressive vocabulary relative to receptive vocabulary, learning under high degrees of referential ambiguity using cross-situational statistics, accelerating (vocabulary explosion) and decelerating (power-law) learning rates, fast-mapping by mutual exclusivity (and differences in bilinguals), improvements in familiar word recognition with development, and correlations between individual differences in speed of processing and learning. Five theoretical points are illustrated. 1) Word learning does not require specialized processes – general association learning buttressed by dynamic competition can account for much of the literature. 2) The processes of recognizing familiar words are not different than those that support novel words (e.g., fast-mapping). 3) Online competition may allow the network (or child) to leverage information available in the task to augment performance or behavior despite what might be relatively slow learning or poor representations. 4) Even associative learning is more complex than previously thought – a major contributor to performance is the pruning of incorrect associations between words and referents. 5) Finally, the model illustrates that learning and referent selection/word recognition, though logically distinct, can be deeply and subtly related as phenomena like speed of processing and mutual exclusivity may derive in part from the way learning shapes the system. As a whole, this suggests more sophisticated ways of describing the interaction between situation- and developmental-time processes and points to the need for considering such interactions as a primary determinant of development and processing in children.

Corresponding Author: Bob McMurray Dept. of Psychology E11 SSH University of Iowa Iowa City, IA 52242 319-335-2408 (voice) 319-335-0191 (fax) bob-mcmurray@uiowa.edu.

Introduction

The word is fundamental to language. Words serve an organizing role in syntactic parsing (MacDonald, Pearlmutter, & Seidenberg, 1994; Tanenhaus & Trueswell, 1995), speech perception (Goldinger, 1998; McClelland, Mirman, & Holt, 2006) and semantic organization (Elman, 2009; Lupyan, Rakison, & McClelland, 2007; Mayor & Plunkett, 2010; Samuelson & Smith, 2000; Waxman, 2003). Lexical items live at a critical juncture in language processing, linking sound, articulation, syntax and meaning and as a result, the acquisition of words has attracted enormous attention (P. Bloom, 2000; Carey, 1978; Fenson, Dale, Reznick, Bates, & Pethick, 1994; Golinkoff et al., 2000; Mayor & Plunkett, 2010; Xu & Tenenbaum, 2007).

Whether such research examines the growth of the lexicon as a whole (e.g., Fenson, et al., 1994; Ganger & Brent, 2004) or conducts micro-investigations of single word learning (e.g., Carey & Bartlett, 1978; Horst & Samuelson, 2008), the fundamental questions concern *word knowledge*: 1) whether children know a word, 2) how they come to know it, and 3) how many they know. The typical article addressing this starts with the scale: children acquire about 60,000 words in about 18 years. It then describes why this is so hard. Famously articulated by Quine (1960), in any naming situation there are infinite interpretations for an unknown word. Thus, children face a daunting task of ambiguity resolution that they must solve thousands of times.

Such a paper then proposes an explanation for how children solve this problem, but often skips a primary question—what does it mean to know or learn a word? A canonical finding is that toddlers comprehend more words than they produce: 75% of 12-month-olds understand *all gone* but it takes 8 more months before that many say it. When do we consider *all gone* known? Children can often identify the referents of novel words on their first exposure (Mervis & Bertrand, 1994), yet their ability to recognize familiar words develops over some time (Fernald, Pinto, Swingley, Weinberg, & McRoberts, 1998). Again, when do we consider a word known?

Perhaps it is not possible to quantify when, or if, a word is known. If so, the problem of lexical acquisition may be better framed in terms of how children learn to *use* words. After all, we can measure *word use* directly. Commonly, this idea calls to mind the process of producing words, but we mean something broader. To the extent that a word links sound and meaning, any time that link is used to guide behavior, a word is being used. Thus, word use also includes processes like comprehending known words, and even determining referents for new words.

If we ignore the uncertainty of knowledge and focus on only children's word use, children must still solve a set of difficult problems. Yet, the concept of *using* a word does not appear in classic descriptions of word learning. Rather, the focus is on the information needed for learning, the amount that must be gathered and the difficulties in gathering it. This has led to theoretical views that emphasize knowledge-based processes in accounting for learning, but inadvertently deemphasize use (Golinkoff & Hirsh-Pasek, 2006; Golinkoff, Mervis, & Hirsh-Pasek, 1994; Mayor & Plunkett, 2010; Woodward & Markman, 1998; Xu & Tenenbaum, 2007). Our purpose here is to advance an account of the development of *word use*, both novel and familiar, over multiple timescales. We demonstrate its power with a computational model.

In developing our account, we start by discussing the standard view of word learning and the theoretical tensions surrounding it. We then distill word learning to the minimal computational problem to frame our account. Finally, we present an account based on associative learning and dynamic competition and demonstrate its power to illuminate

lexical behavior. These simple mechanisms operate on different timescales – dynamic competition describes the situation-time process of selecting a referent of a word, while associative learning describes the developmental time process of slowly forming mappings between words and concepts. Critically, the interactions of these timescales can yield emergent power to describe lexical behaviors that does not derive from either one alone.

The standard view: Acquiring lexical knowledge

Early approaches to word learning used measurement studies to examine the number of words known over development (e.g., L. Bloom, 1973; Dore, Franklin, Miller, & Ramer, 1976; Fenson, et al., 1994; Reznick & Goldfield, 1992). A key finding was that word learning appears to accelerate. The source of this is debated (P. Bloom, 2000; Ganger & Brent, 2004; McMurray, 2007; Nazzi & Bertoncini, 2003), but clearly children are efficient learners and may become more efficient over development. This contrasts with the apparent difficulty of word learning. A major obstacle to acquiring words is *referential ambiguity* (Quine, 1960): in any naming event, a novel word can refer to any object present, its properties, the speaker's feelings or intentions for it, an impending action, or to something else all together. Even considering only the smaller problem of which *object* or category of objects is being referred to, this is still challenging.

The *constraint approach* offers a meta-theory for solving referential ambiguity: children have (perhaps innate) constraints, principles or biases that help them infer a word's meaning by providing information not available in the situation (Golinkoff, et al., 1994; Woodward & Markman, 1998). The most elementary constraints simply restrict the possible interpretations of a novel word (Markman, 1990), positing that new words refer to whole objects (not parts), or to basic-level categories (rather than super- or sub-ordinate categories). More complex constraints like social cues may go further, pinpointing the correct referent (Baldwin, 1991; Baldwin, Markman, Bill, Desjardins, & Irwin, 1996; Tomasello, Strosberg, & Akhtar, 1996).

Particularly relevant to the present study is the mutual exclusivity constraint (Markman & Wachtel, 1988) and the similar novel name-nameless category principle (N3C, Mervis & Bertrand, 1994) which describe how children infer the referent of a word on the basis of which other objects they have names for. For example, when presented with a familiar spoon and a novel whisk, children infer that *whisk* refers to the latter, if they know the word *spoon*. The form of this inference has been debated (Grassmann & Tomasello, 2010; Halberda, 2006; Jaswal & Hansen, 2006; Markman & Wachtel, 1988; Mervis & Bertrand, 1994), yet it is clear that children can make inferences that integrate available context with the contents of their lexica.

In the constraint approach, such inferences become the primary route to learning, and some have argued that it is the onset of these constraints or the related social/pragmatic skills that create the sudden acceleration in word learning (Golinkoff & Hirsh-Pasek, 2006; Golinkoff, et al., 1994; Markman, 1990; Nazzi & Bertoncini, 2003). However, the constraint approach has been challenged on a number of theoretical and empirical grounds.

Theoretical Challenges to Constraints

A classic concern with the constraint approach is its ability to scale up. Once children master basic-level terms, how is the taxonomic constraint relaxed to learn super-ordinates? When can children violate the whole-object constraint to learn properties? And, how do children ignore mutual exclusivity to learn synonyms or super-ordinates? Similarly, the constraint approach doesn't provide a clear framework for how constraints interact, or how conflicts are resolved.

This has led some to cast word learning in terms of general inference processes used for reasoning or social/pragmatic behavior. Mutual exclusivity, for example, applies to concepts (Behrend, Scofield, & Kleinknecht, 2001; P. Bloom & Markson, 1998; Markson & Bloom, 1997; Waxman & Booth, 2000), and can be described by principles like deductive syllogism (Halberda, 2006). These more general-purpose problem-solving skills may avoid issues related to the relaxation of hard constraints. A powerful formulation of this is to frame the problem in terms of probabilistic (Bayesian) inference (Frank, Goodman, & Tenenbaum, 2009; Xu & Tenenbaum, 2007). In this view, constraints are prior probabilities that interact with context and existing knowledge using the laws of probability to determine the optimal solution. This can handle interacting constraints, and allows constraints to be violated when the data permits. Others have cast the problem in terms of social inference (Akhtar & Martinez-Sussman, 2007; Moore, 2006; Tomasello, 2001). This transforms the problem from determining what a word means, to determining what a speaker is referring to. Mutual exclusivity, then, becomes a social inference process, in which speakers are expected to follow social conventions and use familiar words (e.g., Diesendruck & Markson, 2001). Social and Bayesian accounts are not mutually exclusive. Both approaches are cast fundamentally in terms of acquiring *knowledge* about words, but say little about how words are used once (or while) this knowledge is acquired (though see Frank, et al., 2009) or how these inferences and knowledge relate to long-term learning.

Empirical Challenges to Constraints

The foundations of the constraint approach have been shaken by research on four topics: the vocabulary spurt, fast-mapping, cross-situational learning, and familiar word recognition.

Vocabulary Spurt—The sudden acceleration vocabulary growth has been seen as indirect evidence for constraints by implying their sudden availability or a change in children's approach to word learning. However, Ganger and Brent (2004) argued that if the vocabulary spurt was a singular change, then the velocity profiles of individual children should show a sudden shift in velocity. Yet, for 33 of 38 children, a smoothly accelerating function fit bitter.

Alternatively, it is possible that principles like mutual exclusivity are available all the time (Akhtar & Martinez-Sussman, 2007; Markman, Wasow, & Hanson, 2003; Tomasello, 2001), but children simply don't have enough words (or other knowledge) to use them. As the first few words are acquired, these exert greater leverage, allowing the rate of acquisition to increase (c.f., Elman et al., 1996; van Geert, 1991). Recently, however, McMurray (2007; Mitchell & McMurray, 2009) demonstrated that acceleration is possible without such mechanisms. As long as learning proceeds in parallel and the distribution of easy and hard words includes few easy words, acceleration is guaranteed. Thus, while a change in underlying learning mechanism or constraints could account for acceleration, acceleration is not evidence for it.

Fast-mapping—In her original discussion, Carey (1978) contrasted children's quick mapping of a novel word to a novel referent (which was illustrated with the first demonstrations of mutual exclusivity), with a slower, phase of learning the word's full meaning. It is not clear if this "fast-mapping" refers to partial, early stages of learning or purely in-the-moment referent selection (though the word "mapping" implies learning). Nevertheless, if word learning is due to fast-mapping, then the act of selecting the referent should result in something being retained.

The most compelling test of this would first ask if the child selects the correct referent, and then examine retention when the child is re-tested in a neutral context. Prior studies failed to

do this (see Horst & Samuelson, 2008): some studies re-test with a subsequent trial of the same sort, allowing children to simply solve the problem again (e.g., Mervis & Bertrand, 1994; Wilkinson & Mazzitelli, 2003); and others provide a short review of the name-object linkages before testing retention (e.g., Goodman, McDonough, & Brown, 1998).

To address this, Horst and Samuelson (2008) presented children two known and one novel object and asked for the referent of a novel name. Children successfully selected the referent via mutual exclusivity. Five-minutes later, however, they were unable to map that same to its referent when it was presented with other novel items they had just seen. This suggests that the linkage was not retained. Thus, the use of mutual exclusivity does not necessarily result in long-term learning. It is uncertain if this generalizes to other task variants or other ages (e.g., Kucker & Samuelson, 2012; Spiegel & Halberda, 2011) or to other constraints. Nevertheless, it suggests that mutual exclusivity may simply bias the child toward the referent in the moment, and is not synonymous with learning. This questions a fundamental assumption of the constraint approach, that resolving referential ambiguity (via constraints) is the same as learning.

Cross-situational learning—If solving the problem of referential ambiguity is not the same as learning words, how do children do it? One possibility is statistical learning. In any novel naming situation the intended referent may be ambiguous. However, *across* situations, there may be only one object consistently paired with a word. For example, while the word *dog* may occur with a dog, a ball and a leash in one situation; later on it may be heard without ball or leash and with other objects. Over time the referent, dog, is likely to be the most frequently co-occurring object. Thus, at any given time, the child may not need to determine the referent—the child only needs to accumulate co-occurrence statistics to learn the mappings (Horst, McMurray, & Samuelson, 2006; McMurray, Horst, Toscano, & Samuelson, 2009; Siskind, 1996; Smith & Yu, 2008; Yu & Smith, 2007). If true, associative mechanisms (MacWhinney, 1987; Merriman, 1999; Regier, 2005) may suffice for word learning.

This idea had been examined computationally (Horst, et al., 2006; McMurray, Horst, et al., 2009; Siskind, 1996), but Yu and Smith offered the first empirical tests. Adults (Yu & Smith, 2007) and infants (Smith & Yu, 2008) were exposed to small artificial lexica that contained such regularities across trials. Both groups successfully learned the word-object linkages from this alone. It is still not known whether such learning can handle categories of objects, and this may change the computational problem. However, this provides an important proof of concept, that statistical or associative learning may proceed without solving referential ambiguity and without constraints (though they may facilitate learning or in the moment language-use).

Familiar word recognition—Finally, by focusing on information used to solve referential ambiguity, the constraint approach has little to say once the child has acquired word-object mappings (familiar words). However, familiar word recognition also changes over development, and it is not clear that this is related to constraints. Fernald et al. (1998) measured the amount of time it took infants to fixate the correct object (in a two-alternative forced-choice looking task), as a measure of the recognition of the word's meaning. This decreased dramatically over development, suggesting a tuning process for known or recently learned words (Fernald, Perfors, & Marchman, 2006; Fernald, et al., 1998). This decrease cannot be accounted for by attentional or oculomotor processes, as infants show no changes in purely visual tasks during this time, and performance in the visual task is not related to their speed of word recognition (Fernald, et al., 2006). More importantly, this improvement cannot be characterized as simply refining an existing skill; rather, speed of processing predicts the rate of long term learning (Fernald, et al., 2006) and later linguistic and

cognitive outcomes (Marchman & Fernald, 2008), suggesting it is a more fundamental property of word learning and use.

The constraint approach has no way to describe these changes, as it lacks a theory of how words are used. Since such changes are not likely driven by referential ambiguity, they require a separate developmental mechanism. To be clear, few existing accounts make a strong distinction between learning novel and familiar words. Studies like Merriman, Lipko and Evey (2008) have examined how children decide that a word is novel or familiar (implying that such decisions may help engage the right learning/processing strategies). However, this is not an essential component of any major theoretical accounts. Yet at the same time, no major theories address the improvement in familiar word recognition, nor do they seem to have the theoretical tools do so. Fernald and colleagues' work suggests an important developmental phenomenon that demands an explanation, and should be linked to the word learning literature more broadly. If an account can handle both novel and familiar word learning with the same mechanism this may offer a more parsimonious explanation of word learning.

A New Direction

While none of these findings completely rule out constraints, they paint a picture in which the solution to referential ambiguity is subtly independent of long-term learning. These problems are not limited to the constraint approach—any approach focusing exclusively on referential ambiguity and the information used to solve it will struggle to account for these findings. We need an account that emphasizes how novel and familiar words are used, and builds from there to understand how this ability develops. Such an account cannot ignore referential ambiguity. However it must move beyond it to account for development. Our thesis is that we may make more headway by considering behavior at *two timescales*. Referential ambiguity is a problem that children face in a given situation and must be solved in real-time. This differs substantially from the problem of learning and retaining word-object mappings, which may unfolds over many situations, and indeed over development.

Distilling the Word

To develop this account, we first distill "word use" and "word learning" to their minimal computational components. We define them in terms of association and activation; processes that are independent of the information that contributes to word recognition and word learning. This distinction is not theoretically novel—it builds on constructs from cognitive development that have been most extensively developed by Munakata, McClelland and colleagues working in the connectionist paradigm, and by Thelen, Smith, Schöner and colleagues in dynamic systems theory (Munakata, 1998; Munakata & McClelland, 2003; Munakata, McClelland, Johnson, & Siegler, 1997; Smith, Thelen, Titzer, & McLin, 1999; Thelen, Schöner, Scheier, & Smith, 2001; see also, Elman, 1990; Harm & Seidenberg, 1999; McMurray, Horst, et al., 2009; Spencer, Perone, & Johnson, 2009). Our goal here is to translate these concepts to word learning, and to use them to develop an account that stands independent of a strongly theoretically connectionist stance, as our ideas are conceptually compatible with other approaches that are distinct from connectionism. This account necessarily oversimplifies many things. We discuss this later. However, it allows us to be precise about mechanisms to frame them computationally.

Word use and learning fundamentally concern the relationship between a phonological pattern and a semantic category. For present purposes, we ignore the complexities of mapping sounds to word-forms, and assume that the auditory system can identify discrete word-forms. This is not trivial, but by the middle of the second year, many of the basic properties of auditory word-form recognition are in place (Fernald, Swingley, & Pinto,

2001; Swingley, 2009; Swingley & Aslin, 2002). Similarly, we assume that infants can analyze a visual scene and categorize referents. This too is not trivial, but again, by the middle of the second year, children appear adept at it (Bauer, Dow, & Hertsgaard, 1995; Behl-Chadha, 1996; Mareschal & Tan, 2007). To be clear, word learning involves mapping words to categories, not merely to individual objects (and there are excellent models that capture aspects of this: Mayor & Plunkett, 2010; Samuelson, 2002). Our goal here is to strip out these important processes to investigate the power of the associations themselves and the real-time processes that operate over them to form the basis of interesting word learning behavior. Thus we'll assume some categorization ability and focus on the mapping between word-forms and categories, and for simplicity's sake we will often refer to objects and/or referents when what is meant is a category of objects or referents.

Figure 1 shows our distillation. Circles represent representations; and their shading indicates how strongly each is considered. Panel A shows the process of identifying the referent of a word in situation-time, in a situation with two visual competitors. Initially, the system starts with every word-form under partial consideration or activation (Figure 1A, left side), as nothing has been heard yet. Two objects (*dog* and *tree*) are active reflecting the visual scene. As the word is heard, the system moves toward considering one word and object (bottom). This shift in activation represents the process of deciding what was heard and what should be attended, as in many interactive activation models (McClelland & Elman, 1986; Spivey, 2007).

Thus, resolving referential ambiguity is a matter of moving from consideration of multiple objects to one. This demands a solution in terms of activation or attention to referents, not learning. Such changes in real-time consideration of the referents could derive from external forces that decrease consideration or activation for incorrect object(s) or increase consideration of the correct one. In this example, if *dog* is unknown, eye-gaze could add consideration for *dog*; mutual exclusivity could reduce consideration of *bug* (if it is known); or a context demanding animacy could rule out the *tree* (Figure 1B). Such external forces may also include attention processes that both facilitate (Fulkerson & Waxman, 2007; Samuelson & Smith, 1998) and/or interfere with word and category learning (e.g., C. W. Robinson & Sloutsky, 2004, 2007).

Crucially, none of this has to involve a mapping, nor does it entail learning. As long as the activation of the objects reaches the correct state, there is no need to retain anything—the child has arrived at the right inference. So how does learning occur? Taking typical connectionist assumptions, initially, the system starts with word-forms and objects randomly connected – many possible mappings are under partial consideration or activation (Figure 1D, left side), as nothing has been heard or seen yet. Over time, some of these connections will be pruned, and others will be strengthened. For example, the simultaneous consideration of a word (e.g., *dog*) and an object (a dog) could lead to links being strengthened if learning is associative, and the link between *dog* and tree (which is not present) being weakened (Figure 1D). Over time, such changes could build a system of links that encompasses many words and objects. This changing and building of linkages is the result of *learning*, not external forces like attention, or pragmatics that guide activation to solve the referential ambiguity problem in the moment. If such links become strong enough, such pathways will be employed when the word is heard again, allowing the word-form, *dog* to activate the appropriate concept without external support (Figure 1C).

This suggests two distinct processes: the use or recognition of a familiar or novel word (changing activation states), and the changing of the links between the word-form and visual referent (learning or the changes of connection weights). These processes can be described on different time-scales. The problem of determining the referent of a novel word is a

problem of usage. It must be solved very rapidly in *situation-time,* that is, within the context of a single naming event, and it could take advantage of learned mappings or external support. In contrast, the problem of learning is solved over *developmental-time*: it is a problem of acquiring lasting linkages between sound patterns and meanings that may take weeks or years. For familiar words, deploying these mappings to understand or produce a word is a *situation-time* phenomenon; but enhancing the efficiency of this occurs over *developmental-time*. These problems rely on different theoretical mechanisms—situation-time processes involve changes in activation, or the strength of consideration of particular words/objects, while developmental-time learning processes involve changes in knowledge, that is, associations between words and objects (Harm & Seidenberg, 1999; McMurray, Horst, et al., 2009; Munakata & McClelland, 2003; Munakata, et al., 1997; Smith, et al., 1999; Spencer, et al., 2009; Thelen, et al., 2001)

While both problems are fundamentally about matching words and referents, the demands of these tasks are clearly different. The problem of finding a referent in the moment does not necessitate learning. Simply arriving at a state in which one word and one object are under consideration is sufficient, and does not require changing of the strength of the linkage. That's not to say that children can turn off learning, or that learning may not occur in some circumstances—indeed it is more parsimonious if learning is "always on." Rather, we are arguing these are not problems that are solved by learning processes, and it may not matter how much or little is learned in a situation for a child to find a referent. If word-object linkages are ever created in situation-time, these linkages do not need to be complete. Thus, children could make use of contextual cues to solve referential ambiguity for their immediate communicative needs, but not necessarily commit to a given mapping from one event.

Conversely, learning doesn't require the child to solve referential ambiguity. If multiple objects are under consideration, multiple linkages can be laid down. If this is done in small increments over multiple naming events, the more consistent ones could rise to the top, as in cross-situational learning (Smith & Yu, 2008; Yu & Smith, 2007). Such a process would need to be slow. If all available referents are strongly associated with a word in a single event, many erroneous linkages will be considered. If learning is too fast, these linkages could become solidified, permanent… and wrong. This fits with the fact that the average child hears 17,000 words a day (Hart & Risley, 1995), and even at their peak rate of learning, children may acquire only a handful of words in that same period (Sénéchal & Cornell, 1993). Thus, children must learn words slowly.

This framing yields enormous flexibility as situation-time processes can be optimized for the demands of speaking, comprehending and inferring, while developmental-time processes can be optimized to the demands of learning. Indeed, by moving much of the sophisticated inference of novel word meanings (classically described as constraints) to situation-time, it may allow simpler mechanisms of learning to have complex effects as learning is not entirely independent of such situation-time processes. For example, if the system attends to a referent longer in some circumstances than in others, more learning may result; or if competition between referents resolves faster (in situation-time), the system may be able to acquire more unambiguous associations. Similarly, changes in situation-time familiar word recognition could be produced by simply improving the strength of the links between referents and their auditory word forms or by eliminating unnecessary connections.

This approach also addresses the relationship between familiar and novel words. To the extent that any partially formed mappings are available for a novel word, the system may use those partial mappings to increase consideration to the correct object and decrease consideration to erroneous objects, a form of mutual exclusivity. Thus, novel word

recognition may take advantage of familiar word processes. Similarly, familiar word recognition may be enhanced by the fact that well-learned words would have stronger associations than newly learned ones, a type of frequency effect. Indeed, the process of tuning these connections to augment familiar words can derive from the same learning mechanisms that establish them. If both novel and familiar words harness the same processes, there's no reason to ignore external support (e.g., pragmatics, semantics) in even familiar word recognition, something of recent interest in adult sentence comprehension (e.g., Chambers, Tanenhaus, & Magnuson, 2004; Hanna & Tanenhaus, 2004; Tanenhaus & Brown-Schmidt, 2008).

Bringing together novel and familiar word learning can unify the literature. It also reframes how we think about classic findings by treating processes that have been described as constraints as descriptors of referent selection, not learning. Finally, it eliminates the classic distinction of fast- and slow-mapping. Carey (1978) described fast-mapping as a special stage of initial learning where links between words and meaning were first forged. In our view it may reflect purely situation-time referent selection. Indeed, as words are acquired, there are complex changes as connections are built and pruned, but there is no reason to assume the underlying processes are any different the first time. Thus, as the term "fast-mapping" conflates learning and situation-time processing, we avoid it. Instead, we use referent selection to refer to situation-time behavior in ambiguous naming situations, and learning to refer to developmental-time changes.

Mechanisms of ambiguity resolution and learning

Our *dynamic associative* account makes two claims. First, novel word processing (referent selection) is not distinct from familiar word recognition. Second, and more importantly, word learning is the interaction of situation-time processes that give rise to referent selection and familiar word recognition (word use), and developmental-time processes that give rise to retention and the improvement of these abilities. Here we propose mechanisms. Situation-time processes arise out of dynamic competition between referents and/or words, and developmental-time processes arise from associative learning harnessing cross-situational statistics.

Dynamic Competition

In connectionist models, any time multiple items are considered in parallel, some form of competition is present. Indeed, several models of word learning hypothesize probabilistic representations with this property (MacWhinney, 1987; Merriman, 1999; Regier, 2003, 2005; Xu & Tenenbaum, 2007). These provide a good starting point; however, they do not describe how these probabilities unfold over situation-time. While this is important for modeling behavior like reaction times or eye-movements, one could in principle derive simple linking functions to map these probabilistic representations to reaction times. However, this may not be sufficient. Computational models have shown when competition unfolds dynamically over time, unexpected effects can occur – gangs of weakly active representations can suddenly inhibit a more active one (Spivey, 2007). Since we want to explore situations like mutual exclusivity characterized by such ambiguity, and investigate how these unintended consequences can shape learning, it is crucial to implement competition as a dynamic process that unfolds over time.

Dynamic competition has been postulated in a number of domains involving constraint satisfaction and ambiguity resolution including music (Bharucha, 1987), syntactic parsing (MacDonald, et al., 1994; McRae, Spivey-Knowlton, & Tanenhaus, 1998), comparison (Goldstone & Medin, 1994), visual scene organization (Vecera & O'Reilly, 1998), visual categorization (Spivey & Dale, 2004), visual search (Spivey, 2007), speech perception

(McClelland & Elman, 1986), and language production (Dell, 1986). Many of these problems are clearly relevant for our domain. These computational approaches all incorporate simple neural-like units with graded activation. Activation feeds forward (from perceptual inputs to higher-level representations) and backward (from higher-level representations) while competing within a level over time. The resulting pattern of activation usually represents the best solution given the constraints imposed by the input, top-down expectations, and the structure of the network. Typically, after this process, a single unit is active, and activation for competitors is suppressed, offering a close fit to our distillation of referent selection. There are a number of computational formalisms for this including interactive activation (Dell, 1986; McClelland & Elman, 1986), normalized recurrence (Spivey, 2007) and dynamic field theory (Schutte, Spencer, & Schöner, 2003; Thelen, et al., 2001), which share these basic properties.

In our model, referent selection is modeled with dynamic competition using Normalized Recurrence (McRae, et al., 1998; Spivey, 2007). Normalized Recurrence has been applied to problems that are related to referent selection including speech perception (McMurray & Spivey, 2000), visual categorization (Spivey & Dale, 2004) and visual search (Spivey, 2007, chapter 8), and has been used to map lexical activation for known words to objects (Spivey-Knowlton & Allopenna, 1997; Spivey, 2007, pp. 187-200), embodying our distillation of word recognition for familiar words. Most importantly, versions of this architecture combine this competition with unsupervised learning (McMurray, Horst, et al., 2009; McMurray & Spivey, 2000).

To model referent selection, words and referents are modeled as localist units. On any trial, one word and multiple objects are active. Words and referents pass activation to a lexical layer (what Spivey, 2007, terms a "decision layer"), and recurrent competition among all three layers forces the network to suppress activation for the objects that do not map to the word. We discuss the motivation for localist representations in the general discussion. However, at a purely practical level, virtually all of the above referenced competition architectures use localist units, and it is difficult to implement competition in a distributed representation.

Associative Learning

A number of researchers argue that word learning cannot be associative: the fundamental mechanisms are social (Golinkoff & Hirsh-Pasek, 2006; Nazzi & Bertoncini, 2003), referential/conceptual (Waxman & Gelman, 2009), or constraint-based (Woodward & Markman, 1998). While such accounts describe important sources of information, and/or important representational issues, it is not clear what these mean for learning because terms like social, referential or conceptual learning do not have clear definitions in learning theory.

Some of these non-associative accounts still argue that *early* word learning may be associative (e.g., Golinkoff & Hirsh-Pasek, 2006; Nazzi & Bertoncini, 2003; see also Namy, 2012). This is almost a necessity—there is little lexical knowledge to facilitate mutual exclusivity; and social skills like the use of eye-gaze are still developing (e.g., Moore, 2008). These accounts typically argue that more complex mechanisms like constraints or social pragmatics take over later. Thus, such accounts posit a discontinuity in the learning process, but even with this discontinuity, they do not offer an explanation for improvements in familiar word recognition.

Such accounts critique a straw-man version of associationism in which raw perceptual inputs are directly associated without processing or intervening representations (c.f., McMurray, Zhao, Kucker, & Samuelson, in press). Indeed, this critique seems to focus on the information that is associated, not in the mechanisms by which the linkages are made

(c.f., Smith, 2000). In contrast, modern learning theory admits internal representations as a basis of association, and allows attention and other factors to shape the strength of these associations (Livesey & McLaren, 2011; Shanks, 2007). This is also central to connectionist learning. Our model does both: associating visual and auditory inputs to a lexical concept and allowing competition to shape their strength. If associative learning uses abstract representations and sophisticated situation-time processing, there is no reason to abandon it after the initial words. Indeed as we described, by allowing social inference or constraints to shape in-the-moment processes, learning may still be associative at its base while leveraging these richer sources of information.

Under our view, learning is the same whether children are in the so-called association phase, or the so-called constraint-based, referential or social/pragmatic phase. The distinction highlighted by these theories is in terms of the *information used* during learning and novel word inference, not the learning *mechanism*. This makes a simple story. Basic learning mechanisms handle the retention of information. Initially co-occurrence may be the only source of information available to them, but later, as the child learns to use other information in the environment and make more robust decisions about what the referents of words are, these form more precise activation patterns in situation time, which enables richer and faster associative learning—but the associative learning is the same. In this way, the time-scale distinction allows us to distinguish and relate these processes. Principles and constraints identify the relevant information in situation-time for the purpose of using words; while associative learning processes build the correct mappings over developmental-time. By buttressing associative learning with dynamic competition to handle the situation-time ambiguity resolution, we may achieve a significantly more powerful word learner as in any novel naming instance, the learner will have more activation for words and/or referents (enabling stronger associations to be built more quickly) and will have less activation for competing referents, preventing the formation of spurious associations. Crucially, this can be accomplished without having to posit qualitative distinctions between learning mechanisms at different ages.

This framework may also help connect vocabulary learning to classic findings in learning theory that are directly relevant to words. These include phenomena like the power-law of learning (Heathcote, Brown, & Mewhort, 2000; Logan, 1992; Newell & Rosenbloom, 1981; and see Section 2), the role of similarity (Palmeri, 1997; Storkel, Armbruster, & Hogan, 2006; Swingley & Aslin, 2007; Wifall, McMurray, & Hazeltine, submitted), the role of statistics (Yu & Smith, 2007), and even old phenomena like cue-neutralization (Apfelbaum & McMurray, 2011; Bourne & Restle, 1959; Bush & Mosteller, 1951; Rost & McMurray, 2010).

Thus, our goal was to investigate the consequences of associative learning when embedded in this richer framework of multiple timescales and internal representations. Our model uses perhaps the simplest form of associative learning, Hebbian learning. Inputs (words or objects) will be associated with an internal lexical unit if both are active; otherwise, the association decays. As in the case of competition both implementing and understanding such learning makes the most sense with localist units.

Even within this simple approach, there are layers of complexity. First, associations connect word-forms and object-categories to lexical concepts, not to each other. These lexical concepts function something like lemmas – abstract representations that connect other representations. Their presence means that learning requires at least two connections (word→lexicon; lexicon→object). Second, learning must not just build connections, but also avoid or eliminate unnecessary ones (c.f., Regier, 1996). Consider the connections between visual and lexical units (Figure 2). If the network heard *dog* in the presence of a

dog and a bug, the most salient connection is the positive association between the object-category *dog* and its lexical unit (thick line). However, this system should also learn that the object-category *dog* is *not* associated with the word *tree* (which was not heard), a negative association (dashed line). It also needs to reduce the negative association between the object-category, *tree*, which is not present, and the word, *dog* (dotted line). Thus, for successful learning to occur, it must increase one association and decrease two. In a larger lexicon, there will still be one positive connection, but there will now be hundreds of spurious connections to prune[1].

What is the source of these spurious associations? As we described, these may exist from the earliest stages of learning. In connectionist models connection weights start from small random values, a necessity in unsupervised learning (e.g., Rumelhart & Zipser, 1986). Some may also be formed during learning when incorrect objects appear with the referent (e.g., *bug*). Either way these spurious associations will also need to be pruned. This pruning must occur in a way that preserves plasticity for new words. If connections too all words and referents were pruned equally on each naming event, potential positive connections for novel words and categories that have not been heard or seen would be pruned as well, making such words difficult to learn. Rather, we need a form of pruning which preserves potential connections for novel words/referents, but still allows the system to refine its connectivity for familiar words/referents. Thus, this pruning process, which is *not* often what comes to mind when we think of learning a word, could be an important determinant of development, if only for its massive scale.

Overview of Architecture

At the broadest level, our dynamic associative model captures short- and long-timescale dynamics by embedding a model of learning in a model of use. It ignores the complexity of auditory and visual processing, to link word-forms to object-categories. However, it does not link them directly. Rather, word-forms and categories each link to intermediate lexical representations. We implemented this in a hybrid connectionist/ dynamic systems approach. Word-forms and categories are represented by localist units, which compete in situation-time using normalized recurrence to push the network toward a single interpretation. During competition, connections between inputs and the lexical layer are tuned via associative (Hebbian) learning. As we discussed, localist representations are the most transparent way to implement these mechanisms. However, they also offer a theoretical advantage: by stripping out the emergent power of distributed representations they allow us to isolate these mechanisms and observe their consequence in a more or less pure state. Our goal is to investigate the power of this unique combination of basic mechanisms.

Relationship to standard connectionist approaches

By situating learning in connection weights and situation-time processing in changing activation, our approach is broadly consistent with classic connectionist thinking on development (Elman, 1990; Harm & Seidenberg, 1999; Munakata, 1998; Munakata & McClelland, 2003; Munakata, et al., 1997). In models like these, such a description can divorce emerging knowledge from the ability to use it in a real task. For example in Munakata's models, the model may have some latent ability to represent an object under occlusion, but this can be overwhelmed in the moment.

Our model shares these broad properties, though it differs in theoretically important ways. First, as in many models, situation-time processing occurs over recurrent connections

---

[1]We use the term "prune" here as a vivid metaphor. However we intend a more graded sense in which spurious connections are slowly and gradually reduced, not irrevocably eliminated in one step.

between units. However, we argue that competition is the essential element of this situation-time processing (in word learning), something other models (Elman, 1990; Harm & Seidenberg, 1999; Munakata, 1998) have not explored. Second, many models of this sort make the simplifying assumption that each time-step corresponds to the presentation of a single input, what Schlesinger and McMurray (in press) term "trial-time"[2] (though see Harm & Seidenberg, 1999). In contrast, in this domain it is important to model ongoing processing after the presentation of an input, both to model things like reaction time, and also because settling can cause the network to change its interpretation of the input, not just strengthen whatever is already the most active (c.f., McClelland & Elman, 1986; Spivey & Dale, 2004). Third, all of the prior networks have used more powerful supervised learning, while we argue that unsupervised learning may be fundamental to word learning. Thus, our framework is built on classical connectionist thinking about learning and processing, but makes specific and theoretically motivated decisions about how to use these constructs that have not appeared in prior models of cognitive development.

Why Another Computational Model of Word Learning?

There are a many computational models of components of word learning, examining topics ranging from the sequencing of phonological material (Gupta & Tisdale, 2009; Sibley, Kello, David, & Elman, 2008); to generalization of category membership (Colunga & Smith, 2005; Li & MacWhinney, 2004; Mayor & Plunkett, 2010; Samuelson, 2002); to embodiment (Roy & Pentland, 2002; Yu, H., & Aslin, 2005). There are several models of the word-referent mapping problem (Frank, et al., 2009; Li & MacWhinney, 2004; MacWhinney, 1987; Mayor & Plunkett, 2010; Merriman, 1999; Regier, 2005; Siskind, 1996; Xu & Tenenbaum, 2007). These models make valuable contributions, highlighting the information that is relevant to the problem (Siskind, 1996), the power of associative mechanisms (Colunga & Smith, 2005; Li & MacWhinney, 2004; Mayor & Plunkett, 2010; Merriman, 1999; Regier, 2005; Samuelson, 2002); and how constraint-like behavior emerges from simpler systems (Merriman, 1999; Regier, 2005).

A complete analysis of these models is outside of the scope of this article (see Frank, et al., 2009; Regier, 2003; 2005 for useful reviews), and we are not proposing a competitor to them. Our goal is not to develop a complete model of word learning, but rather to use computational tools to investigate the emergent consequences of theoretical ideas (c.f., Schlesinger & McMurray, in press). In that sense, it is important to address what questions these models have been used to answer, and any limitations that may inhibit their ability to address our questions.

First, by not explicitly capturing both time-scales, many models do not succeed in modeling certain phenomena, or are forced to treat problems like referent selection as developmental-time phenomena. For example, MacWhinney (1987), Merriman (1999) and Regier (2005) incorporate something resembling competition in their probabilistic representations. But they do not incorporate situation-time dynamics, nor distinguish between aspects of the problem that must occur in situation-time (e.g., referent selection) from those that occur over developmental-time. Consequently they discuss things like mutual exclusivity as a limit on *learning* (e.g., it's difficult to learn a second label for an object), rather than a principle of referent selection. Moreover, without implementing dynamic competition, these models cannot investigate the emergent interactions between competition and learning.

---

[2]Though this is a misnomer with respect to the Munakata models as each trial consists of multiple input presentations (the sequence of visual inputs). The fact that there is not significant cycling of the network after each presentation of the input is the relevant factor here.

Bayesian models (Xu & Tenenbaum, 2007), in their focus on how interacting constraints lead to accurate inferences, offload many constraints onto priors (and add new ones). This makes it difficult to understand how these behaviors develop or could arise from simpler processes. More recent Bayesian approaches (Frank, et al., 2009) do address independent time-scales, acknowledging the demands of both long-term learning and referent selection, and offer an example of how statistical learning can be buttressed by social cues. But lacking situation-time dynamics or a developmentally plausible learning model, such approaches only serve as a meta-level description of the information used for word learning (Jones & Love, 2011).

Regier's (2005) and Mayor and Plunkett's (2010) models are both associative and perhaps closest to ours. In some respects they go further, examining the contributions of auditory and visual similarity. In others they are more limited. At the level of associative learning, both models include rich representations of auditory word-forms and basic categories, but do not include abstraction between them, so they may not be able to generalize word knowledge to other processes such as action systems, spatial processes or orthography. Moreover, competition only arises from probabilistic representations—not true dynamic competition. As a result, they have not investigated the consequences of situation-time competition (although these approaches are likely not opposed to this). Further, these models emphasize issues in word learning that are now understood differently. Regier focuses on accelerating vocabulary growth, when McMurray (2007) suggests this is a property of many parallel learning systems. Both models simulate fast-mapping by mutual-exclusivity as one-shot learning, but Horst and Samuelson (2008) suggests this may not be required, a model could perform well in the moment even if little learning occurred (Horst, et al., 2006). Finally, neither deal with referential ambiguity, assuming a form of ostensive naming. This, however, is a limitation of implementation – such models may be able to cope with many of the phenomena we examine here, given their theoretical overlap.

Finally, Samuelson (2002) and Colunga and Smith (2005) present semi-supervised associative learning models that do incorporate settling dynamics (though not competition). However, they do not attempt to model the actual learning problem. Rather than examining referential ambiguity, they focus on how associative learning can lead to generalization and to the identification of relevant dimensions for object categorization. Moreover, they do not use the settling dynamics to explain situation-time behaviors like referent selection.

Thus, our model is consistent with features of many models: it incorporates associative learning (MacWhinney, 1987; Regier, 2003, 2005; Samuelson, 2002), and examines the role of statistical structure (Frank, et al., 2009; Siskind, 1996) using graded or probabilistic representations (MacWhinney, 1987; Regier, 2005; Xu & Tenenbaum, 2007). It examines the emergence of constraints (Colunga & Smith, 2005; MacWhinney, 1987; Regier, 2005; Samuelson, 2002), and its architecture combines situation-time and learning processes (Colunga & Smith, 2005; Samuelson, 2002). However, it builds on these notions to examine a central new issue: the emergent power of how mechanisms interact across two time-scales. The power of this combination is illustrated by distilling the problem to the point where such mechanisms operate over a minimally informative set of representations (e.g., word forms and concepts stripped of their phonetic, visual, and conceptual processes), and by embedding them within a cross-situational learning framework in which words can only be learned via co-occurrence statistics.

Our goal is not to present a complete model of word learning, but to use our dynamic associative model to learn how these processes interact. This can test the sufficiency of these processes to account for a range of phenomena. More importantly, we can use the model to clarify how these processes are related and develop a theoretical framework on which to

base empirical investigations. This allows us to ask whether associative learning can cope with referential ambiguity, and if children must solve this problem to learn words; how online processes and learning interact; whether processes that underlie familiar word recognition give rise to mutual exclusivity; and whether constraints like mutual exclusivity emerge without being built in. Answering these will help develop a theoretical approach bigger than any one model.

Specific architecture

Our dynamic associative model (Figure 3) has two layers of localist inputs, for auditory word-forms and visual objects. Each auditory unit corresponds to a single word; and each visual unit corresponds to one category of possible referent. During processing, the auditory and visual layers are normalized such that the sum across each layer is 1.0. Thus, if a single node were fully active, its activation would be 1.0; if two were active, each would be 0.5; and when all nodes are inactive (the resting level), they are set to 1/N where N is the number of nodes. After normalization, the vector of activations across a layer can be read as the distribution of likelihoods that the auditory (or visual) hypothesis represented by that node is present.

There are no direct connections between auditory and visual units. Interactions occur because both connect to a hidden layer of lexical units. These weights are initially random such that each auditory and visual unit is connected to each lexical unit (with differing strengths; Figure 1A). However, after learning, these weights generally reflect one-to-one mappings between word-forms and lexical units, and between referents and lexical-units (Panel B).

The model contains more lexical units than needed to represent all the words. This leads to better learning. Generally, Hebbian learning reinforces existing associations and prunes unnecessary ones. Thus the first few connections formed during learning are crucial. If the network initially randomly associated two inputs with the same lexical node, this mis-mapping could be reinforced over subsequent inputs, and prevent the network from ever discriminating them. Greater numbers of lexical units make this much less likely (McMurray & Spivey, 2000).

On each trial, a single auditory unit (a name) is activated (set to 1.0). An array of visual units (the objects in the scene) is also activated. Input layers are normalized and activation from each layer is sent concurrently to the layer of lexical units (Equation 1).

$$\Delta l_x = \left(\sum_{z \in a} w_{xz} a_z + \sum_{z \in a} u_{xz} v_z \right) \quad (1)$$

The change in activation of a lexical unit is based on the net input to that node: the sum of all of the auditory units (a_z, for the zth auditory unit) multiplied by their connection weight (w_{xz} which connects auditory node x to lexical node z), and the corresponding weighted (u_{xz}) activation of the visual input units (v_z). This, is multiplied by the feed-forward temperature (τ_{ff}), which controls the weighting of the prior activation in setting the new activation (we use the superscripted $l^{(t)}$ to denote l at time t).

$$l_x^{(t+1)} = l_x^{(t)} + \tau_{ff} \cdot \Delta l_x^{(t)} \quad (2)$$

After updating the activations, lexical units inhibit each other using squared normalization (3).

$$l_x^{(t+1)} = \frac{\left(l_x^{(t)}\right)^2}{\sum\limits_{z \in L}\left(l_z^{(t)}\right)^2} \quad (3)$$

Since activation in the lexicon is always between 0 and 1, the result of this is that highly active units become more active and dominate less active ones[3]. When run repeatedly over several cycles, this approximates winner-take-all competition.

The resulting lexical activation then feeds back to the input layers, (Equations 4, 5). The change in activation of an auditory (or visual) unit, y, is proportional to the product of its previous activation and the weighted activation (w_{yz}) of the lexical units (l_z).

$$\Delta a_y = a_y \cdot \sum\limits_{z \in d} l_z w_{yz} \quad (4)$$

$$a_y^{(t+1)} = a_y^{(t)} + \tau_{fb} \cdot \Delta a_y^{(t)} \quad (5)$$

The fact that feedback is multiplied by the current activation means that feedback from the lexical layer only affects *active* input units. This prevents the network from activating perceptual inputs on the basis of top-down evidence alone and it also introduces nonlinearity into the system. The temperature parameter, τ_{fb}, is not required to be the same as for feedforward activation and was not for the simulations reported here.

After updating, the activations of the auditory and visual layers each undergo a small amount of inhibition (Equation 6) and are normalized.

$$a_y^{(t+1)} = \frac{\left(a_y^{(t)}\right)^I}{\sum\limits_{x \in A}\left(a_x^{(t)}\right)^I} \quad (6)$$

Here, I represents the degree of inhibition. At I=2, this would yield strong inhibition as in the lexical layer; at 2>I>1 there is less inhibition; at I=1, no inhibition; and at I<1 the activation collapses back to the resting state. After this step, activation feeds forward to the lexical layer and the cycle continues. Activation cycles in this way until the lexical layer settles (i.e., the change in lexical activation from time-step to time-step is close to zero).

Typically on any trial the network is presented with multiple visual units to simulate a cluttered scene. Throughout cycling, the network partially considers each visual competitor simultaneously, but after many cycles, the competition and feedback generally result in a single object having more activation than the others. This active visual unit is the network's response—the network is allocating more attention to this referent. Thus, recurrent cycling causes the network to settle into an activation pattern across all three layers that reflects the present constraints (the auditory and visual inputs) and partial knowledge (in the weights).

Connections are modified at each cycle with Hebbian learning (Equations 7, 8). The network increases the strength of the connection between simultaneously active input and lexical units, and decreases the connection in other cases.

---

[3]This particular form of inhibition instantiates a form of lateral inhibition in which the ability of each unit to inhibit the other units is a function of its proportion of the total activation.

$$\begin{aligned}
\Delta w_{xy} = \quad & a_x l_y \left(1 - w_{xy}\right) \\
& -.5 \cdot (1 - a_x) \cdot l_y w_{xy} \quad (7) \\
& -.5 \cdot a_x \left(1 - l_y\right) \cdot w_{xy}
\end{aligned}$$

In Equation 7, the first line represents the positive term. If a_x (auditory unit x) and l_y (activation for lexical unit y) are active, the weight is increased proportional to its distance from 1 (its maximum value). The second and third terms are decay terms. A given weight decreases if 1) the input unit is active and the lexical unit is not, or 2) if the lexical unit is active and the input is not. If neither is active there is no decay. By restricting weight decay to only connections between units that are actually used at that point in time, the model maintains plasticity in weights connecting input and lexical nodes that are not used. This is crucial for learning new words in the future (for a similar learning rule, see Grossberg, 1976). Weights are updated with (8)

$$w_{xy}^{(t+1)} = w_{xy}^{(t)} + \eta \cdot \Delta w_{xy}^{(t)} \quad (8)$$

Here, η is the learning rate and is typically very small—on the order of .0005. This is because learning occurs on each cycle of competition (and with many cycles/input this will add up). By learning continuously, rather than at the end of processing, we need no homunculus controlling when learning can occur, much as children may not differentiate between training and test trials in the laboratory from other learning opportunities.

An important question is what regularities in the input drive functional learning. We examine one possibility here: co-occurrence between words and referents, or cross-situational statistics (Yu & Smith, 2007). We implemented this style of learning by ensuring that among the set of visual competitors active on any given trial, one was consistently paired with the auditory target while the others were randomly selected.

General Methods

For most models, a 35-word lexicon was used. Though the network can learn larger lexica (Online Supplement, S2), 35 was sufficient to be interesting while allowing the network to run reasonably quickly[4]. Thus, models were initialized with 35 input units and 500 lexical units. Weights were initialized to random values, generally between 0 and .5 (the *wtsize* parameter).

Training

Models were generally trained for 200,000 epochs, where an epoch is one presentation of a word (though this entails many cycles of competition as the word is processed). While the model can *perform* quite well after only a handful of exposures to a word, it may take several thousand for the weights to settle on a single strong association. Thus, a long period of training was important for understanding both early and late stages of learning.

On training trials, a single auditory unit was activated, accompanied by the corresponding visual unit and a variable number of competitors. The average number of visual units active across trials is the degree of *referential ambiguity* present in naming situations for that model. Thus, if on average 14 of 35 units were active, the model faced 40.3% referential ambiguity. On any given trial, the active visual units were determined by first choosing a

---

[4]A typical model completed training in 30 minutes to two hours, but for each simulation we typically ran many repetitions of each model in several conditions, requiring several days.

vector of 35 independent random values between 0 and 1. From this, any unit whose random value was less than the level of referential ambiguity would be active (e.g., for a referential ambiguity of 20%, if the random value for a unit was .2 or lower, this unit would be activated). Consequently, the number of competitors was not constant over trials, though the mean level of ambiguity was.

Testing

While many tests of the model were particular to a simulation, most models were tested in the following ways: 1) a simulation of an N-alternative forced-choice referent selection task; 2) a production task; and 3) an analysis of the weight matrix.

The N-alternative forced choice (NAFC) task is similar to what is used with children (e.g., Golinkoff, Hirsh-Pasek, Bailey, & Wenger, 1992; Mervis & Bertrand, 1994). In this procedure, a single auditory unit is activated with N visual units (typically three). One of these is the target and the others are randomly selected. Activation settles through the model until the lexical layer stabilizes. The activation in the visual units then determines the model's response. To correctly identify the word, the model must pass activation from the auditory layer to the visual layer, strengthening activation for the correct visual unit, and suppressing the competitors. This can be interpreted as the model attending to one object (a look or a reach in a child). This is repeated for each word to estimate the number of words known. This test emphasizes the model's observable behavior in a simulated experiment, rather than analyzing its unobservable intermediate states (the lexical units).

We also created a production task, to simulate laboratory naming tasks, and to simulate the "child says" version of the MacArthur-Bates Communicative Development Inventory (MCDI), which is commonly used to assess which words a child knows. Here, we run the model in reverse. This time, a single referent is active and the system must select from *all* the auditory word-forms—all of the auditory units were partially active (1/N). In this task, there are no constraints on which word-forms are considered and there will be typically many more auditory competitors than visual competitors in a comprehension task. Activation settles through the model until lexical layer stabilizes. Here, the most active auditory unit serves as the response. This is repeated for each word to estimate the model's productive vocabulary.

In addition to these tasks, we also analyzed the weight matrix. This is not comparable to anything that can be done with children, but it assesses learning more directly. For each word, we first examined the auditory→lexical weights to determine which lexical unit was most strongly connected to that auditory unit. This was repeated for the visual inputs. If this was the same unit, it would imply that the model correctly formed the association between auditory and visual representations for this word and this word was scored as correct. Occasionally some instantiations of the model selected the same lexical unit for two word/ object pairs, implying that both words meant the same thing. Because this was incorrect (in the lexicons used in the first three sections), only one of the two words would be scored as correct. Thus, this analysis asks if the model has achieved a representation of the word that is close to an idealized state in which each word corresponds to one object (and vice versa).

Each of these tests was often repeated over training, raising the possibility that the model could learn during test. Thus, unless we've noted otherwise, any weight changes that accumulated during testing were discarded before continuing with training.

Overview of Simulations

Our first goal is to show that this dynamic associative model can account for a variety of empirical results. We do not do this to emphasize the fit of this specific model. Rather our model is representative of a broad class of interactive and associative mechanisms. Thus, its ability to fit the data really emphasizes the strength of these kinds of approaches, and of approaches operating at two timescales more generally. A second, more important, goal is to use the model to understand the processing mechanisms that underlie word learning phenomena like mutual exclusivity, familiar word recognition, and statistical learning. Thus, our simulations alternate between simulating empirical phenomena and unpacking the models' information processing.

Table 1 shows a summary of all of the simulations presented here with citations to relevant empirical studies. Section 1 examines learning. We ask whether the model can learn under referential ambiguity (Simulation 1.1), and whether it shows differences between comprehension and production (Simulation 1.2). We next ask how much referential ambiguity the model tolerates (Simulation 1.3) and the shape of vocabulary growth (Simulation 1.4). These demonstrate the basic phenomena of word learning, and the role of word-use in explaining them.

We next examine situation-time phenomena. Simulation 2.1 examines the timecourse of familiar word recognition and Simulations 2.2 and 2.3 examine the use of mutual exclusivity for referent selection. Both are assessed over development and our analysis suggests that even these *situation-time* phenomena are fundamentally shaped by *developmental* forces.

Section 3 examines the interaction of situation- and developmental-time processes. Simulation 3.1 examines the apparent independence of these timescales, examining referent selection and retention (Horst & Samuelson, 2008). Next, we examine phenomena arguing for more dependent timescales: the interaction of task and mutual exclusivity (Simulation 3.2), and individual differences in familiar word recognition (Simulation 3.3). Finally, Simulation 3.4 manipulates situation-time processes in the model to show that they are necessary for learning.

Section 4 scales the model up in two important ways. First, Simulation 4.1 trains the model on both basic-level and super-ordinate labels to show that despite the model's use of mutual exclusivity, it can learn multiple labels for a referent. This also affords the opportunity to examine basic-level advantages in a system that is not hierarchical. Simulation 4.2 generalizes this finding, training the network on multiple referents for a word (e.g., homonyms or polysemy), or multiple labels for a given object (e.g., bilingualism, synonymy). As these can disrupt the one-to-one word-object mappings commonly thought to underlie referent selection by mutual exclusivity, we evaluate both learning and mutual exclusivity.

A number of additional simulations were also conducted and reported in the online supplement. These support the simulations presented here, and do not introduce new theoretical points. Commented Matlab code for all simulations is also available in the supplement.

Section 1: Developmental time processes

We first ask if the model can learn words under referential ambiguity, and whether differences in real-time comprehension and production tasks account for the commonly observed delay in productive vocabulary growth relative to comprehension. Next we assess how much referential uncertainty can be tolerated, and the shape of learning.

Simulation 1.1: Learning, measured by comprehension

Our first simulation validated that the model can learn under referential uncertainty, using the simulated laboratory comprehension task. We initialized ten models to learn 35 words under 50% ambiguity (on average 17 competitors were present on every trial—a substantial degree of ambiguity). Every 1000 epochs we tested the model in the 3AFC and 10AFC tasks, and analyzed the weights (Table 2 shows the parameters for all simulations in Section 1).

Results—Figure 4 shows the number of words learned as a function of time (training). The perceived rate of learning is largely a function of the task. At 50,000 epochs, the model appears to know most of its lexicon when tested in the 3AFC task, and can do fairly well in the 10AFC task. However, this model has not finished learning. The weight analysis shows only 1.5 words are known—for most words, the corresponding auditory and visual units do not have strong associations to the same lexical unit. This suggests that competition can make the model outperform its stored knowledge, and even after the model can identify the referent, there is still significant learning to do. Some have likened this "slow-mapping" process to an elaboration of the meaning, syntax and phonology (e.g., Capone & McGregor, 2005). This undoubtedly occurs, but this model has no semantics, syntax or phonology. Thus, the raw word/object mappings may need to undergo a similar process in which competing weights are gradually eliminated, and the correct connections sharpened, even after the model demonstrates understanding of the word (Riches, Tomasello, & Conti-Ramsden, 2005). Crucially, however, we also observe a disconnect between the model's knowledge (which is poor at early points in training) and its performance (which is simultaneously quite good). Here, task constraints, such as the reduced number of competitors, actually allow the model to perform better than its knowledge, in an apparent reversal of the performance competence distinction.

Simulation 1.2: Production and comprehension

Simulation 1.1 suggests a dissociation between what a model "knows" (the associative connections), and what it "does" (performance on comprehension tasks). This raises the possibility that vocabulary growth may appear to follow a different developmental timecourse when measured with production and comprehension tasks.

Such differences are observed empirically: children typically comprehend more words than they can produce (Fenson, et al., 1994; Reznick & Goldfield, 1992). While this is often attributed to memory demands, difficulty planning articulation, or the earlier age at which speech perception develops, in the model, such differences cannot arise from these factors. However, as we've described them, comprehension and production are different tasks, as comprehension requires selecting from a small number of referents, while production requires selecting from a vast number of word-forms. Thus, this simulation asks if task differences alone account for part of the delay in production. This is not to say that such differences are artifacts of the tasks used to assess them—they are fundamental to comprehension and production more generally.

Twenty models were trained at 50% referential ambiguity (Table 1). Every 200 epochs they were tested on a number of comprehension tasks: 3AFC, 5AFC, 10AFC and 35AFC (as if the entire lexicon were visually present) and a 35AFC production task.

Results—Figure 5 shows the number of words learned over time for each task. As before, even when 50% of the lexicon was present on any trial (~17 competitors), the model acquired the full lexicon no matter how it was assessed. Moreover, the 3, 5, and 10AFC tasks suggested that like children, in comprehension tests, the network appears to know

more words than in production tests. However, by the end of training the network performs equivalently for both tasks. Interestingly, the production measure matches the estimates based on the weight analysis. Thus, more difficult naming tasks may tap this level of competence.

The number of response alternatives plays a critical role in predicting performance and the apparent rate of learning. To the extent that the response-set for comprehension must include fewer objects than the full lexicon, task differences can account for differences between comprehension and production. Similarly, Huttenlocher (1974) describes this in information processing terms: comprehension is a recognition task, while production is a recall task. Our model instantiates both processes as variants in the same underlying competition dynamics.

This might suggest that the number of response alternatives alone dictates success. However, this is not entirely the case as a comparison of the 35AFC comprehension and production tasks show that now *comprehension* is delayed. This is due to an asymmetry during training. Auditory units are always presented singly, while visual units are not. As a result, the network has experience suppressing unnecessary visual units, but has never suppressed unnecessary auditory units[5]. Nonetheless, this small difference illustrates that task differences can arise both from differences in situation-time factors (the number of response alternatives) and developmental-time factors (the history of suppressing competitors).

Critically, however, this distilled account of production vs. comprehension suggests that 1) the number of response alternatives during testing can dictate how many words a child appears to know; and 2) this somewhat obvious fact can give rise to differences in production and comprehension vocabulary.

Simulation 1.3: Learning under referential ambiguity

Thus far, we've held referential ambiguity at 50%. This is substantial, yet we found excellent learning. It is important to determine the robustness of this learning, particularly when buttressed by dynamic competition. Thus, we varied the degree of referential ambiguity from 20% to 95%, and trained 10 models at each level. Note that 95% means that on average 33.25 objects were present with the referent in any naming situation, and on 16.5% of the naming instances all 35 words are active. Models received 200,000 training trials, and we assessed performance in the 3AFC and 10AFC tasks as well as a weight-matrix analysis every 25,000 epochs.

Results—Figure 6A displays the number of words identified in the 3AFC task as a function of referential ambiguity. Chance is 33% (11.7 words). At low levels of noise, the model acquired most words within about 25,000 epochs and learned all of them by 100,000 epochs. At 100,000 epochs an effect of referential ambiguity is seen: the model's performance drops off as with more competitors (though not very far). However, this is overcome with additional training: at 200,000 trials, the model performed at 100% even at 95% ambiguity.

This success is emphasized by our more conservative analysis of the weights (Figure 6B). Here chance is much lower: the probability of randomly mapping a single auditory and visual unit to the same lexical unit is $1/500*1/500= 0.0004\%$. Moreover, to pass this test, the model cannot rely on competition to arrive at the best guess if the weights are imperfect (as

---

[5]Of course, with a more realistic auditory representation, multiple auditory word-forms will be active in parallel (e.g., similar sounding words Allopenna, Magnuson, & Tanenhaus, 1998; Marslen-Wilson, 1987). This may minimize the differences seen when the number of alternatives is equated.

it can in the 3AFC task). By this criterion, it takes much longer to learn a word. At 50,000 epochs, the model's 3AFC performance is good, but its underlying competence (the weight matrix) is far from complete. For example, at 50% referential ambiguity the model has only learned 10 words by this point in time. However, with enough training, the model performs well on this conservative measure even at the largest degree of ambiguity. With 90% of available referents co-present during learning, all models learned all 35 words; at 95% ambiguity they averaged 27.4, and may not have stabilized yet. This suggests that the simple statistics of co-occurrence (between visual and auditory events) can be extremely powerful. Even when most of the lexicon is available on any given naming situation, the model eventually determines the correct mappings between words and referents. This undercuts claims that associative learning cannot cope with the high degree of referential ambiguity faced by real children. Given sufficient time, such mechanisms may be sufficient when embedded in situation-time competition framework.

Simulation 1.4: Acceleration

In the prior simulations, the model appears to start with high rate of acquisition and taper off (e.g., the 3AFC and 10AFC tasks—Figure 4, 5). This contradicts the consensus that word learning accelerates (P. Bloom, 2000). While acceleration was observed in production (Figure 5) and in the weight analysis (Figures 4,5), in children comprehension also accelerates (Reznick & Goldfield, 1992). More importantly, the McMurray (2007; Mitchell & McMurray, 2009) analysis suggests that acceleration should be observed in most parallel learning systems. So, while it is not distinctive of word learning, it was unexpected that it was not consistently observed.

There may be several reasons for this. First, perhaps we are not sampling at a high enough density. The vocabulary explosion typically appears during the second year of life. Given that it takes the model about 100,000 epochs to learn its adult lexicon, the explosion would probably occur in the first 10,000 epochs, while the above simulations only sampled twice during this time-window. Second, McMurray (2007) predicts acceleration only when words vary in difficulty such that there are fewer easy words than moderate or difficult words. As a simplifying factor in our simulations, all words were equally difficult—they were equally frequent and the degree of ambiguity was the same across words.

Third, when measuring children's vocabulary size we do not subject them to a 3AFC task for each word. Rather, we use a parent questionnaire like the MCDI, (Dale & Fenson, 1996; Fenson, et al., 1994) which probably measure something like children's ability to use a word in a variety of contexts, that is, more of an average performance across time for that word. In fact, ongoing work (Mitchell & McMurray, 2008) using a stochastic version of the (McMurray, 2007) model has shown that, acceleration is only observable when words require several exposures to learn. When words can be learned in only one exposure, *deceleration* is guaranteed. Given the heavily constrained 3AFC task, it may only take a small number of repetitions to learn a word by this criterion (particularly at low-levels of ambiguity) and we may not see acceleration.

Thus, we ran 10 repetitions of the model with three changes. First, we sampled every 200 epochs. Second, for each word, in addition to the usual 3AFC and 10AFC tests, the model was tested five times and had to get the right answer on at least four (to simulate understanding a word in a variety of contexts). Third, we increased the referential ambiguity to 75% (so all words took longer to learn). We also explored the difficulty distribution by manipulating the frequency of the words such that there were few easy words, and many harder ones. This version of the model also showed acceleration. It is discussed in the online supplement (Simulation S1)

Results—Figure 7 displays the results for the 3AFC and 10AFC tasks when all of the words were equal frequency. The thin lines show when the word was considered "known" if the model was correct in a single 3AFC or 10AFC task. The thick lines require the model to be correct on 4/5 trials of these tasks. Requiring a word to pass multiple tests clearly slowed initial performance, and led to a steeper learning curve.

There is also more noise in these measurements (from epoch-to-epoch) than was seen earlier. At this stage in development the model has not really "learned" the words, and its performance is affected by fluctuations in the weights of competitors (that are still settling down) and the choice of competitors in the comprehension task (see S. R. Robinson, Adolph, Young, & Gill-Alvarez, 2008 for a discussion of sampling issues in development). This suggests that even using standardized measures like the MCDI the implicit "task" (the real-time behavior leading up to the measure) may shape the outcomes we measure (see also Sandhofer, Smith, & Luo, 2000).

However, no matter how we measure it, the model undergoes a period of slow learning followed by a period of fast growth (note that the apparent threshold in the 10AFC task is only a momentary plateau—by the end of training, all models learned all 35 words). Thus, the critical factor for acceleration is difficulty—as long as words are fairly difficult to learn, the model shows acceleration. Measurement clearly plays a role, because it acts as sort of a threshold: the number of repetitions required to pass a 3AFC task is less than the number required for a 10AFC task. This is in accord with (Mitchell & McMurray, 2008) model of the vocabulary explosion.

Section 1 Discussion

The first simulations suggest that this dynamic associative model can acquire a lexicon under substantial uncertainty. Even with 90% referential ambiguity, the model acquired all of the words. However, the model's knowledge is a function of how it is tested. Comprehension tasks with small numbers of competitors show earlier learning than those with larger numbers or production tasks. As a result, we see differences between comprehension and production in a system that doesn't do either. Similarly the growth curve is a function of testing – some tests can show decelerating learning, while more realistic assessments show the predicted acceleration.

Virtually all the tasks suggest the model knows more words than are reflected in its connections (knowledge). This offers a paradigm for thinking about the integration of short- and long-timescale processes. Long-term knowledge (the weights) develops slowly and may be incomplete for substantial portions of development. However, fast competition processes allow the model (or child) to augment these weak representations in the moment and act more intelligently. This throws a novel spin on performance/competence. Typically, children are thought to have better underlying competence than their performance. Indeed, even in connectionist networks that do not make a strong distinction between performance and competence we still see situations where the networks' apparent "knowledge" can be overshadowed by in-the-moment task demands (Munakata, 1998; Munakata, et al., 1997). Yet here, we see the opposite: situation-time performance compensates for rather lousy competence. This is not a contrived product of the model. In most real situations the environment (including the caregiver) may provide information or support that allow the child to perform significantly above their level of competence in many situations (c.f., Peterson & McCabe, 1991; Reese & Fivush, 1993). Thus, performance constraints, while usually thought of as impeding our ability to see the true developmental level of the child, may actually augment it in some circumstances.

These simulations also suggest that referential ambiguity may not be as problematic as typically implied. Even in vastly ambiguous environments, associative learning can be successful over the long term. However, learning must be slow, because the relevant information can only be extracted across naming events—in fact, fast learning may cause the model to over-commit to an incorrect interpretation (as we demonstrate in this model in the Online Supplement, S4). Here again, however, the interaction of timescales helps: if fast, task-constrained processes buttress poor knowledge, the model can perform well despite imperfect knowledge in its weights.

Section 2: Situation time processes

This section examines children's ability to *use* their lexica in situation-time. The existing literature tends distinguish familiar and novel word processes. Work on novel words focuses on accuracy: can children infer, in the moment, the referent of a novel word? For familiar words, the emphasis is on the efficiency or speed of referent selection. Thus, Simulation 2.1 examines developmental changes in the efficiency of children's familiar word recognition and Simulations 2.2 and 2.3 examine referent selection by mutual exclusivity, a form of referent selection for novel words. Both require children to use available information, in the moment, to identify the referent of a word. Beyond offering a model of these phenomena individually, these simulations also argue that mutual exclusivity emerges from the same processes as changes in familiar word recognition, and they help reveal fundamental properties of learning that make both possible.

Simulation 2.1: The development of word recognition and the power law of learning

Fernald and colleagues (Fernald, et al., 2006; Fernald, et al., 1998; Swingley & Aslin, 2000) have examined the timecourse of children's mapping of familiar words to their referents using fixations. In this paradigm, children see pictures of two objects (e.g., a *ball* and a *car*), and are instructed to look at one. The speed at which they fixate the right object is taken as a measure of processing speed, and this tends to decrease over development (Fernald, et al., 1998; Hurtado, Marchman, & Fernald, 2005). Fixation time is also correlated with lexicon size (Fernald, et al., 2006; Zangl, Klarman, Thal, Fernald, & Bates, 2005), and is stable and predictive among children (Fernald, et al., 2006; Marchman & Fernald, 2008). The goal of this simulation was to investigate this computationally, in order to identify the potential loci of these effects.

Ten models were run at three levels of referential ambiguity (25%, 50%, 75%, Table 2 for parameters). Every 250 epochs, the model was tested on its entire lexicon in a 3AFC task to assess the number of words known and the time it took the model to settle on a referent for each of them. Similar to Fernald et al (2006), RT was only saved for trials in which the model selected the correct referent.

Results: Development of RT—Figure 8 shows the settling time in cycles as a function of training for each of the three levels of ambiguity. There is a dramatic drop early in training, from 20-30 cycles at 250 epochs, to 5-6 cycles by the end of training. There are small effects of referential ambiguity. These likely derive from the fact that different levels of referential ambiguity offer the model different amounts of exposure to visual competitors, which may alter how competing associations for a given word can be pruned (as we will discuss, this is a crucial determinant of RT). Optimal learning, then, may require a mix of low-competitor situations (e.g., ostensive naming) to establish the words and rule out strong competitors, and high-competitor situations to improve processing (c.f., Horst, Scott, & Pollard, 2010; McMurray, et al., in press).

So what causes the decrease in settling time?—Changes in the dynamics of activation flow cannot account for the decrease in RT because the parameters that control it (temperature and the degree of lateral inhibition) did not change over learning. The only things that did change were the weights. Figure 9 shows a representation of the weights connecting the visual and lexical layers over development. Along the x-axis are the 35 visual units. Along the y-axis are 35 of the 500 lexical units. The strength of the connection between each unit is represented by the darkness of the patch at their intersection. At the beginning of training (Figure 9A) these connections are random and there is no clear structure, but even 10 epochs in, we can see that the overall strength of the connections has been reduced (although there are a few strong connections for words that have been heard). Over the next 1000 epochs, unnecessary connections are further pruned and positive ones enhanced. By around 10,000 epochs (Panel D), the model only has a handful of strong connections. However, some of these are ambiguous. For example, lexical unit 11 (circled) has strong connections to at least 3 input units (and those units are also connected to lexical unit 17). At 100,000 epochs these competing representations have been eliminated—each visual unit connects to only one lexical unit and the bulk of the connections are near zero[6].

We can roughly characterize these changes in terms of the entropy of the weight matrix: a "peaky" distribution of weights characterized by a few strong connections and many weak ones (low entropy) describes a more developed model, while a flatter distribution (high entropy) characterizes a less developed model. We tested this by running an additional ten models with a referential ambiguity level of 50%. Every 500 epochs, models were tested in a 3AFC and 10AFC task and we evaluated both the auditory and visual weight matrices using three measures. First, we computed entropy, converting weight values to probabilities. By this measure, relatively random weight matrices will have large entropies, while peakier weights will have small entropies. Second, at each point in training, we found the single strongest connection linking each input unit to the lexical layer and recorded its strength as a measure of how strong the positive connections were. Third, we took the average of all of the other connections as an indicator of how weak the irrelevant connections were.

Entropy showed strong relationships with both learning and reaction time. We found a negative correlation between entropy and time during learning (R=−.92), and between entropy and the number of words known (10AFC task: R=−.77; weight analysis: R=−.92). Thus, entropy captures some aspect of overall learning or development. Figure 10A relates entropy to the log of the RT in the 3AFC task. Though the relationship is non-linear, high entropies (relatively random weights), predict slower reaction times (Linear: R=.43; Hyperbolic: R=.92).

The strength of the positive connections was not highly predictive of performance on either of our tasks (3AFC: R=−.27; 10AFC: R=−.41) or settling time (R=.31), and analyses of the scatter plots suggested this was not due to a non-linear relationship (Figure 10B). It was moderately related to number of words known (R=−.60), but negatively. That is, lower connection strengths tended to indicate more words known. All of these results derive from the fact that the positive connections fluctuated over training but ultimately decreased slightly (the negative ones decreased a lot more; see Figure 10C). The average of the negative connections was much more predictive. It was highly negatively correlated with performance on the lexical identification tasks (3AFC: R=−.94; 10AFC: R=−.89), and with the log of settling time (R=−.92, see Figure 10D). In both cases, performance increased with smaller irrelevant weights.

---

[6]Note that over development, even the positively associated weights tend to drift downward using this learning rule (they do eventually stabilize—see Figure 10B). This accounts for the fact that the final weight matrix (Figure 9F) shows lower weights than some of the earlier panels.

As a whole, then, the pruning of unnecessary connections is the driving force behind both acquiring new words *and* recognizing them faster. Unnecessary connections cause an auditory input to activate multiple lexical units. These lexical units compete, moving in the direction of winner-take-all. Since the network does not settle until this competition is resolved, it is the presence of these momentarily active competitors (driven by unnecessary connections in the weights) that ultimately leads to a longer settling time. This suggests that empirical correlations between reaction time and vocabulary size (Fernald, et al., 2006; Zangl, et al., 2005) may be driven by the fact that children who know more words may also have fewer spurious associations at that point in development. More sophisticated eye-movement paradigms may be able to test this by evaluating more precisely degree of competitor activation.

Acceleration and Deceleration—Developmentally these simulations suggest a steep decrease in settling times early, followed by a flattening. This pattern, is commonly seen in power-law or exponential decay function in the literature on general learning principles and has appeared in a variety of motor and cognitive learning tasks (Anderson, 1982; Heathcote, et al., 2000; Logan, 1992; Newell & Rosenbloom, 1981; Wifall, et al., submitted), suggesting that word learning may operate by similarly general principles. The power-law has always been interpreted as demonstrating that learning slows throughout training, which would seem to violate the acceleration commonly observed in word learning. However, our model does both.

To understand this we looked for a relationship between the greatest change in RT and the change in number of words known. Figure 11A shows the change in reaction time and words-known for the models learning under 25% referential ambiguity. Here, both RT and words- known undergo their greatest change at around 5000 epochs, implying some fundamental process that affects both at this time. However, Figure 11B shows the same data for the models learning under 50% referential ambiguity. Here, the time-course of lexical acquisition is pushed much later (peaking between 60,000 and 70,000) but the pattern of RT barely moves. Thus, the timing of changes in RT does not clearly map to changes in the number of words known.

What can explain apparent acceleration and deceleration in learning? As we discussed in the prior section, the best predictor of RT is the magnitude of the spurious connections between auditory and visual *competitors* and the current lexical unit. These constitute the bulk of the weight matrix. If there are $35 \times 500 = 17,500$ weights connecting auditory and lexical nodes, only 35 of them are correct—17465 must be suppressed over learning. However, weight decay is also widespread. Every time a word is heard, thousands of weights are not used and therefore suppressed. Thus, there are quick RT gains to be made for suppressing unnecessary competitor weights, and this can be done in virtually all circumstances. Any word, with any degree of referential ambiguity will result in some weights being decayed.

The number of words known is more complex. It requires two positive weights (auditory→ lexical and visual→lexical) and an absence of competing weights. Neither the irrelevant weights nor the positive weights are singly related to it. Even if the network had a very strong connection between an auditory word-form and its lexical unit, if there were other strong competing connections, the word would still not be learned. Thus, in contrast to reducing RT, actually learning a word requires a confluence of events, and is much more difficult. This is clear in the correlations observed in the previous section: words-known was best correlated with entropy, a global measure of the weights (R=−.91) and less so with either spurious (R=−.56) or relevant (R=−.60) weights; while RT was best correlated with spurious weights (R=.91) and less so with the other measures (relevant weights: R=.21; entropy: R=.42). Thus, despite the fact that the development of both RT and vocabulary size

rely on fundamentally the same learning mechanism, the fact that each must be measured through a situation-time measure means that they may tap different aspects of learning, and hence show different learning curves.

Discussion—This simulation shows similar results to Fernald et al's studies: the dynamic associative model's familiar word recognition improves over time. This improvement, which appears as gains in the efficiency of a skill, actually derives from changes in knowledge (connection weights). Crucially, the suppression or pruning of the irrelevant weights is the dominant factor—the bulk of word learning may consist of learning which objects and words *do not* go together (c.f., Regier, 1996). We also show that word learning does not have to differ from general approaches to learning which predict decelerating reaction times, even as it show acceleration in vocabulary size. This underscores the fact that word learning may operate by general learning principles and challenges the utility of drawing strong conclusions based on the shape of vocabulary growth (McMurray, 2007). Critically, each of these measures derives from different changes in the underlying associations, showing the multi-faceted nature of association learning in this context. At a broader level, however, the acquisition of word/object linkages, a clearly developmental-time learning phenomenon, is directly implicated in the ongoing development of processing speed, a situation-time measure. In fact, changes in processing derive entirely from a non-obvious component of learning—suppression of irrelevant connections.

Simulation 2.2: Referent Selection by mutual exclusivity

If the development of familiar word recognition derives from the same learning mechanisms as the acquisition of new words, this raises the question of whether the processes that infer the referents of novel words can also arise from these mechanisms. The plethora of proposed constraints and specialized inference processes imply that children deploy additional mechanisms in novel naming situations. Yet, how does the child first determine that the word is novel and then which constraints to apply? While some have argued for a decision process of some kind (Merriman, et al., 2008), an alternative and perhaps more parsimonious account is that these biases emerge out of the same dynamics that give rise to familiar word recognition.

In our dynamic associative model, both novel and familiar words undergo the same competition and associative learning. Given that early in training all words are novel, the fact that this model is able to learn words at all suggests that it is not *a priori* necessary to separate novel and familiar word processing to solve the referential ambiguity problem. However, it is not clear whether the model also shows systematic biases in how it interprets novel words.

Simulation 2.2 evaluates mutual exclusivity. Mutual exclusivity is the idea that a novel word cannot refer to a referent that has a previously established word-referent link; that is, words are mutually exclusive. In terms of behavior in referent selection tasks, however, children are said to be following the mutual exclusivity constraint when they exclude objects with known names as the referent of novel words (Mervis & Bertrand, 1994). Note that there are debates over whether the inference children make in such cases is best described as mutual exclusivity (Markman & Wachtel, 1988), a process of matching novel words to novel objects (Mervis & Bertrand, 1994), or another form of inference (Halberda, 2006). Likewise there are debates over whether this constitutes learning (Horst & Samuelson, 2008; Spiegel & Halberda, 2011). Here, we are only using mutual exclusivity as a moniker for a behavioral phenomena that happens in the moment when children are confronted with a novel name, a novel object and several familiar objects. We are not implying any specific inference principle (in our view it arises from competition dynamics), nor are we implying learning—

in terms of either an initial, quick link (as the term "fast-mapping", often applied to such situations, does) or a longer, more robust connection. Rather, we will simply refer to the phenomenon of selecting the object that does not have a name when confronted with novel and known objects and a novel name as *referent selection by mutual exclusivity* or *M.E. reference selection* for short.

Crucially, M.E. reference selection minimally requires a range of available objects, a novel word and some partially learned weights. All of these are present in this model. Thus, while many of the proposed biases and constraints "live outside" of the simple architecture of this model, mutual exclusivity is clearly within the domain of our model, raising the possibility that it could arise in the context of the competition dynamics.

To examine this, 20 networks were initialized with 40 auditory and visual units and 500 output units (Table 2). Of the 40 input units, 30 were used during training; the remaining 10 novel units were never heard nor seen. Thus by the end of the 100,000 training trials, the model had a lexicon of 30 words, but an additional 10 novel words, whose weights were largely unchanged. After training at 50% referential ambiguity, the models were tested in three ways. For ease of description, we describe these as three-letter strings, with the first letter representing the status of the target. First, we used a 3AFC task with all familiar words (F_tFF) for comparison with previous models. Second, we tested *M.E. referent selection* trials, using a novel target, with two familiar objects (N_tFF). Finally, we used the same configuration of competitors, but with a familiar target (F_tFN). Here, if the model always selected the novel object we should see a performance decrement. To construct five novel and familiar trials, the network needed to know at least 10 words, so we tested all 30 words in a production task prior to constructing these test trials, and models without a 10-word vocabulary were not tested.

Results—The models ultimately acquired 28.7 of the 30 words (by the production task). Figure 12A shows the models performance on the three primary tests. Models were 99.8% correct on the 3AFC task, and 100% correct on familiar word trials and M.E. referent selection trials. Thus, fully trained models were effective at identifying familiar words (both with and without the novel object) and at matching the novel word with the unnamed visual object.

Given that the model has no explicit mechanism for mutual exclusivity, how did this emerge? The answer lies in the interaction of the competition and the learning rules. Figure 12B shows the visual-to-lexical weight matrix for a single run of the model after 100,000 training epochs. The x-axis shows the index of each of the 40 visual units (units 1-30 are familiar words and 31-40 are novel words). The y-axis shows the index of the lexical units. The model had 500 lexical units, but we only show lexical units that were strongly connected to one of the familiar words along with a random sample of 20 output units for the novel words[7].

For familiar words each object is strongly connected to a single lexical unit. This can be seen in Region 1, which resembles the weight matrices displayed earlier for a trained network. Similarly, Region 2 shows that the connections between these familiar objects and the unused lexical units were eliminated so it would be difficult to assign a novel word to one of these objects. Region 3 shows that *lexical* units that have been assigned to a visual input are also not connected to any other visual input, including the novel objects. Most importantly, however, Region 4 shows that the novel objects all have small and variable

---

[7]We also changed the order to group the "assigned" lexical units together (a typical run of the model would recruit output units from across the array with no relationship to the order of the inputs).

connections to the remaining, unused lexical units. As a result, it is highly likely that there will be an associative path that allows activation to spread from a novel word to a novel object, even though these associations are random. Conversely, it is quite unlikely that a novel word could activate known objects.

This particular structure in the weights comes from the weight decay in the learning rule.

$$\Delta w_{xy} = \begin{aligned} & v_x l_y \left(1 - w_{xy}\right) \\ & -.5 \cdot (1 - v_x) \cdot l_y \cdot w_{xy} \\ & -.5 \cdot v_x \left(1 - l_y\right) \cdot w_{xy} \end{aligned} \quad (9)$$

Here, weights decay in two circumstances. First, they decay whenever a visual unit is on and the corresponding lexical unit is not. Thus, the connections in Region 2 are pruned because the familiar objects have developed connections with a single lexical unit, and lexical inhibition prevents a second one from becoming active. Second, they decay if a lexical unit is active, but the visual input is not. Thus, the weights in Region 3 decay because the network has not encountered these objects but has used the corresponding lexical units for other words. The weights connecting novel inputs to unused lexical units *never decay* because neither class of units is ever active. As a result, connections in Region 4 maintain their original (small, random) values. Then, when a novel word is encountered, these connections permit the network to activate (to various degrees) a large number of lexical units. This activation feeds back to the visual units, but since the familiar objects are not connected to any of these now-active lexical units, only the activation of the novel visual objects is amplified. This allows the network to select the correct (novel) object.

Thus, the selection of the novel object is dependent on the learning rule, but not because the network needs to learn something about that object/word. Rather, the weights between the known word/objects and the unused lexical units must decay, and the weights between the novel ones must not in order to create a platform upon which real-time competition dynamics can select the right object. A different type of weight decay (for example, if all weights decayed on each epoch) would not preserve the right form of the weight matrix. However, learning is not the whole story: this pattern of connectivity could not be harnessed in situation time without the gradual settling process represented by the inhibition and feedback dynamics. Moreover, the model's ability to learn from M.E. referent selection may also depend on this competition/feedback cycle. The model must select a single lexical unit and selectively amplify the novel object in order to eventually turn a word-referent link created during M.E. referent selection into a known word by associating the novel object with the novel word over many instances. Thus, while as a real-time process mutual exclusivity is likely to impact learning, it is really more the product of learning than a mechanism of it. This implies that some types of learning environments may make it more difficult for children to engage in this by eliminating this particular structure of associative weights. This will be examined in Simulation 4.2.

Simulation 2.3: The development of M.E. referent selection and visual familiarity

In some ways, the previous model performs mutual exclusivity too well—children rarely approach 100%. Yet, this was a fully trained adult model, so it was important to examine the model developmentally. There have been few comprehensive developmental investigations of mutual exclusivity. We do know that in a 2AFC task children can succeed at about 18 months, depending on vocabulary size (Markman, et al., 2003), but fail at 5AFC novel-word tasks until after the vocabulary spurt (Mervis & Bertrand, 1994). Further, Halberda shows a clear developmental timecourse with 14 and 16 months old failing in a 2AFC looking version of the task, but 17 month olds succeeding. Thus mutual exclusivity is not an innate

constraint, but develops over time. Given the prior simulation demonstrating the dependence of mutual exclusivity on learning, we investigated this by re-running the 20 models described in the previous section but measuring performance every 5000 epochs.

Results—Figure 13A shows the results. The lines marked by open circles show the model's performance on the 3AFC (F_tFF) and production tasks. These were run regardless of how many words the model knew and show a steady improvement over learning. The models knew enough words to be tested on M.E. referent selection by around 45,000 epochs, and at this point performed at nearly 100%: by the time model knew enough words to be tested on mutual exclusivity, that ability was present. Indeed, runs of this model with fewer novel word trials (hence requiring fewer known words), show even earlier abilities, suggesting that this model may be able to do this task with very little experience.

This was unexpected. Apparently it did not take much learning to set up the right structure in the weight matrix (and given the dramatic changes in irrelevant connections seen in the first thousand trials in Figure 9, this may be sufficient). One factor that may moderate this is visual familiarity. Our analysis of the weight matrix suggests that good M.E. referent selection derives from the fact that the novel visual units have never been active to any degree. Yet, most experiments do not use stimuli that are completely unfamiliar. Typical novel objects such as whisks and juicers, while unlikely to be named, have likely been seen before, or may be similar to things that children know. Thus we ran an additional set of simulations in which the novel visual units were seen (but never named) on some proportion of the trials. The likelihood of seeing a novel object varied from 5% to 50% (since the referential ambiguity rate was 50%, so this last condition was equivalent to the unnamed objects being as familiar as the known objects).

The effect of familiarity—Results are shown in Figure 13B-F. Panel B shows the lowest level of familiarity—novel objects appeared 5% of the time; Panel F shows the highest, in which novel objects were as likely to appear as known objects (though never named). The familiarity of the novel objects does not seem to influence responding on the familiar word trials (F_tFN)—performance was equally good in all simulations. Importantly, however, even a small amount of visual familiarity impedes M.E. referent selection at early points in development. Panel B shows that a marginal amount of familiarity brings initial performance down to 55%, and any more can bring it down to chance. Very quickly after that, performance seems to develop to full capacity. However, when the novel foils are highly familiar, M.E. referent selection is never very good. Thus, at least for this model, mutual exclusivity is not solely about lacking a name (as accounts like N3C suggest). The familiarity of the object may play a role as well.

Thus, there may be effects of visual novelty in these tasks that can be seen developmentally – the use of more novel objects could lead to better referent selection by mutual exclusivity. Indeed, Horst, Samuelson, Kucker and McMurray (2011) showed that in a completely unconstrained referent selection task (three novel objects) children were heavily biased toward objects that were not seen in a brief familiarization. This raises the possibility that at the early stages, mutual-exclusivity is really more of a novelty preference, rather than a complex inference. Moreover, beyond visual novelty there may be other ways to slow down the development of M.E. referent selection. For example, we've found in ongoing work with this model that including some quantity of low-ambiguity (ostensive naming) trials along with higher ambiguity trials can give rise to a similar effect (McMurray, et al., in press).

Section 2 Discussion

These simulations capture a number of classic findings in word learning, including the improvement in efficiency of familiar word recognition, and referent selection by mutual exclusivity. They show that changes in familiar word recognition, while manifested in online processing, are better characterized by the shape of the learned connections. Moreover, this same learning can give rise to both a deceleration in RT and acceleration in the number of words known. Word learning can be a general learning process. Similarly, referent selection by mutual exclusivity, while it may appear a specialized inference process, can arise out of the same competition dynamics as familiar word recognition, when this process plays out over connections established during learning. This allows us to capture how exposure to objects alters referent selection by mutual exclusivity. More importantly, embedding this within a system that recognizes familiar words and learns word/object linkages allows for a richer explanation.

In both of these simulations, these ostensibly situation-time processes are the product of learning—but not any simple version of learning. With respect to familiar words, the most important predictor of processing speed is how the *unnecessary* connections decay. Similarly, mutual exclusivity fundamentally relies on a learning rule that describes a particular pattern of weight decay. Thus, suppressing competing associations is essential to multiple aspects of word learning. Similarly to Fernald et al's account of changes in processing speed, Mervis and Bertrand (1994) suggest that the number of words is the critical factor that predicts the onset of M.E. referent selection. However, this doesn't offer a clear mechanism of change in this context because the competitors are always familiar in mutual exclusivity task. Our model suggests that development has more to do with the pruning of weights (which is likely correlated with the number of words known, and was in the model). Across both simulations, however, the more important message is that to take advantage of the explanatory power inherent in this version of associative learning, we must consider *both* the positive and negative associations.

Section 3: The relationship of situation- to developmental-time processes

The previous simulations demonstrate that apparently situation-time processes are the product of learning. This section addresses the converse: How do the details of processing impact learning? Simulation 3.1 models data suggesting independence of timescales: Horst and Samuelson's (2008) work on retention after M.E. referent selection; and Simulation 3.2 examines task effects on mutual exclusivity over development. Simulation 3.3 returns to familiar word recognition, and examines longitudinal work showing that recognition time predicts the future rate of acquiring new words. Finally, Simulation 3.4 asks if learning can occur without processing.

Simulation 3.1: Referent selection by mutual exclusivity and retention

Horst and Samuelson (2008) showed that children do not retain words after referent selection by mutual exclusivity. This suggests that this behavior is a situation-time process and not synonymous with learning. Our dynamic associative account is ideal for capturing such effects: referent selection emerges out of online competition, while learning is slow and may not be able to acquire a word in one exposure. Our model can extend these findings by asking whether anything is retained from referent selection, and what circumstances may be necessary to see it.

We simulated Horst and Samuelson (2008) by initializing 20 models with 40 input units, but only training them on 30 words (Table 4 for parameters). This left 10 novel words/objects that did not receive any training. Five were used to test mutual exclusivity and the other five were held-out. Referential ambiguity for familiar objects (words 1-30) was set to 50%, with

novel objects appearing as competitors 8.75% of the time (but novel names were never heard).

Models were tested on several tasks. First, we assessed which words were known with the production task. Only words that passed this test were used in subsequent testing. Next the model received five novel (N_tFF) and five familiar (F_tFN) word trials similar to Simulations 2.2 and 2.3. However, unlike those simulations, the model learned on these trials, enabling tests of retention for these words. Finally, on retention trials, each of the five novel objects named on the prior M.E. referent selection trials was paired with another novel object and a held-out object (N_tNH). While learning (weight change) occurred throughout the sequence of test trials, the total learning over a single batch of testing was not carried back to training.

Results—Figure 14A shows the model's performance on each type of trial over the course of training. As before, familiar and novel word performance was excellent after the emergence of mutual exclusivity at around 45,000 epochs. However, the model was not able to retain the words that were tested in the M.E. referent selection trials, averaging 38% correct retention.

Based on the prior simulations, we were concerned that the visual familiarity of the novel objects may have created this effect. Thus, we replicated these simulations under two conditions: one in which the novel and held-out objects were never seen during training; and one in which they occurred frequently (p=.375). When the novel objects were completely unfamiliar, results were similar (Figure 14B): referent selection performance was at ceiling (as in 3.1), but retention was at chance. However, when the novel objects were highly familiar (but unnamed), there was a period during which the model did well in referent selection, but not retain, followed by later points in development when could do both (Figure 14C). This fits with recent work by Kucker and Samuelson (2012) showing that 24-month-old children can retain links created in an M.E. referent selection context if they play with the objects prior to the mutual exclusivity trials. It is also relevant to Spiegel and Halberda's (2011) recent finding that older children (30 month-olds) appear to retain word/object mappings, though in an easier looking task.

Overall, then, the model fits the pattern of Horst and Samuelson (2008) showing excellent performance in referent selection by mutual exclusivity, but a failure to retain when contextual support is removed. This raises the question of how much, if anything, the model learned from a single mutual exclusivity trial. While gross performance did not yield evidence of learning, there may be a small amount of learning that is insufficient to drive overt retention.

To assess this, we examined the amount of change in subsets of each weight matrix at various points in training using the root mean squared (RMS) difference between the weights at two points in time (e.g., before and after the mutual exclusivity trials). Weights were divided up (Figure 14D) into weights connecting lexical units to 1) familiar words; 2) novel words and to 3) held-out words. We then computed the weight change (learning) in each group that occurred during learning and during mutual exclusivity trials. Including the held-out units allow us to determine how much change to expect for completely unused items.

Figure 14E and F shows the results. Panel E shows the amount of weight change up to the point where the model was tested at 100,000 epochs. There is quite a bit more change in the weights for familiar words (which are being trained) than the novel or held-out words, particularly in the auditory weights. This makes sense—the novel and held-out auditory units are never activated, while the novel visual ones occasionally appeared as competitors.

In contrast, Panel F shows the amount of weight change during the mutual exclusivity trials. There was some learning on these trials and generally more learning for the novel words than the others. However, the amount of learning on these trials is far less than what was learned about those words over the course of training… when they were never heard! It is also far less than what a truly familiar word would have received. Moreover, this small amount of learning is not responsible for the excellent performance in referent selection by mutual exclusivity—when learning is turned off during these trials (Simulation 2.2), the model still performed at 100%.

Nonetheless, this offers a clue to how M.E. referent selection relates to learning. The model learns a little something from each these trials, and over the course of many such trials, this accumulates to yield complete word learning (see also, Horst, et al., 2006). But crucially, that tiny amount of learning we observed on that first exposure to a novel word is not different from what would be observed on the 2nd, 3rd or 4th exposures. Moreover, this learning consists not only of building or maintaining correct associations, but also (and to a much larger extent) suppressing unnecessary ones. Thus, what happens during this first referent selection is quite different from what earlier views (Carey, 1978) may suggest. Thus, referent selection by mutual exclusivity, while a primarily in the moment process, leaves a small trace in the weights which can accumulate to achieve real knowledge.

In retrospect, the training used in all of the simulations thus far likely included many mutual exclusivity trials. Since competitors were randomly chosen on each epoch, there were likely many epochs in which the model "knew" all of the words except the target (or knew more about the competitors than the target). In this way, there is nothing fundamentally different about familiar and novel words.

Simulation 3.2: Effect of task—Section 1 suggests that task has a critical effect on the model's performance (e.g., the delay in productive vs. receptive vocabulary). Similarly, mutual exclusivity also has the characteristics of a task effect: the two familiar words constrain the task, permitting the model to perform well despite no knowledge of the novel word. This predicts that task variables like the number of alternatives may affect mutual exclusivity, particularly early in development.

Mervis and Bertrand (1994) measured referent selection via mutual exclusivity as a function of vocabulary size. They found that only children with relatively large vocabularies (greater than 90 words) consistently performed well in the mutual exclusivity task (despite the fact that the familiar objects were known to each child). Moreover, for the children who failed at this task, once their vocabularies reached this level, they too could select the correct referent. Thus, some critical quantity of words may be required before this ability emerges. However, Markman, et al., (2003) pointed out that the four and five-alternative mutual exclusivity tasks used by Mervis and Bertrand, may have simply been too difficult. They used a two-alternative task (one familiar and one novel object), and demonstrated that, regardless of vocabulary level, all children could succeed at the mutual exclusivity task. Horst et al., (2010) further showed that by 30 months, the number of competitors made little difference in children's performance (though it did affect retention). Thus, task differences may disappear later in development.

To examine this, 20 models learned 30 words at 50% referential ambiguity. As before, 10 novel words were held out for testing. Novel visual units appeared as competitors with a likelihood of 8.75%, to obtain non-ceiling performance on mutual exclusivity trials. Every 2000 epochs, the model was tested in both 3AFC and 5AFC familiar and novel word tasks. For both familiar and novel-word tasks, the competitors consisted of two or four familiar objects and one novel object (F_tFN, F_tFFFN, N_tFF or N_tFFFF). As before, words were

screened using the production task, and models were not tested unless sufficient words were known.

Results—Figure 15A shows that, as expected, models performed equivalently and at ceiling on both 3AFC and 5AFC familiar word tasks. In the M.E. referent selection tasks, models performed better in the 3AFC task than the 5AFC task early in development, also as predicted. Panel B shows results from a second set of simulations in which the novel objects appeared 25% of the time. It suggests that this task difference was enhanced when the objects were more familiar. In both cases, however, the effect of task was eliminated quickly after M.E. referent selection got off the ground. Thus, as in children, task differences in referent selection by mutual exclusivity are only observed at a narrow window in development.

Simulation 3.3: Processing speed and rate of acquisition

We've described referent selection by M.E. as primarily a situation-time inference process, but one that is fundamentally based to prior changes in the weights due to learning. We now ask whether such interactions between timescales are also seen in familiar word recognition.

A study by Fernald, Perfors and Marchman (2006) offers an intriguing platform for this. They examined 63 infants longitudinally between 12 and 25 months and assessed both the number of words known (using the MCDI), and the infants' speed of processing familiar words using the looking-while-listening task. They found first that speed of processing in this lexical task (reaction time) is stable across individuals (though decreasing) from month to month. Second, there was a correlation between speed of processing and accuracy in preferential looking. Third, RT was negatively correlated with the number of words known, particularly for the older children (25mos). These first findings could be accounted for by simply assuming that processing speed is a function of learning, much as we showed in Simulation 2.1; learning influences processing. However, Fernald et al. also found that the children with significantly shorter RTs at 25 months showed more acceleration in the number of words acquired. This suggests the converse – that processing influences learning. Fernald et al., conclude that processing speed may be a property of the child that is fundamentally related to word learning. This motivates a compelling link between online processing and learning. They argue that children who can identify words faster have more resources (or time) to process subsequent words, allowing further opportunities for learning.

This simulation asks if our model shows these same dependencies. This can validate the model by simulating a complex set of phenomena, and extend it to individual differences. It also affords the opportunity to understand the mechanisms that give rise to this particular relationship between learning and processing. In particular, our model does not have the sequential processing demands of real children so if it still shows such dependencies it may help illuminate alternative accounts. Moreover, as processing speed can derive from both differences in activation flow and from differences in learning, we can start to understand the range of causes that can give rise to this crucial descriptor.

To simulate this, we needed individual differences across models. While some variation is created by the random initial weights and the random sequence of training, this was not sufficient. Thus, across simulations we also varied the free parameters of the model as an additional source of variation (c.f., McMurray, Samelson, Lee, & Tomblin, 2010). We initialized 100 networks and varied parameters like the temperatures, the inhibition, and the learning rate by adding Gaussian noise to the means used in prior simulations (Table 3).

Models were trained on a lexicon of 35 words. They were tested every 1000 epochs on all 35 words in a 2AFC task and we recorded both the number of words known (by this

measure) and the average settling time. Similar to Fernald et al. (2006), settling time was only recorded for trials on which the network answered correctly. Models were also tested on a 20AFC comprehension task. Our analysis follows the findings of Fernald et al, (2006). We assess the stability (within a time-slice) of our measures of RT and accuracy, and then the correlation of these measures with the rate of long-term learning.

Results: Stability of reaction time and accuracy measures—Table 5 shows the pair-wise correlations in RT and accuracy in the 2AFC task for adjacent tests. Both accuracy and RT were highly correlated across time with an average correlation of .9 (RT) and .57 (Accuracy). These are higher than the correlations found by Fernald et al., (2006) who report correlations in the range of .2 to .4 for RT and .25 to .5 for accuracy, but not unexpected for two reasons. First, children have a range of processes outside of the model that may introduce variability. Second, our test included all 35 words, while Fernald et al., only tested 4-8 words, a much smaller sample of a much larger lexicon. As a result, the models' estimates of RT and accuracy were likely to be closer to the true values than behavioral work can derive for children.

Despite these correlations, however, the model is not perfectly stable. Figure 16A shows 2AFC performance over training for six representative runs of the model. Model number 09, for example, starts with one of the worst performances, but is quite successful by the end. Model 07 is the worst, but by about 10,000 training trials (Log 4), it tracks quite closely with 11, the best. There is also considerable variation in when the models start to perform quite well. Panel B shows a similar pattern for RT. Models 07 and 10, for example, start with the worst reaction time, but end up with the lowest; while models 11 and 13 start low and stay low. Model 09 starts similar to 11 and 13, but ends high. Thus, despite these remarkably high point-wise correlations, looking at the whole timecourse of development (and sampling at a much higher rate), we see that the underlying instability of the developmental timecourse is not well captured by the correlations. This supports the kinds of micro-analyses advocated by Robinson et al., (2008).

Results: Relationship of speed to accuracy—Fernald et al., also found negative correlations between speed and accuracy at any given month: children who settled faster got more words correct (R=−.3 to −.5). This was also observed in the simulations. Table 6 (Accuracy column) shows the correlations between the number of correct 2AFC trials and the log of the settling time. At 5,000 and 10,000 time-steps there was no significant correlation between these two factors. However, this is expected, because the models knew an average of 3.8 words at 5000 and 4.9 at 10,000 (chance on our 20AFC task would predict 1.75 words). By 15,000 epochs, however, the models knew an average of 7.5 words (by our 20AFC assessment) and correlations between RT and accuracy were significant (R=−.30, p<.01) and increased throughout development (to R=−.85 by 100,000 epochs). These correlations were negative: models with lower settling times knew more words.

Results: Relationship of speed to vocabulary growth—In the longitudinal study, RT at 25 months predicted the *acceleration* in word learning across the period studied. To assess this, Fernald et al., fit quadratic functions to the number of words known at each month. They compared each term of this function to the RT at 25 months and found a significant correlation with the quadratic term (but not the linear term or the intercept), suggesting that settling time is related to acceleration in vocabulary growth.

The model showed the same behavior. For each model, we fit a quadratic function to the number of words known (in the 20AFC task) over the first 25,000 epochs. This roughly corresponds to the period of early learning studied by Fernald et al. (2006)—by 25,000 epochs, the models averaged 11 of the 35 words and were through their first period of

acceleration. As Table 6 shows, the quadratic term was correlated with settling time at every time-step, while the linear and intercept terms were not (except at the first time-step). Thus, the model shows the same relationship between speed of processing and vocabulary growth as children.

Fernald et al., (2006) posit that in running speech, children who finish processing a word quickly can move on and learn from subsequent words. This doesn't seem to be the case here—the model is reset between words. While this does not rule out this sort of boot-strapping in children, it does suggest that such relationships can arise from other causes. For example, in our model, the parameters controlling settling dynamics (RT), may also alter the networks' ability to resolve referential ambiguity, which could affect learning. Or conversely, as we've shown settling time is primarily a function of learning, so both effects may derive from the same cause.

Results: What parameters influence outcome measures?—Fernald et al (2006) describe *speed of processing* as a fundamental parameter describing variation among children. At the level of description this is undoubtedly correct, though the underlying mechanisms are not clear. Our model offers a set of candidate parameters, but no single one maps directly to speed of processing. Rather, the speed with which the model processes input is an emergent property that derives from multiple components: parameters like the temperature and the degree of inhibition that directly affect the dynamics of settling are clearly important, but the biggest component may be the weight matrix, a product of learning.

We examined the relationship between our behavioral measures and each of the parameters that were manipulated to create individual differences across development. Correlations are shown in Table 7. A number of factors were significant predictors. The number of visual competitors, for example, was correlated with learning: models with more competitors learned more slowly. The stability point (e.g., how little change was required to end a trial) was moderately correlated with RT (models with high thresholds tended to settle slower), but inversely related to learning (the models with the lowest thresholds knew the most words). Input inhibition played a role in settling time, but not so much in learning until the end of the simulation at which point models that had stronger inhibition showed more words learned. This suggests that the ability to inhibit competing representations in the visual layer is important both for quickly settling on a target and for ultimately learning the correct mapping. Learning rate was not a strong predictor, but later simulations (see Online Supplement, Simulation S3) suggested that the range of learning rates we tested may not have been sufficient.

The most important predictor was feedforward temperature, the rate that activation accumulated in the lexical layer from input layers. It was highly correlated with settling time (at all points in development), and with the number of words known at mid-to-late points in development. Oddly, however, the correlations with RT were the inverse of what we expected: models with higher temperatures settled slower (and learned worse). Follow-up simulations (reported in supplementary note S2) demonstrate that this is due to the fact that models at a high temperature settle slower because they artificially increase activation for the competitors (as well as targets), and thus take longer to suppress them.

Given the complex role of temperature and other parameters contributing to settling time and learning, we suggest that concepts like "speed of processing" do not reflect a unary dimension of the underlying architecture. Rather, they are emergent on a complex interplay of system dynamics, the performance in the tasks, and the developmental history. Crucially, even the parameters controlling dynamics were correlated with number of words learned

(which in turn is a predictor of settling time), and thus many of these effects may be mediated via learning.

Discussion—Our dynamic associative model provides a compelling complement to Fernald et al. (2006). The model shows the same stability of RT across development as children, but suggests that there may significant instability when we look closer. It demonstrates the link between RT and accelerating learning but without any simple causal mechanism (e.g., processing capacity). Rather, the relationship derives from the fact that both processing and learning derive from changes in the weight matrix.

More importantly, the model offers a way to interpret RT. Reaction time may not be isomorphic to some elemental individual difference. Rather, it emerges from the interplay of the properties of both learning and the dynamics of competition. These create a fairly stable measure, but one which affects RT and learning at different points in time, as we saw in Simulation 2.1 with respect to acceleration an deceleration. While processing time is clearly an emergent property of network dynamics and learning, it also reflects individual differences in things like the learning rate, the temperature and the like. And, because these things also affect learning, it suggests a highly circular and mediating relationship between learning and processing. As simulation 2.1 showed, processing time is determined in large part by the nature of the weight matrix (and the irrelevant connections in particular), so it should not be surprising to find such a relationship. Only if processing time is treated independently of lexical knowledge does this seem surprising. This suggests that explanations of individual differences based on speed of processing (e.g., Kail, 1994), while perhaps behaviorally stable, may oversimplify the problem, particularly when learning is involved.

Simulation 3.4: Is online processing required for learning?

The previous simulations suggest interactions between learning and processing, deriving in part from the nature of the weight matrix and how it influences real-time competition. In the present simulation, we take this to the extreme, asking if real-time processing is necessary for learning.

Models of unsupervised learning in other domains suggest that unsupervised learning may require some form of competition (McMurray, Aslin, & Toscano, 2009), and most unsupervised architectures include some form of it (e.g., Kohonen, 1982; Rumelhart & Zipser, 1986). Perhaps, then, competition is required for learning. If the competition/ feedback dynamics allow the model to improve upon ambiguous inputs, it would seem more efficient to use the results of this processing as the basis of association, rather than the more ambiguous inputs to it.

There are three components of competitive real-time processing in this model. First, inhibition between lexical units allows more active units to suppress activation for competitors. Second, feedback between the lexical layer and the inputs helps the network suppress competing inputs (visual competitors) as it makes a decision about the word. Finally, inhibition among input units helps the network suppress competing inputs. All three contribute to activating the correct lexical and visual units in situation-time; but, it is not clear if they are necessary for learning.

Thus, we ran a series of simulations which factorially varied whether feedback, lexical inhibition and input-layer inhibition was used. Each combination was run at three levels of referential ambiguity (20%, 50%, 80%), yielding 24 simulations. This was repeated 10 times for 240 simulations (Table 4). Each model was tested every 5000 epochs. Models without

feedback cannot adjust activation in the visual units, making the NAFC tasks useless. Thus, our primary measure was the analysis of the weight matrices.

Results—Inhibition at the input layer was a fairly small contributor to learning, thus we averaged across models with and without it for most of the analyses. Figure 17A shows the number of words learned over training in each of the four permutations. Lexical inhibition was required for learning. The models without it acquired an average of 1.27 words, while all of the models that used it acquired all 35 words. Associative learning of this type cannot proceed without the ability to suppress competitors at the lexical level.

The effect of feedback (assuming the presence of lexical inhibition) was more nuanced. The models with feedback (the full model) acquired a few words very quickly, followed by a delay before learning the rest (Figure 17B). Models without feedback took longer to get started, but once they did, they quickly outpaced the models with feedback.

While competition is clearly required, is there an advantage for feedback? Possibly. It may help the model to acquire a small working vocabulary quickly (Figure 17B). It may also benefit online processing. Figure 17C shows the settling time in a 3AFC task of models with and without feedback. The processing ability of the model with feedback improves much more rapidly and stably than those without. Moreover, while feedback slows the speed of learning, it may improve the quality. Figure 17D shows the average of the weights connecting any given lexical unit to the "incorrect" visual and auditory units. These should be 0 by the end of learning. As can be seen in the figure, situation-time feedback allows the model to suppress these connections faster over developmental-time. Finally, feedback enables the model to change activation in the visual layer as a lexical unit is selected, the basis of our "behavioral" tests. Yet, this is not just a computational convenience. For children, word learning must be integrated with behaviors like selecting referents (focusing attention). It makes more sense to harness the word-object connections established by word learning to guide these behaviors than to rely on a completely feed-forward system that would have to acquire new mappings to do this.

In Simulation 1.1, we described how online processing allows the network to perform better than its partially learned knowledge (weight matrix). However, the current simulations suggest that online processing is much more important than that. Associative learning in this model is simply not possible without some type of lexical inhibition—the model must make a decision about what word it was hearing. Online processing is not merely shaped by learning, nor does it merely buttress performance. It is essential for, and integrated with, learning.

Section 3 Discussion

This section simulated three empirical findings regarding the interaction of situation- and developmental-time processes. Across all three, we sought to determine 1) if situation-time processes are independent of development, 2) if development, particularly word learning, impacts situation-time processes and/or 3) if situation-time processes impact learning.

Our dynamic associative account captured some of the evidence for (1): the failure of children to retain recently novel words from mutual exclusivity trials. However, it suggests that while learning may be slow (requiring multiple trials) some learning occurs during referent selection by M.E.. It also accounted for (2): the model's performance on 3- and 5AFC mutual exclusivity tasks was a function of its development. However, unlike prior explanations based on number of words known, the pruning of unnecessary connections was the determining factor. Our examination of (3) was more ambiguous. While we modeled the Fernald et al. (2006) longitudinal work, the model suggested that processing and learning are

both emergent from the whole network; however, when we manipulated the component processes of the model in Simulation 3.4, we showed that processing is required for learning.

This last finding suggests the model is not entirely unsupervised, it is self-supervised. The model doesn't just associate the inputs it sees and hears. Rather, it performs something like an inference process (implemented via competition and feedback) and then uses the output of such a process as the basis of association. The power of this model to both learn under high degrees of ambiguity and account for a wealth of data speaks to the power of such a scheme.

Section 4: One-word/One-Object?

Thus far, we have restricted our simulations to "pure" word/object sets in which every object has one label and every label has one referent. However, real language contains many polysemous words (with two meanings). Conversely, most objects can be labeled in multiple ways; most vividly, objects have labels at multiple levels (e.g., basic level, superordinate). In this final section, we begin to probe the limits of such situations in two ways to determine how much of the models performance can be attributed to its idealized "language".

First, in addition to its use in describing real-time referent selection, mutual exclusivity has also been described as a hindrance to learning second labels for objects (e.g., Markman, et al., 2003; Regier, 2005). As Xu and Tenenbaum (2007) point out, this constraint must be relaxed for children learn a second name for an object (e.g., its super-ordinate names). Our model shows the first sense of mutual exclusivity (referent selection in ambiguous situations), but it is not clear if this will also impede learning second labels. This is examined in Simulation 4.1 by training the model on both basic and super-ordinate labels. Crucially, this allows us to study basic-level advantages in a system that does not represent taxonomy hierarchically.

Second, mapping multiple words to objects disrupts the one-to-one mapping between words and objects. This consistency may be essential both for learning in general, but also for the development of mutual exclusivity. However, it is unclear whether a purely associative system can generalize a principle across multiple words. Thus, Simulation 4.2 examines two situations which disrupt this mapping: 1) when words can refer to multiple objects (e.g., polysemy) and 2) when the same object can be referred to by multiple words.

Simulation 4.1: Multiple labels

Learning multiple labels for an object may challenge both constraint and associative approaches. Constraints like the taxonomic constraint or mutual exclusivity must relax to learn properties, synonyms, or other taxonomic categories (e.g., super-ordinates) for the same object. Similarly, associative learning could commit to a single label for an object and have a hard time linking a second one. Xu and Tanenbaum (2007) argue that to solve this problem, the system must be sensitive to statistical distributions and show graded constraint satisfaction. They argue that Bayesian inference uniquely has these properties. Indeed, the localist representations and strong inhibition in our dynamic associative account may make it difficult to assign multiple labels to a one word.

Thus, this simulation examined the ability of the model to handle both basic and superordinate names. Models were initialized with 25 auditory word-form units and 25 object-category units (Table 7). An additional 5 auditory units corresponded to five super-ordinate categories. There were no visual units for these: each super-ordinate was associated with five of the 25 objects. On each training trial, we first selected one of the 30 auditory

units (25 basic-level and five super-ordinate units). Thus, the likelihood of hearing a super-ordinate term was the same as each of the basic-level terms. If the auditory unit was a basic-level name, the corresponding visual unit was active (along with several competitors). If the auditory unit was a super-ordinate, one of its five corresponding basic-level visual units was activated. Consequently, basic-level names and visual units should be strongly associated with each other, while the association between super-ordinate names and their category members may be smaller (since it will be spread among five objects).

Basic-level performance was assessed by presenting one basic-level name, its referent and two competitors (for each of the 25 words). Super-ordinate performance was assessed by selecting a super-ordinate name along with a target from that category, and competitors from two different categories. Each super-ordinate name was tested 5 times (once for each member).

Results—Figure 18A shows the network's performance on both tasks. By the end of training, the model mastered both super-ordinate and basic-level labels performing at 100% on both tasks. Such performance requires the model acquire multiple names for each object, suggesting that the model displays the necessary flexibility. The model also learned basic-level names before super-ordinates, an example of the commonly reported advantage for basic-level terms (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976)

The Role of Frequency—One factor that could contribute to the basic-level advantage is frequency. The network was more likely to hear a basic-level name than a superordinate. To examine this, we ran an additional simulation in which super-ordinate and basic-level labels were equally likely by boosting the frequency of individual super-ordinates (Super-ordinate: p=.1; Basic: p=.02). Figure 18B shows the results (results of the prior simulation are shown in gray). As expected, the network continued to perform well on basic-level categories. This time, super-ordinate performance was closer to basic-level, yet there was still a basic-level advantage.

Spreading Associations—An additional contributor to the basic-level advantage is the fact that superordinate names must share associations with five different objects. For every five exposures to a basic-level name the corresponding object will be seen five times; in contrast, for five exposures to a super-ordinate name, any given basic-level object will be seen once. Thus, super-ordinate names may have weaker connections to each of their members. To test this, we conducted a third simulation in which there were 27 basic-level names and only 3 super-ordinate terms. Each super-ordinate category now had 9 members. This should enhance the spread of associations and lead to even worse performance. To ensure that frequency was not a factor, two versions of the model were run. In the first, the overall frequency of super-ordinate names was matched to the first simulation in this section where super-ordinates were as frequent as basic-level terms. That is, the probability of any of the 3 super-ordinates was 5/30, or 0.167; thus the probability of any individual one was 0.167 / 3 =.055. In the second version, the frequency of individual super-ordinate names was the same as the frequency for individual basic-level names (given the smaller number of super-ordinates. That is, probability of any of the 3 super-ordinates was 3/30=.1, thus the probability of any individual one was .1 / 3 = .033. Figures 18C and D show the results. Regardless of the super-ordinate frequency, networks learned the super-ordinate categories, even with 9 members, however, the basic-level advantage in both was larger than in 18A, confirming that the spreading of association contributes to this effect.

Discussion—These simulations demonstrate that this dynamic associative model is capable of learning both basic-level and super-ordinate categories on the basis of co-

occurrence statistics alone. Thus, the model is considerably more flexible than the one-word-to-one-object mappings we have largely focused on, and the ability to use mutual exclusivity in the moment does not necessarily constrain learning (nor does it rely on such a constraint).

Given the intermediate lexical representations, our model has two routes to achieve a mapping between one object and two words. It could associate a super-ordinate word-form with multiple lexical units, each of which is associated with a single object; or it could associate a super-ordinate word with a single lexical unit that is associated with all of the category members. It is unclear what the consequences of one or the other are, but a preliminary analysis of several networks' suggested that the latter was the dominant pattern —each super-ordinate word was associated with a new lexical unit, which in turn was associated with multiple visual units.

More importantly, however, this model illustrates that behaviors like mutual exclusivity and the use of multiple labels can co-exist comfortably in an associative architecture. The key innovation is that mutual exclusivity is not a constraint on learning (as in Regier, 2005; Xu & Tenenbaum, 2007). Rather, it is a constraint on online behavior (referent selection), that has consequences for an unconstrained learning system. In the moment, the network is nudged away from assigning a new name (the super-ordinate) to a known category (as predicted by mutual exclusivity). However, it must be nudged toward another object in the scene. Across trials, however, the available visual objects are not consistent, so mutual exclusivity never nudges it to consistently select one object for the superordinate name. As a result, much stronger cross-situational statistics take over and establish the correct associations.

Simulation 4.2: Violating one-word-one object

The previous simulation shows that mutual exclusivity need not be a constraint on learning; the model can learn multiple labels for a given word despite the use of mutual exclusivity as an in-the-moment referent selection strategy. The final simulation examines three related issues. First, Simulation 4.1 did not actually test the model's ability to use mutual exclusivity. It is possible that even with the disruption in the one-word/one-object nature of the word/object mapping the model could still learn, but that its ability to use mutual exclusivity is hampered. Many approaches to M.E. referent selection assume that such inferences are built on a realization by the child that each word refers to one object, and therefore a novel object must have a novel name (Halberda, 2006; Markman & Wachtel, 1988). While this generalization must be acquired from the statistics of word/object relationships (which our model is sensitive to), our model has no way to store such a principle or strategy. If violating the one-word/one-object assumption impairs the model's ability to use mutual exclusivity this would be a powerful demonstration that even associative systems can show principled generalization across words.

Second, Simulation 4.1 examined a special case in which objects have multiple names with a clear hierarchy. However, there are also cases in which objects have two equally probable names. An extreme example is bilingual children who learn two words for most objects. An equally important property of real languages is the converse, in which a name can refer to two objects or categories. This property of polysemy is common – most words have multiple meanings, but it may have different consequences for both learning and M.E. referent selection from the many-names/one-category situation.

To examine this, we ran a series of simulations. Two versions were run and varied parametrically. In the *multiple-meanings* models, there were 30 objects and some number of auditory units referred to two of them, while the remainder referred to one. This was varied

in increments of five from 30 unique words (30 words each referring to a single objects—the equivalent of the prior models), to 0 unique words (15 words, each referring to two of 30 objects). In the *multiple-labels* models, there were 30 words, and some number of visual units had two names, varying from 0 objects with two labels (the equivalent of the prior model) to all 15 objects having two labels.

It was not clear how to test this network using our analysis of the weight matrix, so we conducted a 10AFC task. During testing, foils were restricted such that a word would only have one of its referents present on a test trial (this restriction was not present during training). In addition to this, 5 auditory and visual units were not trained and used to test M.E. referent selection as in Simulation 2.2. Parameters (Table 8) were similar to prior simulations, and novel objects appeared (during training) at 17.5% of the 50% referential ambiguity rate (which yielded more realistic M.E. referent selection in the prior simulations).

Results—Figures 19A and 19B show performance on the trained words over the last 25,000 epochs of training for both models. This is shown as a function of the number of words with one-to-one mappings, and separately for words that had a unique mapping, and those that did not. In the *multiple-meanings* model (Figure 19A) all of the models learned both types of words well. Though words with one referent were learned slightly better than words mapping to two objects, both types showed accuracy above 95% across all the simulations. Similarly, in the *multiple-labels* model (Figure 19B), we also see a benefit for learning objects that only have one word, and learning is somewhat lower when no objects have a single label. However, again, performance is excellent with accuracy in the worst condition at 94.6%.

Figures 19C and 19D show mutual exclusivity performance. Chance (33%) is indicated by the dashed line; the black curve shows performance at the end of training; and the gray curve shows performance early in training. At the end of training, the *multiple-meanings* model shows no problem in mutual exclusivity – even the model which completely violated the one-word/one-object mapping (with no words with one referent), performed at 97.9%. This implies that "understanding" this systematicity is not a pre-requisite for mutual exclusivity in this model. This is underscored by our analysis of the models' mutual exclusivity performance early in training (the gray curve). Here the model with no words referring to one object actually performed better than models with more unique mappings. This was somewhat surprising, and awaits empirical testing as there are few analyses of the number of polysemous words for which children are exposed to both meanings. However, it powerfully underscores the fact that in this system mutual exclusivity behavior need not rely on a systematic one-word/one-object bias in the input. Rather, as we described in Simulation 2.2., what is required is that the learning rule preserves some pathway through the weights to get from the novel visual to the novel auditory units. Having more than one referent for each word does not disrupt this (since weight decay relies on exposure to the objects and word individually), thus preserving the ability to use mutual exclusivity. Intriguingly, however, the multiple-labels model showed a very different pattern of performance. Here we saw that even at the end of training, mutual exclusivity performance dropped as there were fewer objects with only a single label. Indeed, in the extreme (with no words with one label) the model barely performed above chance.

Discussion—These simulations show that first and foremost, disrupting the one-word/one-object mapping does not disrupt overall learning performance. While words that had two referents (the *multiple-meanings* model) and objects that had two words (the *multiple-labels* model) were learned slightly less well than those using a one-to-one mapping, this performance decrement was negligible. Thus there's no reason to assume that these

principles cannot be scaled up these more realistic situations, or to situations like bilingual word learning in which most objects will have multiple names.

Second, mutual-exclusivity, as a situation-time process, was largely spared in these models. When words could have two meanings, mutual exclusivity was fine at every level tested, and appeared to develop *faster* with more polysemous words. There has been little work looking at the number of words with multiple meanings that children are exposed to or know, and it is unclear where this benefit derives from in the model. However, this counter-intuitive prediction may be a useful hallmark of this model. Mutual exclusivity was also largely preserved when models had multiple words for each object (the *multiple-labels* model). While performance was degraded with more of these words, it was still above chance (until there were no uniquely mapped words). Thus, situations like the basic/super-ordinate situation described above should not pose a problem for referent selection by mutual exclusivity.

Third, in spite of this overall performance, the decrement seen in the *multiple-labels* model is important. Byers-Heinlein and Werker (2009) suggest that young bilinguals and trilinguals show decrements in using mutual exclusivity. As bilingualism represents the sort of extreme version of a multiple-labels system, our model offers an explanation for these findings. However, the contrast with the *multiple-meanings* model, which did not show this decrement, is revealing. It suggests that in an associative system, it is not a strict one-word/one-meaning structure that is necessary. Words may have any number of referents, as long as many objects are largely named by one label. This suggests an important new dimension to principles like mutual exclusivity and the novel-name-nameless category that have been used to describe this behavior.

More importantly, however, these simulations make the broader point that in a way, this ostensibly associative system has derived a principle from across its training experiences, a principle that it can harness in the moment to make decisions about novel words. However, it does so without any real capacity to represent principles like this. Rather, it seems to have set up its associative weights in such a way that this capacity emerges, in the moment, during real-time competition between words and objects.

General Discussion

There are two levels on which to evaluate this (or any) model. First, we can consider the range of empirical findings it captures. Second we can evaluate the theoretical advance made by the model – does the broader theory tell us something new about word learning.

With respect to the first issue, Table 1 presents a summary of our findings. To briefly summarize, we found that that this model could learn words under conditions of very high referential ambiguity. It shows that differences between production and comprehension arise in part due to the fundamental differences in these tasks. It can model the improvement in familiar word recognition over time, including the structure of individual differences, without any need for semantic elaboration, or for boot-strapping type processes. It shows hallmarks of the power-law of learning while simultaneously showing accelerating vocabulary growth. It can model referent selection via mutual exclusivity, its development, the lack of retention observed by Horst and Samuelson (2008), and differences in multi-linguals (Byers-Heinlein & Werker, 2009). It can also learn multiple names for categories, multiple meanings for words, and it shows a basic-level category advantage.

This model can clearly capture substantial data from very diverse domains of word learning. However, such findings only matter to the extent that they shape our theories of word learning, and our intuitions about the nature of the problem. This discussion will focus on

these issues. We start by describing some critical limitations of our approach, and ask if there is an (even) simpler alternative. We'll then boil down the most important theoretical contributions of this range of work and end by discussing predictions and new directions.

Limitations

Despite the range of phenomena this dynamic associative model can account for, it is not intended as a complete model of word learning and has a number of limitations. Here, we distinguish limitations of the theoretical approach from those of the simple model we developed to illustrate it. There is no reason that similar mechanisms could not be implemented in more complex models to capture an even broader range of behaviors.

First, our model uses localist representations for words and objects, treating each word or category as maximally different from the others. This was deliberately chosen for a number of reasons. Localist representations are easy to interpret and theoretically fairly transparent. More importantly, our twin theoretical claims are best instantiated in this framework. Inhibition between representations is straightforward and easy to implement as a dynamic process in localist representations, while there are no clear ways for doing this with overlapping distributed representations. Similarly, associative learning has long been modeled as linking discrete, localist units, and thus this architecture best captured our approach to learning as well.

Moving to a distributed representation for inputs (e.g., feature vectors for objects) could make it difficult to solve the problem of referential ambiguity. In the current localist scheme, to indicate that multiple objects are present, we simply activate each of their units. However, if objects were represented by a distributed representation across multiple feature vectors, it would be difficult to code more than a few objects – as the feature vectors add up, gradually most if not all of the feature units would be active. As a result, the distributed representation for multiple objects simply becomes a concatenation of all features present in a scene. This makes it very difficult to cope with high levels of referential ambiguity. The problem is that in this simple architecture there is no way to bind the features, to know that activity in some features goes together (apples are red and round, blocks are brown and square). This binding problem is a classic issue with distributed representations (Hinton, McClelland, & Rumelhart, 1986), and work must be done to determine if any of the solutions proposed for this can be integrated with this framework. Perhaps some form of localist (or near localist representations, as seen in topographic maps) is necessary at some level of the system? Additional information in the input (beyond the presence or absences of a feature) may also be helpful, and approaches using spatial location of features to bind them together seem a promising new direction (Johnson, Spencer, & Schoner, 2009; Samuelson, Smith, Perry, & Spencer, 2011). In the meantime, however, our goal was to examine the power of competition and associative learning, and localism presented a clear platform in which to do so without solving these historically difficult problems.

Second, our choice of representations completely ignores a fundamental issue—that word learning must link *categories* of words and/or objects. We've attempted to bypass this by assuming that the auditory and visual inputs really represent the output of some other categorization process, but, we did not model these processes. However, Hebbian Normalized recurrence has already been applied to problems like speech categorization (McMurray, Horst, et al., 2009; McMurray & Spivey, 2000), and non-learning versions have been used to model visual categorization (Spivey, 2007; Spivey & Dale, 2004). Thus, it may be possible to simply chain together such models. Indeed work in progress (McMurray, et al., in press) has already developed a network that first categorizes visual feature vectors, and then uses these categories as the inputs to the network described here. Such an extension may allow the model to capture more of the interesting social, attentional and conceptual

processes that guide children to the right referent in real-time, and our preliminary work on this (McMurray, et al., in press) suggests it can model complex phenomena by which words sometimes impede visual categorization (C. W. Robinson & Sloutsky, 2004) and sometimes facilitate it (Lupyan, et al., 2007).

Even if we allow for such chained models, using localist representations of words and objects appears to make the erroneous assumption that similarity among objects or word-forms is irrelevant to word learning. This obviously cannot be so. Fortunately, such similarity relationships can be implemented in localist scheme. Minimally, one would expect that similar categories should be partially co-active due to their overlapping inputs. For a real child, when *bug* is heard, similar sounding words like *bus* will be partially active (Swingley & Aslin, 2002). Similarly, when a bug is identified in the visual scene, similar categories like ant or spider may be also active. All these competing, partially active categories, could significantly raise the number of spurious associations that would have to be considered and eliminated. However, it is not clear if this would ultimately be problematic: the set of similar sounding words (*deer→gear*, *deal, tear*) is distinct from the set of similar looking objects (deer→*horse, cow, antelope*), so cross-situational statistics may quickly rule out these associations. Conversely, competition dynamics at the level of visual categories could eliminate some of these competitors. Thus, the problem of co-active categories or ambiguous inputs may not be hugely problematic, though it remains to be investigated.

Given all these issues with localist inputs, the self organizing map approach (Li & MacWhinney, 2004; Mayor & Plunkett, 2010) may be a natural bridge between localist representations which ignore similarity and distributed representations which naturally reflect it. Such networks capture similarity relations well, and are based on similar Hebbian learning rules to our network. However, they also have enough topography and competition that input representations are precise. The simple form of one-step competition they use could likely be modified to be more dynamic, but they may need additional cues (like space) to cope with multiple inputs. Thus, using self-organizing maps as the input to our model, or using our settling dynamics in such approaches may offer a useful hybrid.

A third limitation is the scale of our simulations. Most of our simulations used only 35 words, which is small by comparison to the real problem. Simulations reported in the online supplement (Note S3) show the network can learn 150 words with few modifications. We have not yet tested larger lexica than that, but there is no reason in principle why this would fail. Moreover, larger lexica may create more optimal statistics for learning (Sibley, et al., 2008). With a thousand words, the chance of any given competitor appearing is miniscule, so more invariant (correct) associations may pop out quicker. Of course, the larger number of erroneous connections to suppress may also slow learning.

Fourth, our focus on concrete nouns is a limitation, but not problematic. The localist visual units could easily be treated as tags for properties of objects, allowing the model to learn adjectives. Moreover, if the child or model can segment events from the scene (Reynolds, Zacks, & Braver, 2007), visual units could serve as tags for events or actions allowing the model to learn verbs. Siskind (1996) has shown how cross-situational statistics can be used to acquire word meanings from text, potentially enabling this mechanism to be applied to abstract nouns or verbs as well; and Scott and Fisher (in press) have shown that cross-situational statistics could be involved in verb learning as well. More broadly, the fact that lexical representations are situated between multiple layers of input could allow other sources of information (e.g., conceptual) to interact with existing auditory/visual inputs to guide learning—something that would be difficult to accomplish if auditory units were associated directly to visual ones.

Finally, our model illustrates the properties of learning necessary to give rise to behaviors like referent selection via mutual exclusivity (the pattern of weight decay). However, our learning rule is one instantiation of a variety of Hebbian rules, and there may be other versions that are capable of learning more quickly or with fewer lexical units. One could also explore the possibility that supervised (error-driven) learning plays a role. Work in language acquisition more generally suggests that children do receive some feedback from caregivers ranging from quite overt error signals, to more subtle cues like the way in which sentences are repeated back, with or without modification (Bohannon, 1988; Chouinard & Clark, 2003). More directly, in word learning children are corrected in various ways for naming mistakes or incorrect referent-selection (Chouinard & Clark, 2003; Gruendel, 1977; and see additional unpublished evidence cited in Chapman, Leonard, & Mervis, 1986), and there is evidence that children benefit from such feedback in word learning (Chapman, et al., 1986; O'Hanlon & Roberson, 2007). Even beyond this, connectionist models suggest that simple prediction error (e.g., hearing a word, predicting which object(s) are likely to be present, and learning on the basis of the discrepancy) can be an extremely powerful way to use supervised, error-driven learning in an essentially unsupervised context (Elman, 1990). Error driven learning (even on a handful of learning events) may buttress some of the slowness of purely associative processes. However, error-driven learning is likely to have very different consequences for the conditions under which irrelevant associations are suppressed, which will have ramifications throughout the system. We also will have to explore, when, during processing the error signal is available and this could have ramifications for learning. Such effects may ultimately be quite diagnostic, allowing us to identify when and where unsupervised and/or supervised learning contribute to word learning.

At their core these limitations are largely limitations of the simple model we used to explore our broader dynamic associative account. There are clearly more sophisticated competition algorithms, better input representations, and richer approaches to learning that could be incorporated for a more realistic model. However, what is startling is how much of the word learning literature we could to capture by stripping many of these factors out. The delay in productive vocabulary acquisition can be partially accounted for by the nature of the task, without any recourse to articulation, phonology, or perception; relationships between processing time and learning can be accounted for without resource limitations or bootstrapping; so-called "slow-mapping" effects can be observed without the need for semantic elaboration; a general principle (mutual exclusivity) can emerge in a purely associative framework; and basic-level categorization advantages can emerge with no hierarchical semantics. While these explanations are only part of the story for these phenomena, we would have missed them in a more complex model. Thus, the broader theory concerning the linkages of real-time competition and associative learning may have much explanatory power by itself.

Can we get any simpler?

Given this, we might ask if this model could get any simpler. Could this range of processes derive from even more basic principles? There is some impetus to think about word learning in this way; for example, the McMurray (2007) model of the vocabulary spurt modeled learning as simply accumulating points, and discovered that acceleration falls out of parallel learning.

In terms of information processing (e.g., Marr's first level of description), perhaps the core of our model is cross-situational learning. We modeled this via Hebbian associations, but one can think of this in even simpler terms as simply co-occurrence counts between words and objects. Yu and Smith (Yu & Smith, 2012) modeled adult cross-situational learning with exactly such a model (their "bare bones" Dumb Associative Model) and found versions of it

performed quite similarly to a hypothesis testing model. Medina, Snedeker, Trueswell, and Gleitman (2011) have argued that co-occurrence counts are not consistent with the fact that people reap more benefit from low-ambiguity statistics early in training, since by the end of training, the accumulated statistics are the same (though we have modeled such effects using our dynamic associative network: McMurray, et al., in press). But given this discussion, it is worth asking if our approach offers anything substantive over and above a simpler co-occurrence counter.

To some extent, many of our effects can derive from statistics alone. Yu and Smith (2012) have argued that under some circumstances Hebbian learning can compute a co-occurrence matrix quite directly, and our model's ability to learn cross-situationally derives from this. It is likely that our findings of a basic-level advantage (stronger word-object associations for basic-level terms which have one referent than super-ordinate terms which have multiple) also derive from this. However, interestingly, Yu and Smith's (2012) associative model only reached about 40% correct in their most difficult condition (18 words / 4 presented per trial) that corresponds to a referential ambiguity rate of 17.6%. Our model is at ceiling under these circumstances (though with more training). Of course, our model also includes real-time processing (which consequently enables slower learning) which may enable better learning.

Other findings do not directly fall out of co-occurrence statistics, but would require real-time processing. Differences in word learning or referent selection based on the task (number of alternatives) could not be accounted for with co-occurrence alone. However, they could if co-occurrence statistics were used as the input to some kind of read-out rule (as in Yu and Smith, 2012). Our finding of decreasing RTs over training would also require some sort of decision rule that converts co-occurrence information into RTs (maybe something like Ratcliff & Rouder, 1998; Usher & McClelland, 2001). However, these decision rules would need to be built or tuned by hand to account for things like exponential decay in RTs, whereas in our model, this is an emergent property of our core theoretical principles, competition and learning. Moreover, at this point, the model would start to look quite similar to ours, and it is not clear what would be gained. Perhaps most importantly, these decision rules would ultimately be just a read-out of learning, and would not interact substantively with it – that is, the real-time decisions made by the model during training would have no effect on learning. This could be problematic for modeling of Fernald et al's (2006) longitudinal work. We've shown in multiple simulations here how important this linking between real-time and developmental processes can be, and in our other work (McMurray, et al., in press) suggest it may make all the difference to modeling results like Medina et al. (2011). If we are going to have to add some sort of dynamic decision process, we may as well be modeling the interaction of this with learning.

Our work on mutual exclusivity is perhaps the most difficult to account for on the basis of co-occurrence. A novel name will have no co-occurrences data for either the familiar or novel objects in the scene, and thus no principled way to show a bias. As a result, any pure co-occurrence counter would need some decision rule to decide when to go with the novel object. Again, this would have to be built in, to account for these results, rather than emerging out of the same process that recognizes novel words. But how much evidence counts as no evidence? If the model had seen an object once, or twice how would it make this decision? And what about retention? Once the model has a single piece of evidence for a word/object pairing, the most optimal thing is to return to that on retention trials, and yet children apparently do not. And how would this "decision rule" develop? Finally, our results on mutual exclusivity when words have more than one referent (or referents have more than one word) suggest that minimally the co-occurrence statistics are not symmetrical: having two words for an object can hurt mutual exclusivity while the converse does not. This lends credence to our use of internal (lexical) representations to mediate these co-occurences. But

more importantly, the fact that the multi-lingual models do not show mutual exclusivity suggest that the development of this behavior is sensitive to the word/object statistics in general, and suggests that our associative system can learn a principle (or will fail to if the statistics demand). It is not clear how this could emerge out of pure co-occurrence counts. The bottom line: in order to do mutual exclusivity and capture the range of effects we have using only co-occurrence counts, one would have to build it into the model, and build it in such a way to that it can explicitly capture these effects. This treats behaviors like mutual exclusivity as fundamentally different from familiar word recognition – they use different decision rules – while our model shows how both of them emerge from the same competition scheme. This would lead one to a very different theoretical conclusion that what we have shown here.

Thus, our model is doing something more than just co-occurrence statistics. It is the unique interaction of learning and processing, embedded in an environment with such statistics, that enables complex behavior to emerge from these mechanisms. Indeed, reductionistically simplifying the model further would require us to put substantial content in the situation-time decision rules, and it may not be able to model all of these effects anyways. More importantly, it may lead to very different theoretical conclusions.

Theoretical Insights

While our model was able to capture numerous empirical findings, its strength lies in its ability to highlight new theoretical conclusions about word learning, conclusions that are substantially broader than the rather narrow model we have presented.

Learning can (and should) be slow—Most accounts of word learning stress its effortlessness and speed: children appear to acquire words very rapidly. This is based on phenomena like so-called fast-mapping and the vocabulary explosion. Such learning is seen as difficult for associative accounts, and thus associative accounts of word learning often stress their rapidity (Mayor & Plunkett, 2010; Regier, 2005).

But is word learning really that fast? Children hear approximately 17,000 words/day (Hart & Risley, 1995). By one year, when the first word is produced, an average child will have heard 6 million words. Even at the height of the vocabulary explosion, children may show evidence of having acquired 15-20 words per week. So slow learning may be required, just to account for the actual acquisition curves.

One might argue that within these thousands of words the child is hearing, there may be few tokens of the individual words that a child appears to learn in a given week, so this still necessitates fast learning. However, it is important to point out that under associative accounts like this one, much of the problem is simply suppressing competing associations, something that is less dependent on the specific words being heard. So thousands of exposures to *duck* and *goose* may help children improve their ability to process and acquire *chicken*. Despite apparently quick learning in the lab, real learning may be fairly slow, but it also may benefit from much broader experience than we normally "count" – experience with other words, and with objects alone is all relevant for learning a particular word.

This contrasts with more inferential or hypothesis testing approaches which consider hypothetical word/meaning in parallel, and wait for the right data to update (Medina, et al., 2011). In a sense, due to random initial connections, associative systems start by considering all hypotheses (with some variation in strength). A true hypothesis-testing system could never rule them all out, working in serial. However, by suppressing many connections in parallel at each naming instance this becomes more feasible. In a sense, at each word, there are global, albeit small, changes in the hypothesis space. This is made even more feasible by

smarter situation-time processes that allow the child to behave in the moment on the basis of partial knowledge (Simulations 2.2, 2.3, 3.2, 4.3). Considering familiar words in the same framework as novel words only underscores this: developmental changes in familiar word recognition take years to unfold (e.g., Fernald et al, 1998; 2000), even as the child is generally able to select the correct referent fairly early in life. Thus, slow, but global changes on each naming instance may yield a fairly robust system that accounts for multiple developmental phenomena.

And why it should be otherwise? The situations that children find themselves in are inherently ambiguous—there are multiple visual referents, and multiple possible interpretations for any given word. Even much vaunted social cues are not consistently available and do not consistently disambiguate words (Frank, et al., 2009). Thus, cross-situational statistics may constitute a good portion of the information available to link words to objects. If this is the case, then slow learning may be more optimal in that it prevents children from committing too strongly to a single (perhaps erroneous) mapping before they have enough data. Indeed, we investigated this in simulations reported in the online Supplement (S4) and found quite poor learning when the learning rate of the model was too high.

Our simulations on mutual exclusivity (3.1, 3.2 and 4.2) underscore this. Simulation 3.1 captures Horst and Samuelson (2008), showing that the model can use mutual exclusivity for referent selection but it retains very little from this. The dependence of mutual exclusivity on learning (2.1), task configuration (3.2), and lexical statistics (4.2) argue that this may be an unstable platform for learning; and the fact that multi-lingual children and models (4.2) still learn words suggest it is not required for learning. Indeed, if referent selection by mutual exclusivity was uniquely powerful for learning, one must ask how often children know the name of every item in the visual scene but one. This seems a fairly unlikely event, underscoring the importance of slower, more gradual mechanisms.

But if learning is slow, this raises the question of how children function while they wait for data to accumulate. This is not just theoretical—slow-learning conflicts with the excellent performance we see in many constrained laboratory tasks, and that parents observe every day in their toddlers. Our dynamic associative account suggests that fast, situation-time processes enable children to take advantage of constraints offered by the environment and children's own incomplete mappings to perform impressively in day-to-day and laboratory tasks, even while learning is slow. This is clear when we compared the model's performance on constrained tasks to its underlying weights: the model's knowledge was incomplete, but it still performed well on NAFC tasks. It also appeared when we compared comprehension to production: comprehension is necessarily constrained by the response options and was consistently better than less constrained production.

Even associative learning is multi-faceted—Our approach to associative learning is more complex than commonly considered. This has theoretical consequences beyond our model. First, since auditory and visual units are independently associated with the lexicon, learning on the auditory and visual side can have different effects. This was most apparent in Simulation 4.2 where having two words per object degraded mutual exclusivity performance, but having two objects per word did not. Even in a simple associative framework, the relevant statistics (for example the fact that children encounter many visual competitors and fewer auditory ones) shape situations in which both auditory and visual units may perform different roles.

This was also seen in Simulations 2.3 and 3.1 which demonstrated how familiarity with the visual objects alone can improve both referent selection by mutual exclusivity and retention

(Kucker & Samuelson, 2012), even in the absence of learning any specific content (the objects' properties). This may be an important consideration in future work on the M.E. referent selection task. Similarly, Horst et al. (2011) recently showed that in a referent selection task with completely novel objects pure visual familiarity with some of the objects can bias children's performance[8]. However, scaling these ideas to the real world, it is clear that novelty may be a major factor in early lexical behavior. Familiarity is graded and there are few, if any, words that a child has literally no experience with. Thus, our artificial segregation of objects as familiar or novel (both in models and in typical laboratory tasks) may not capture the situation in the real world. Crucially, even an associative learning framework this shows that not all learning needs to involved both "ends" of the associative link.

This dynamic associative model also shows the surprising importance of suppressing unnecessary associations. This turned out to be the biggest predictor of settling time for familiar word recognition (Simulation 2.1), and the pattern of weight decay was essential for referent selection by mutual exclusivity (Simulation 2.2). This is because the bulk of learning consists of simply suppressing the vast number of irrelevant connections and this can be done for virtually any naming event. This raises the possibility that even when children select the wrong referent, or do not select a referent at all, they may still be doing useful learning. This non-obvious source of learning and has not been considered in prior theoretical, empirical and computational work.

Finally, our work on referent selection by mutual exclusivity suggests that associations, even among localist inputs, can derive a principle that applies to even novel items. The ability to use mutual exclusivity is not built into this model and clearly develops from the input. More importantly, it can be blocked when there are two words for many objects. In a companion piece (McMurray, et al., in press) we've conducted an extensive parameter search of the models ability to use mutual exclusivity and discovered that only models that have both real-time competition dynamics, and that use internal representations (rather than directly linking words to referents) can do this. Thus, when associative learning is embedded in a more realistic system with both real-time processing and abstract representations, much richer, emergent behavior can arise.

Learning and processing are quasi-independent—Our model was built on the theoretical commitment that using words and learning them are different. Learning is accomplished by changing the connections between words, objects and the lexicon; processing occurs when real-time competition allows activation to flow over those weights to arrive at a solution. This quasi-independence is fundamental. It allows the model to show dissociations between mutual exclusivity and retention (Simulation 3.1) and between performance and knowledge (Simulation 1.1, 1.2). Moreover, by considering processing independently we also showed how both novel and familiar word processing can be handled by the same system (Simulations 2.1-2.3).

The quasi-independence has useful functional consequences. By offloading mutual exclusivity to online processing, it no longer blocks learning of multiple names for a given object (Simulation 4.1 and 4.2). More broadly, learning can be less than perfect because processing can get the child the rest of the way. This is what compensates for slow learning. The best system will be one that uses processes optimized for learning to handle developmental-time learning, and processes optimized for in-the-moment demands to handle real-time behavior.

---

[8]In our model, this novelty bias can derive from something as basic as the random initial connections (that persist for novel objects), again pointing to the importance of pruning (or not).

Nonetheless, while computationally these are distinct processes, they are not completely independent in practice: situation-time processes are dependent on learning. The changes in RT for familiar words derive from learning, and the ability of the model to use mutual exclusivity map derives from a weight matrix created by the specific learning rule. Many people describe fast-mapping as the sort of initial stages of a slow learning process (e.g., Capone & McGregor, 2005; Carey & Bartlett, 1978; Golinkoff, et al., 1992), a sort of incomplete learning. In contrast, we suggest that mutual exclusivity behavior (referent selection) is a real-time *product* of the type of learning that has occurred up to that point, but this behavior in turn leaves an associative trace that can build over repetitions to yield word learning. In this light, the learning on the first exposure of a word (what has been termed fast-mapping) is no different from subsequent exposures (slow-mapping).

Simulation 3.4 showed the converse, suggesting that learning is impossible without processing. The necessity of such competitive processes is implicit in many unsupervised learning models, but its importance has not been highlighted before. Competitive learning (Rumelhart & Zipser, 1986), for example, requires winner take all learning; self-organizing maps (Kohonen, 1982; Mayor & Plunkett, 2010) include a competition/interaction kernel; and even the quite unrelated mixture of Gaussians framework for clustering benefits from competition (McMurray, Aslin, et al., 2009). However, while competition is essential, its outcome can be variable. In simulations not reported here we have found that useful learning still occurs even when competition gets the wrong referent. Thus, the presence of competition is necessary for learning—the system must make a choice. But, on any trial, the specific choice is less important.

Finally, Fernald et al's (2006) longitudinal study is perhaps the best evidence for the dependence of learning and processing as they found children's RTs predicted acceleration in learning. Our model also showed a similar pattern of results, but suggested no simple construct to explain it. Learning rate and settling time were the product of parameters that control learning and processing; the same parameter (e.g., temperature) acts differently depending on the referential ambiguity; and the biggest predictor of RT was the nature of the learned weights. Thus, to understand a functional property of the child like "speed of processing" we must understand the myriad of components of both processing and learning.

Word learning does need not be specialized—A number of models suggest that acquiring vocabulary may harness general purpose learning mechanisms. Regier (2005) and Mayor & Plunkett (2010) show the unexpected power of association learning. Xu and Tenenbaum (2007) and Frank et al (2009) use general Bayesian inference mechanisms. McMurray (2007) suggests that acceleration is not the hallmark of a specialized system but should be seen virtually everywhere. There is also empirical work showing that similar principles may span word learning and other types of learning, whether they derive from general reasoning strategies that apply to both facts and words (Behrend, et al., 2001; P. Bloom & Markson, 1998; Markson & Bloom, 1997; though see Waxman & Booth, 2000) or lower level attention (Samuelson & Smith, 2000) and novelty biases (Horst, et al., 2011). Thus, there is a mounting effort to explain vocabulary acquisition in terms of general cognitive and learning processes.

Our dynamic associative account adds to this. First, it shows that novel word inference and familiar word recognition can arise from the same system. Both use the same set of online processes (which themselves are quite general); both operate over the same mappings (weights) that are shaped by the same learning mechanism. There is no need for any sort of monitoring mechanism to route novel words to more constrained or specialized learning mechanism. Moreover, referent selection by mutual exclusivity can be modeled with

something as general as dynamic competition, further emphasizing the generality of these processes.

Second, the patterns of learning observed in vocabulary are not special. At face value, the decelerating learning predicted by the ubiquitous power law of learning conflicts with acceleration during the vocabulary explosion. Our model highlights this conflict, showing both acceleration and deceleration, and both derive from changes in the associative weights (Simulation 2.1). However, the particular changes in the weights that yield gains in RT are not equivalent to those that allow the model to appear as if it acquired a new word: while RT changes largely derive from suppressing unnecessary connections, changes in the number of words known require both the suppression of unnecessary connections and the establishment of the right positive associations. Thus, depending on the measurement (changes in RT or vocabulary size), and the real-time processing that give rise to the behavior, we may reach different conclusions about the shape of learning (accelerative or decelerative), even as the underlying mechanism is the same. This further cements word learning as a general process, but challenges learning theory by suggesting that changes in RT may not fully describe learning.

Finally, our use of associative learning does not entail a particular source of information: our framework is consistent with information sources like attentional or social cues, or conceptual structure that have been widely interpreted as special. These exist outside of the core lexical mapping system, and constrain the settling dynamics (from the outside), or simply determine the type of representations that are associated. Thus, such higher order factors may be fundamental to learning and/or lexical behavior, without needing to be embedded in the learning system. This permits a soft coupling: as children gain sensitivity to things like speakers' intentions (Moore, 2008), these sources of information gradually play a larger role in shaping online behavior (and through it, learning) without fundamentally restructuring word learning.

New Directions

The test of any model is its ability to make predictions and highlight new research questions. Given the simplicity of the model, it is not clear that we are in a position to make precise empirical predictions for new tests and paradigms. Nonetheless, the model and the broader theoretical view suggest a number of important new areas of investigation in word learning.

Indeed during the course of writing this manuscript a number of predictions from the model were tested empirically. Our finding that M.E. referent selection relies on the relative randomness of the weights connecting the auditory and visual units, and the role of simple (purely visual) experience suppressing them (Simulation 2.1) led us to predict, and confirm, that familiarity with objects may bias children away from selecting them in referent selection tasks (Horst, et al., 2011); and our demonstration in Simulation 3.1 that visual familiarity can simultaneously influence retention motivated Kucker and Samuelson (2012), which showed similar results in children. Taken together these simulations suggest that mutual exclusivity may have two components: a component driven by novelty that leads to excellent referent selection but more retention; and something that resembles a constraint satisfaction, which leads to somewhat worse (but still good) referent-selection, but much better retention. This trade-off should be explored, particularly as these components wax and wane over development.

Similarly, while the Simulations in 4.2 match evidence that bilingual children—who have multiple names for many objects—may perform worse in mutual exclusivity tasks (Byers-Heinlein & Werker, 2009), they also suggest the converse—having multiple meanings for many words—may not be problematic. There has been little work on the statistics of word-

object mappings in the child's environment and how they relate to behaviors like mutual exclusivity (analogous to the Perry & Samuelson, 2011, and Samuelson & Smith, 1999, studies of how such statistics predict the shape and material biases), and it is not entirely clear why our model shows this asymmetry, but this is clear avenue for future work.

More broadly, our work points to a host of issues that the statistics of word-object mappings (co-occurrence statistics) may be involved in. Simulation 4.1 demonstrated how this can give rise to an advantage of basic-level over super-ordinate category labels. An important part of this is that super-ordinate terms have their associations spread across multiple objects, while basic-level terms are only associated with one. Conversely, we saw in our simulations of familiar word recognition how suppressing irrelevant connections was crucial to improving performance, and that this could only occur when competitors were present (but variable from trial to trial). In this case, the spreading of association across competitors prevents any of them from becoming strongly linked with the target word. This spreading of associations has been invoked in a number of other domains including early speech perception, where spreading associations block children from associating irrelevant talker cues with words (Apfelbaum & McMurray, 2011), and in semantic memory where "context variability" prevents aspects of the context from serving as a retrieval cue (Steyvers, 2003). More broadly, however, these simulations force us to think more creatively about the co-occurrence statistics of words and objects – our distillation of the problem suggests that they may play a role in numerous domains.

A third avenue of future study is the role of suppressing irrelevant connections. This is pivotal in predicting changes in processing speed and giving rise to referent selection by mutual exclusivity. However, this non-obvious result of learning has not received extensive study. More complex eye-movement paradigms, may allow us to index the strength of competing associations, to look for correlations with these behaviors, and artificial language paradigms may allow us to manipulate it by temporarily creating strong spurious associations. Indeed, these may be ultimately better predictors of behaviors like mutual-exclusivity and word recognition time. Fitneva and Christiansen (2011) recently demonstrated that in a cross-situational word learning paradigm, participants who looked longer to the *incorrect* objects during learning showed better performance. This clearly is consistent with the idea that suppressing competing associations is critical for learning and points to a paradigm in which to investigate these issues.

Finally, and most importantly, our approach suggests that even in an associative account, what the child does and the exact sequence of events will matter. For example, the timing of the events in a learning trial could influence whether or not there is sufficient time for competition among referents to resolve, and this would alter learning dramatically. Similarly, the configuration of items on a learning trial (e.g., the number of competitors), and the behavior of the child can both affect learning, by influencing how positive associations are formed (which require a fairly specific confluence of events), and how negative associations are suppressed (which may be more general, and not require a correct response).

Conclusions

Lexical behavior must fundamentally be considered on two timescales – children learn words over development, but they must also use them here and now. Word learning is not about acquisition of words as a type of knowledge; rather we must study how children acquire the *abilities* to recognize and produce words, and infer the meanings of novel words. By embedding learning within a structure of word use, our model offers a unified account for a range of findings in word learning, word recognition and novel word inference. In this

framework, word learning is the simple product of ongoing interactions between developmental-time processes like associative learning, and situation-time processes like dynamic competition.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

Akhtar, N.; Martinez-Sussman, C. Intentional communication. In: Brownell, C.; Kopp, C., editors. Socioemotional development in the toddler years: Transitions and transformations. Guilford Press; New York, NY: 2007. p. 201-220.

Allopenna P, Magnuson JS, Tanenhaus MK. Tracking the time course of spoken word recognition using eye-movements: evidence for continuous mapping models. Journal of Memory and Language. 1998; 38(4):419–439.

Anderson JR. Acquisition of cognitive skill. Psychological Review. 1982; 89:396–406.

Apfelbaum K, McMurray B. Using Variability to Guide Dimensional Weighting: Associative Mechanisms in Early Word Learning. Cognitive Science. 2011; 35(6):1105–1138. [PubMed: 21609356]

Baldwin DA. Infants' contribution to the achievement of joint reference. Child Development. 1991; 62:875–890. [PubMed: 1756664]

Baldwin DA, Markman EM, Bill B, Desjardins RN, Irwin JM. Infants' reliance on a social criterion for establishing word-object relations. Child Development. 1996; 67(6):3135–3153. [PubMed: 9071774]

Bauer P, Dow G, Hertsgaard L. Effects of prototypicality on categorization in 1- to 2-year-olds: Getting down to basic. Cognitive Development. 1995; 10(1):43–68.

Behl-Chadha G. Basic-level and superordinate-like categorical representations in early infancy. Cognition. 1996; 60(2):105–141. [PubMed: 8811742]

Behrend DA, Scofield J, Kleinknecht EE. Beyond fast mapping: Young children's extensions of novel words and novel facts. Developmental Psychology. 2001; 37(5):698–705. [PubMed: 11552764]

Bharucha JJ. Music cognition and perceptual facilitation: A connectionist framework. Music Perception. 1987; 5:1–30.

Bloom, L. One word at a time: The use of single-word utterances before syntax. Mouton; The Hague: 1973.

Bloom, P. How Children Learn the Meanings of Words. The MIT Press; Cambridte, MA: 2000.

Bloom P, Markson L. Capacities underlying word learning. Trends in Cognitive Sciences. 1998; 2(2): 67–73. [PubMed: 21227068]

Bohannon JNS, Laura B. The issue of negative evidence: Adult responses to children's language errors. Developmental Psychology. 1988; 24(5):684–689.

Bourne LE, Restle F. Mathematical theory of concept identification. Psychological Review. 1959; 66:278–296. [PubMed: 13803353]

Bush RR, Mosteller F. A model for stimulus generalization and discrimination. Psychological Review. 1951; 58(6):413–423. [PubMed: 14900302]

Byers-Heinlein K, Werker JF. Monolingual, bilingual, trilingual: infants' language experience influences the development of a word-learning heuristic. Developmental Science. 2009; 12:815–823. [PubMed: 19702772]

Capone N, McGregor KK. The effect of semantic representation on toddlers' word retrieval. Journal of Speech, Language, and Hearing Research. 2005; 48:1468–1480.

Carey, S. The child as word learner. In: Halle, M.; Brensnan, J.; Miller, A., editors. Linguistic Theory and Psychological Reality. The MIT Press; Cambridge, MA: 1978. p. 264-293.

Carey S, Bartlett E. Acquiring a single new word. Proceedings of the Stanford Child Language Conference. 1978; 15:17–29.

Chambers CG, Tanenhaus MK, Magnuson JS. Action-based affordances and syntactic ambiguity resolution. Journal of Experimental Psychology: Learning, Memory & Cognition. 2004; 30:687–696.

Chapman KL, Leonard LB, Mervis CB. The effect of feedback on young children's inappropriate word usage. Journal of Child Language. 1986; 13(1):101–117. doi: doi:10.1017/S0305000900000325. [PubMed: 3949891]

Chouinard MM, Clark EV. Adult reformulations of child errors as negative evidence. Journal of Child Language. 2003; 30(03):637–669. doi: doi:10.1017/S0305000903005701. [PubMed: 14513471]

Colunga E, Smith LB. From the Lexicon to Expectations About Kinds: A Role for Associative Learning. Psychological Review. 2005; 112(2):347–382. [PubMed: 15783290]

Dale PS, Fenson L. Lexical development norms for young children. Behavior Research Method, Instruments and Computers. 1996; 28(1):125–127.

Dell GS. A spreading activation theory of retrieval in sentence production. Psychological Review. 1986; 93:283–321. [PubMed: 3749399]

Diesendruck G, Markson L. Children's avoidance of lexical overlap: a pragmatic account. Developmental Psychology. 2001; 37:630–641. [PubMed: 11552759]

Dore J, Franklin MB, Miller RT, Ramer AL. Transitional phenomena in early language acquisition. Journal of Child Language. 1976; 3(1):13–28.

Elman JL. Finding structure in time. Cognitive Science. 1990; 14:179.

Elman JL. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. Cognitive Science. 2009; 33:547–582. [PubMed: 19662108]

Elman, JL.; Bates, E.; Johnson, M.; Karmiloff-Smith, A.; Parisi, D.; Plunkett, K. Rethinking Innateness: a Connectionist Perspective on Development. The MIT Press; Cambridge, MA: 1996.

Fenson L, Dale PS, Reznick JS, Bates ET,D, Pethick S. Variability in early communicative development. Monographs of the Society for Research in Child Development. 1994; 59(5):1–173. [PubMed: 7845413]

Fernald A, Perfors A, Marchman VA. Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. Developmental Psychology. 2006; 42(1):98–116. [PubMed: 16420121]

Fernald A, Pinto JP, Swingley D, Weinberg A, McRoberts GW. Rapid gains in speed of verbal processing by infants in the second year. Psychological Science. 1998; 9:72–75.

Fernald A, Swingley D, Pinto JP. When half a word is enough: Infants can recognize spoken words using partial phonetic information. Child Development. 2001; 72(4):1003–1015. [PubMed: 11480931]

Fitneva SA, Christiansen MH. Looking in the Wrong Direction Correlates With More Accurate Word Learning. Cognitive Science. 2011; 35(2):367–380. doi: 10.1111/j.1551-6709.2010.01156.x. [PubMed: 21429004]

Frank MC, Goodman ND, Tenenbaum J. Using speakers' referential intentions to model early cross-situational word learning. Psychological Science. 2009; 20:578–585. [PubMed: 19389131]

Fulkerson AL, Waxman SR. Words (but not Tones) Facilitate Object Categorization: Evidence From 6- and 12-Month-Olds. Congnition. 2007; 105(1):218–228.

Ganger J, Brent M. Reexamining the Vocabulary Spurt. Developmental Psychology. 2004; 40(4):621–632. [PubMed: 15238048]

Goldinger SD. Echoes of Echos? An episodic theory of lexical access. Psychological Review. 1998; 105:251–279. [PubMed: 9577239]

Goldstone RL, Medin DL. The time course of comparison. Journal of Experimental Psychology: Learning Memory and Cognition. 1994; 20:29–50.

Golinkoff RM, Hirsh-Pasek K. Baby wordsmith: From associationist to social sophisticate. Current Directions in Psychological Science. 2006; 15:30–33.

Golinkoff RM, Hirsh-Pasek K, Bailey LM, Wenger NR. Young-Children and Adults Use Lexical Principles to Learn New Nouns. Developmental Psychology. 1992; 28(1):99–108.

Golinkoff, RM.; Hirsh-Pasek, K.; Bloom, L.; Smith, LB.; Woodward, AL.; Akhtar, N.; Hollich, G. Becoming a Word Learner: A Debate on Lexical Acquisition. Oxford University Press; New York, NY: 2000.

Golinkoff RM, Mervis CB, Hirsh-Pasek K. Early object labels: The case for a developmental lexical principles framework. Journal of Child Language. 1994; 21:125–155. [PubMed: 8006089]

Goodman JC, McDonough L, Brown NB. The role of semantic context and memory in the acquisition of novel nouns. Child Development. 1998; 69:1330–1344. [PubMed: 9839419]

Grassmann S, Tomasello M. Young children follow pointing over words in interpreting acts of reference. Developmental Science. 2010; 13(1):252–263. [PubMed: 20121881]

Grossberg S. Adaptive pattern classification and universal recoding, Part I: Parallel development and coding of neural feature detectors. Biological Cybernetics. 1976; 23:121–134. [PubMed: 974165]

Gruendel JM. Referential Extension in Early Language Development. Child Development. 1977; 48(4):1567–1576.

Gupta P, Tisdale J. Does phonological short-term memory causally determine vocabulary learning? Toward a computational resolution of the debate. Journal of Memory and Language. 2009; 61(4): 481–502.

Halberda J. The development of a word-learning strategy. Cognition. 2003; 87:B23–B34. [PubMed: 12499109]

Halberda J. Is this a dax which I see before me? Use of the logical argument disjunctive syllogism supports word-learning in children and adults. Cognitive Psychology. 2006; 53(4):310–344. [PubMed: 16875685]

Hanna JE, Tanenhaus MK. Pragmatic effects on reference resolution in a collaborative task: evidence from eye movements. Cognitive Science. 2004; 28:105–115.

Harm MW, Seidenberg MS. Phonology, Reading Acquisition, and Dyslexia: Insights from Connectionist Models. Psychological Review. 1999; 106(3):491–528. [PubMed: 10467896]

Hart, B.; Risley, T. Meaningful differences in the everyday experience of young American children. Paul Brookes Publishing; Baltimore, MD: 1995.

Heathcote A, Brown S, Mewhort DJK. The power law repealed: The case for an exponential law of practice. Psychonomic Bulletin & Review. 2000; 7:185–207. [PubMed: 10909131]

Hinton, G.; McClelland, JL.; Rumelhart, D. Distributed Representations. In: Rumelhart, D.; McClelland, JL.; T. P. R. Group. , editors. Parallel Distributed Processing. Vol. Vol. 1: Foundations. The MIT Press; Cambridge, MA: 1986. p. 77-109.Vol. 1

Horst, JS.; McMurray, B.; Samuelson, LK. Online processing is essential for learning: Understanding fast mapping and word learning in a dynamic connectionist architecture. In: Sun, R., editor. The proceedings of the 28th meeting of the Cognitive Science Society; Austin, TX: The Cognitive Science Society; 2006. p. 339-343.

Horst JS, Samuelson L. Fast mapping but poor retention in 24-month-old infants. Infancy. 2008; 13(2): 128–157.

Horst JS, Samuelson LK, Kucker S, McMurray B. What's new? Children prefer novelty in referent selection. Cognition. 2011; 118(2):234–244. [PubMed: 21092945]

Horst JS, Scott EJ, Pollard JA. The role of competition in word learning via referent selection. Developmental Science. 2010; 13(5):706–713. [PubMed: 20712736]

Hurtado N, Marchman VA, Fernald A. Spoken word recognition by Latino children learning Spanish as their first language. Journal of Child Language. 2005; 34:227–249. [PubMed: 17542157]

Jaswal VK, Hansen MB. Mutual exclusivity and social-pragmatics Learning words: children disregard some pragmatic information that conflicts with mutual exclusivity. Developmental Science. 2006; 9(2):158–165. [PubMed: 16472316]

Johnson JS, Spencer JP, Schoner G. A layered neural architecture for the consolidation, maintenance, and updating of representations in visual working memory. Brain Research. 2009; 1299:17–32. [PubMed: 19607817]

Jones M, Love BC. Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. Behavioral and Brain Sciences. 2011; 34(04):169–188. doi: doi:10.1017/S0140525X10003134. [PubMed: 21864419]

Kail R. A method for studying the generalized slowing hypothesis in children with specific language impairment. Journal of Speech Language and Hearing Research. 1994; 37:418–421.

Kohonen T. Self-organized formation of topologically correct feature maps. Biological Cybernetics. 1982; 43:59–69.

Kucker S, Samuelson L. The First Slow Step: Differential Effects of Object and Word-Form Familiarization on Retention of Fast-Mapped Words. Infancy. 2012; 17(3):295–323. [PubMed: 22661907]

Li P, MacWhinney B. Early lexical development in a self-organizing neural network. Neural Networks. 2004; 17(8-9):1345–1362. [PubMed: 15555870]

Livesey, EJ.; McLaren, IPL. An elemental model of associative learning and memory. In: Pothos, E.; Wills, AJ., editors. Formal Approaches in Categorization. Cambridge University Press; Cambridge, UK: 2011. p. 153-172.

Logan GD. Shapes of reaction-time distributions and shapes of learning curves: A test of the instance theory of automaticity. Journal of Experimental Psychology: Learning, Memory & Cognition. 1992; 18:883–914.

Lupyan G, Rakison DH, McClelland JL. Language is not just for talking: Redundant labels facilitate learning of novel categories. Psychological Science. 2007; 18(12):1077–1083. [PubMed: 18031415]

MacDonald MC, Pearlmutter NJ, Seidenberg MS. Lexical nature of syntactic ambiguity resolution. Psychological Review. 1994; 101:676–703. [PubMed: 7984711]

MacWhinney, B. The competition model. In: MacWhinney, B., editor. Mechanisms of language acquisition. Lawrence Erlbaum Associates; Hillsdale, NJ: 1987. p. 249-308.

Marchman VA, Fernald A. Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. Developmental Science. 2008; 11(3):F9–F16. [PubMed: 18466367]

Mareschal D, Tan SH. Flexible and Context-Dependent Categorization by Eighteen-Month-Olds. Child Development. 2007; 78(1):19–37. [PubMed: 17328691]

Markman EM. Constraints children place on word meanings. Cognitive Science. 1990; 14(1):57–77.

Markman EM, Wachtel G. Children's use of mutual exclusivity to constrain the meaning of words. Cognitive Psychology. 1988; 20(2):121–157. [PubMed: 3365937]

Markman EM, Wasow JL, Hanson MB. Use of the mutual exclusivity assumption by young word learners. Cognitive Psychology. 2003; 47:241–275. [PubMed: 14559217]

Markson EM, Bloom P. Evidence against a dedicated system for word learning in children. Nature. 1997; 385(6619):813–815. [PubMed: 9039912]

Marslen-Wilson W. Functional parallelism in spoken word recognition. Cognition. 1987; 25(1-2):71–102. [PubMed: 3581730]

Mayor J, Plunkett K. A neurocomputational account of taxonomic responding and fast mapping in early word learning. Psychological Review. 2010; 117(1):1–31. [PubMed: 20063962]

McClelland JL, Elman JL. The TRACE model of speech perception. Cognitive Psychology. 1986; 18(1):1–86. [PubMed: 3753912]

McClelland JL, Mirman D, Holt LL. Are there interactive processes in speech perception? Trends in Cognitive Sciences. 2006; 10(8):363–369. [PubMed: 16843037]

McMurray B. Defusing the childhood vocabulary explosion. Science. 2007; 317(5838):631. [PubMed: 17673655]

McMurray B, Aslin RN, Toscano JC. Statistical learning of phonetic categories: Insights from a computational approach. Developmental Science. 2009; 12(3):369–379. [PubMed: 19371359]

McMurray, B.; Horst, JS.; Toscano, JC.; Samuelson, L. Towards an integration of connectionist learning and dynamical systems processing: case studies in speech and lexical development. In: Spencer, J.; Thomas, M.; McClelland, JL., editors. Towards an integration of connectionist learning and dynamical systems processing: case studies in speech and lexical development. Oxford University Press; London: 2009.

McMurray B, Samelson VS, Lee SH, Tomblin JB. Eye-movements reveal the time-course of online spoken word recognition language impaired and normal adolescents. Cognitive Psychology. 2010; 60(1):1–39. [PubMed: 19836014]

McMurray B, Spivey MJ. The categorical perception of consonants: the interaction of learning and processing. Proceedings of the Chicago Linguistics Society. 2000; 34(2):205–220.

McMurray, B.; Zhao, L.; Kucker, S.; Samuelson, LK. Probing the limits of associative learning: generalization and the statistics of words and referents. In: Gogate, L.; Hollich, G., editors. Theoretical and Computational Models of Word Learning. (in press)

McRae K, Spivey-Knowlton MJ, Tanenhaus MK. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. Journal of Memory and Language. 1998; 38:283–312.

Medina TN, Snedeker J, Trueswell John C. Gleitman LR. How words can and cannot be learned by observation. Proceedings of the National Academy of Sciences. 2011; 108(22):9014–9019.

Merriman, W. Competition, attention, and young children's lexical processing. In: MacWhinney, B., editor. The Emergence of Language. Lawrence Erlbaum Associates; Mahwah, NJ: 1999. p. 351-358.

Merriman W, Lipko A, Evey J. How young children judge whether a word is one they know: A dual criterion account. Journal of Experimental Child Psychology. 2008; 101(2):83–98. [PubMed: 18675430]

Mervis CB, Bertrand J. Acquisition of the novel name/nameless category (N3C) principle. Child Development. 1994; 65:1646–1662. [PubMed: 7859547]

Mitchell, CC.; McMurray, B. A stochastic model of the vocabulary explosion. In: Love, BC.; McRae, K.; Sloutsky, VM., editors. Proceedings of the 30th Annual Conference of the Cognitive Science Society; Austin, TX: The Cognitive Science Society; 2008. p. 1919-1926.

Mitchell CC, McMurray B. On leveraged learning in lexical acquisition and its relationship to acceleration. Cognitive Science. 2009; 33(8):1503–1523. [PubMed: 21585513]

Moore, C. Representing Intentional Relations and Acting Intentionally in Infancy: Current Insights and Open Questions. In: Knoblich, G.; Thornton, IM.; Grosjean, M.; Shiffrar, M., editors. Human body perception from the inside out: Advances in visual cognition. Oxford University Press; New York, NY: 2006. p. 427-442.

Moore C. The Development of Gaze Following. Child Development Perspectives. 2008; 2(2):66–70.

Munakata Y. Infant perseveration and implications for object permanence theories: A PDP model of the A-not-B task. Developmental Science. 1998; 1(2):161–211.

Munakata Y, McClelland JL. Connectionist Models of Development. Developmental Science. 2003; 6(4):413–429.

Munakata Y, McClelland JL, Johnson M, Siegler R. Rethinking Infant Knowledge: Toward an Adaptive Process Account of Successes and Failures in Object Permanence Tasks. Psychological Review. 1997; 104(4):686–713. [PubMed: 9337629]

Namy L. Getting Specific: Early General Mechanisms Give Rise to Domain-Specific Expertise in Word Learning. Language Learning and Development. 2012; 8(1):57–60.

Nazzi T, Bertoncini J. Before and after the vocabulary spurt: two modes of word acquisition? Developmental Science. 2003; 6(2):136–142.

Newell, A.; Rosenbloom, PS. Mechanisms of skill acquisition and law of practice. In: Anderson, JR., editor. Cognitive skills and their acquisition. Erlbaum; Hillsdale, NJ: 1981. p. 1-55.

O'Hanlon CG, Roberson D. What constrains children's learning of novel shape terms? Journal of Experimental Child Psychology. 2007; 97(2):138–148. doi: 10.1016/j.jecp.2006.12.002. [PubMed: 17292387]

Palmeri TJ. Exemplar similarity and the development of automaticity. Journal of Experimental Psychology: Learning, Memory & Cognition. 1997; 23:324–354.

Perry L, Samuelson LK. The Shape of the Vocabulary Predicts the Shape of the Bias. Frontiers in Psychology. 2011; 2:345. [PubMed: 22125547]

Peterson, C.; McCabe, A., editors. Developing narrative structure. Lawrence Erlbaum Associates; Hillsdale, NJ: 1991.

Quine, WVO. Word and object: An inquiry into the linguistic mechanisms of objective reference. The MIT Press; Cambridge, MA: 1960.

Ratcliff R, Rouder JN. Modeling response times for two-choice decisions. Psychological Science. 1998; 9(5):347–356. doi: 10.1111/1467-9280.00067.

Reese E, Fivush R. Parental styles of talking about the past. Developmental Psychology. 1993; 29(3): 596–606.

Regier, T. The human semantic potential: Spatial language and constrained connectionism. The MIT Press; Cambridge, MA, US: 1996.

Regier T. Emergent constraints on word-learning: A computational review. Trends in Cognitive Sciences. 2003; 7:263–268. [PubMed: 12804693]

Regier T. The emergence of words: Attentional learning in form and meaning. Cognitive Science. 2005; 29(6):819–865. [PubMed: 21702796]

Reynolds JR, Zacks JM, Braver TS. A computational model of event segmentation from perceptual prediction. Cognitive Science. 2007; 31:613–643. [PubMed: 21635310]

Reznick JS, Goldfield BA. Rapid change in lexical development in comprehension and production. Developmental Psychology. 1992; 28:406–413.

Riches N, Tomasello M, Conti-Ramsden G. Verb learning in children with SLI: Frequency and spacing effects. Journal of Speech, Language, and Hearing Research. 2005; 48:1397–1411.

Robinson CW, Sloutsky VM. Auditory Dominance and its change in the course of development. Child Development. 2004; 75(5):1387–1401. [PubMed: 15369521]

Robinson CW, Sloutsky VM. Visual processing speed: Effects of auditory input on visual processing. Developmental Science. 2007; 10:734–740. [PubMed: 17973789]

Robinson SR, Adolph K, Young JW, Gill-Alvarez F. What Is the Shape of Developmental Change? Psychological Review. 2008; 115(3):527–543. [PubMed: 18729590]

Rosch E, Mervis CB, Gray WD, Johnson DM, Boyes-Braem P. Basic objects in natural categories. Cognitive Psychology. 1976; 8(3):382–439. doi: 10.1016/0010-0285(76)90013-x.

Rost GC, McMurray B. Finding the signal by adding noise: The role of non-contrastive phonetic variability in early word learning. Infancy. 2010; 15(6):608.

Roy D, Pentland A. Learning words from sights and sounds: a computational model. Cognitive Science. 2002; 26:113–146.

Rumelhart, D.; Zipser, D. Feature discovery by competitive learning. In: Rumelhart, D.; McClelland, JL., editors. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. The MIT Press; Cambridge, MA: 1986. p. 151-193.

Samuelson LK. Statistical Regularities in Vocabulary Guide Language Acquisition in Connectionist Models and 15-20-Month-Olds. Developmental Psychology. 2002; 38:1016–1037. [PubMed: 12428712]

Samuelson LK, Smith LB. Memory and attention make smart word learning: An alternative account of Akhtar, Carpenter and Tomasello. Child Development. 1998; 1:94–104. [PubMed: 9499560]

Samuelson LK, Smith LB. Early noun vocabularies: do ontology, category structure and syntax correspond? Cognition. 1999; 73(1):1–33. doi: 10.1016/s0010-0277(99)00034-7. [PubMed: 10536222]

Samuelson LK, Smith LB. Grounding development in cognitive processes. Child Development. 2000; 71:98–106. [PubMed: 10836563]

Samuelson LK, Smith LB, Perry LK, Spencer JP. Grounding Word Learning in Space. PLoS ONE. 2011; 6(12)

Sandhofer C, Smith LB, Luo J. Counting nouns and verbs in the input: Differential frequencies, different kinds of learning? Journal of Child Language. 2000; 27:561–585. [PubMed: 11089339]

Schlesinger M, McMurray B. Modeling Matters: What Computational Models Have Taught Us about Cognitive Development. Journal of Cognition and Development. (in press).

Schutte AR, Spencer JP, Schöner G. Testing the dynamic field theory: Working memory for locations becomes more spatially precise over development. Child Development. 2003; 74:1393–1417. [PubMed: 14552405]

Scott RM, Fisher C. 2.5-year-olds use cross-situational consistency to learn verbs under referential uncertainty. Cognition. (in press).

Sénéchal M, Cornell EH. Vocabulary acquisition through shared reading experiences. Reading Research Quarterly. 1993; 28(4):360–374.

Shanks DR. Associationism and cognition: Human contingency learning at 25. The Quarterly Journal of Experimental Psychology. 2007; 60(3):291–309. doi: 10.1080/17470210601000581. [PubMed: 17366302]

Sibley D, Kello C, David P, Elman JL. Large-scale modeling of wordform learning and representation. Cognitive Science. 2008; 32(4):741–754. [PubMed: 20107621]

Siskind JM. A computational study of cross-situational techniques for learning word-to-meaning mappings. Cognition. 1996; 61(1-2):39–91. [PubMed: 8990968]

Smith, LB. Avoiding associations when it's Behaviorism you really hate. In: Golinkoff, RM.; Hirsh-Pasek, K.; Bloom, L.; Smith, LB.; Woodward, AL.; Akhtar, N.; Tomasello, M.; Hollich, G., editors. Becoming a Word Learner: A Debate on Lexical Acquisition. Oxford University Press; New York, NY: 2000. p. 169-174.

Smith LB, Thelen E, Titzer R, McLin D. Knowing in the context of acting: The task dynamics of the A-not-B error. Psychological Review. 1999; 106(2):235–260. [PubMed: 10378013]

Smith LB, Yu C. Infants rapidly learn word-referent mappings via cross-situational statistics. Cognition. 2008; 106:1558–1158. [PubMed: 17692305]

Spencer, J.; Perone, S.; Johnson, JS. Dynamic Field Theory and Embodied Cognitive Dynamics. In: Spencer, J.; Thomas, M.; McClelland, JL., editors. Towards an integration of connectionist learning and dynamical systems processing: case studies in speech and lexical development. Oxford University Press; London: 2009.

Spiegel C, Halberda J. Rapid fast-mapping abilities in 2-year-olds. Journal of Experimental Child Psychology. 2011; 109:132–140. [PubMed: 21145067]

Spivey-Knowlton, MJ.; Allopenna, P. A computational account of the integration of linguistic and visual information during spoken word recognition Proceedings of the Computational Psycholinguistics Conference; Berkeley, CA. 1997.

Spivey, MJ. The continuity of mind. Oxford University Press; New York: 2007.

Spivey, MJ.; Dale, R. On the continuity of mind: Toward a dynamical account of cognition. In: Ross, BH., editor. The psychology of learning and motivation: Advances in research and theory. Vol. Vol. 45. Elsevier; San Diego, CA: 2004. p. 87-142.

Steyvers MM, Kenneth J. The Effect of Normative Context Variability on Recognition Memory. Journal of Experimental Psychology: Learning Memory and Cognition. 2003; 29(5):760–766.

Storkel H, Armbruster J, Hogan T. Differentiating Phonotactic Probability and Neighborhood Density in Adult Word Learning. Journal of Speech Language and Hearing Research. 2006; 49:1175–1192.

Swingley D. Onsets and codas in 1.5-year-olds' word recognition. Journal of Memory and Language. 2009; 60:252–269. [PubMed: 20126290]

Swingley D, Aslin RN. Spoken word recognition and lexical representations in very young children. Cognition. 2000; 76:146–166.

Swingley D, Aslin RN. Lexical Neighborhoods and the word-form representations of 14-month-olds. Psychological Science. 2002; 13(5):480–484. [PubMed: 12219818]

Swingley D, Aslin RN. Lexical competition in young children's word learning. Cognitive Psychology. 2007; 54(2):99–132. [PubMed: 17054932]

Tanenhaus, MK.; Brown-Schmidt, S. Language processing in the natural world. In: Moore, BCM.; Tyler, LK.; Marslen-Wilson, WD., editors. The perception of speech: from sound to meaning. 2008. p. 1105-1122.

Tanenhaus, MK.; Trueswell, JC. Sentence Comprehension. In: Eimas, PD.; Miller, JL., editors. Handbook in perception and Cognition. Speech, language and communication. Vol. Vol. 11. Academic Press; New York, NY: 1995. p. 217-262.

Thelen E, Schöner G, Scheier C, Smith L. The dynamics of embodiment: A dynamic field theory of infant perseverative reaching errors. Behavioral and Brain Sciences. 2001; 24:1–86. [PubMed: 11515285]

Tomasello, M. Perceiving intentions and learning words in the second year of life Language development: The essential readings. Blackwell Publishing; Malden: 2001. p. 111-128.

Tomasello M, Strosberg R, Akhtar N. Eighteen-month-old children learn words in non-ostensive contexts. Journal of Child Language. 1996; 23:157–176. [PubMed: 8733565]

Usher M, McClelland JL. The time course of perceptual choice: The leaky, competing accumulator model. Psychological Review. 2001; 108(3):550–592. doi: 10.1037/0033-295x.108.3.550. [PubMed: 11488378]

van Geert P. A dynamic systems model of cognitive and language growth. Psychological Review. 1991; 98:3–53.

Vecera S, O'Reilly R. Figure-ground organization and object recognition processes: An interactive account. Journal of Experimental Psychology: Human Perception and Performance. 1998; 24(2):441–462. [PubMed: 9554093]

Waxman, SR. Links between object categorization and naming: Origins and emergence in human infants Early category and concept development: Making sense of the blooming, buzzing confusion. The Oxford University Press; New York, NY: 2003. p. 213-241.

Waxman SR, Booth AE. Principles that are invoked in the acquisition of words, but not facts. Cognition. 2000; 77(2):B33–B43. [PubMed: 10986366]

Waxman SR, Gelman S. Early word-learning entails reference, not merely associations. Trends in Cognitive Sciences. 2009; 13(6):258–263. [PubMed: 19447670]

Wifall, T.; McMurray, B.; Hazeltine, RE. Similarity impairs motor learning: Implications for the power law of learning?. (submitted)

Wilkinson KM, Mazzitelli K. The effect of 'missing' information on children's retention of fast-mapped labels. Journal of Child Language. 2003; 30(1):47–73. [PubMed: 12718293]

Woodward, AL.; Markman, EM. Early word learning. In: Damon, W., editor. Handbook of child psychology: Volume 2: Cognition, perception, and language. John Wiley & Sons; Hoboken, NJ: 1998. p. 371-420.

Xu F, Tenenbaum JB. Word learning as Bayesian inference. Psychological Review. 2007; 114(2):245–272. [PubMed: 17500627]

Yu C, H. BD, Aslin RN. The role of embodied intention in early lexical acquisition. Cognitive Science. 2005; 29:961–1005. [PubMed: 21702799]

Yu C, Smith LB. Rapid Word Learning Under Uncertainty via Cross-Situational Statistics. Psychological Science. 2007; 18(5):414–420. [PubMed: 17576281]

Yu C, Smith LB. Modeling cross-situational word–referent learning: Prior questions. Psychological Review. 2012; 119(1):21–39. doi: 10.1037/a0026182. [PubMed: 22229490]

Zangl R, Klarman L, Thal D, Fernald A, Bates E. Dynamics of Word Comprehension in Infancy: Developments in Timing, Accuracy, and Resistance to Acoustic Degradation. Journal of Cognition and Development. 2005; 6(2):179–208. [PubMed: 22072948]
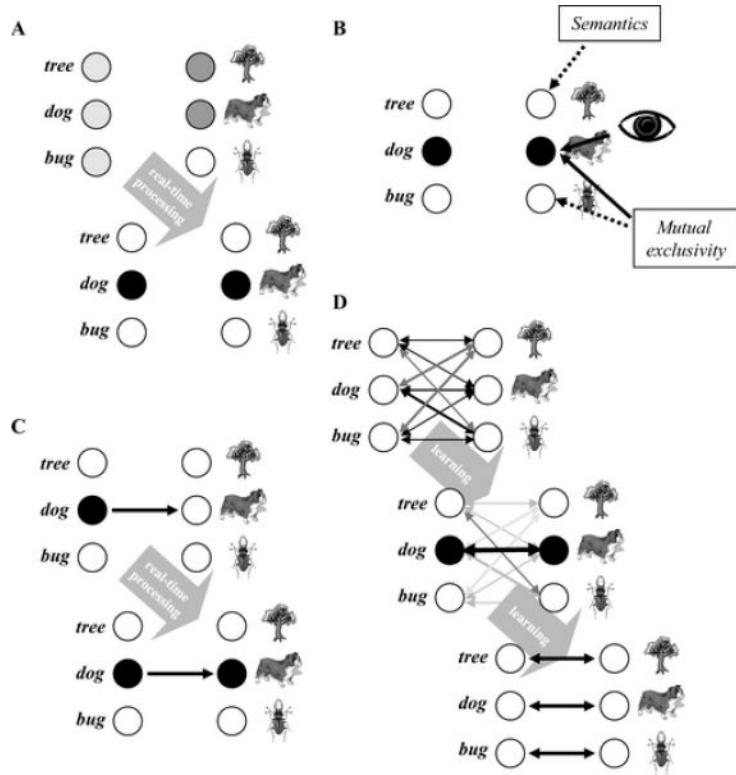
Figure 1.
A schematic of word learning. On the left side of each panel are units showing how strongly a particular auditory word form is under consideration; on the right are units showing the strength of a visual object under consideration. A) Solving the problems of word recognition and referential ambiguity require a transition from a state in which multiple word-forms and object categories are considered to a state in which only one is. B) Constraints like pragmatics and mutual exclusivity can act by simply changing the degree of consideration without affecting long term linkages. C) Familiar word recognition takes advantage of learned associations to activate object representations from spoken words. D) Learning is instantiated as long-term linkages between words and objects which are strengthened when both word-forms and objects are considered simultaneously.
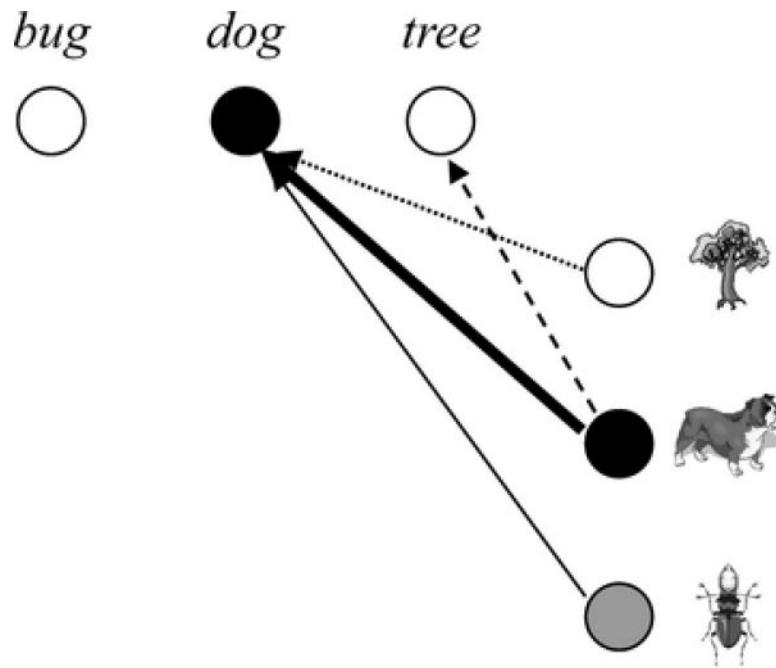
Figure 2.
A schematic of the sorts of connections that would need to be acquired or pruned during learning
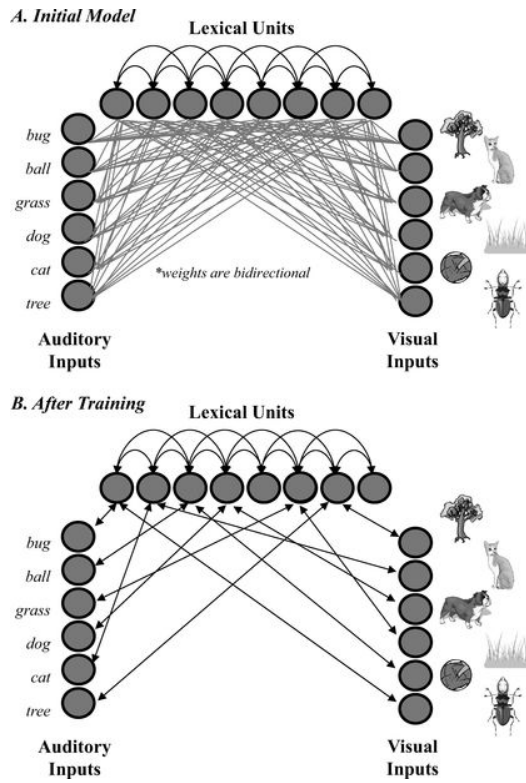
Figure 3.
Architecture of the model, both before (Panel A) and after (Panel B) learning.
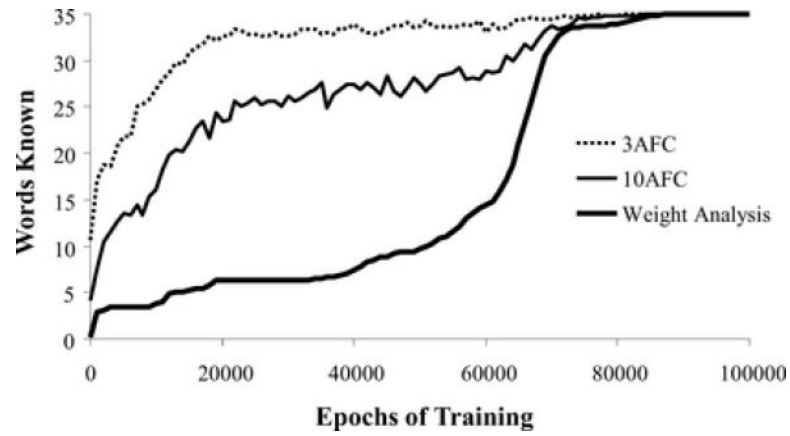
Figure 4.
Number of words known over the course of training as measured by 3AFC and 10AFC tasks as well as an analysis of the weights. B) Same data more finely sampled over the first 50,000 trials.
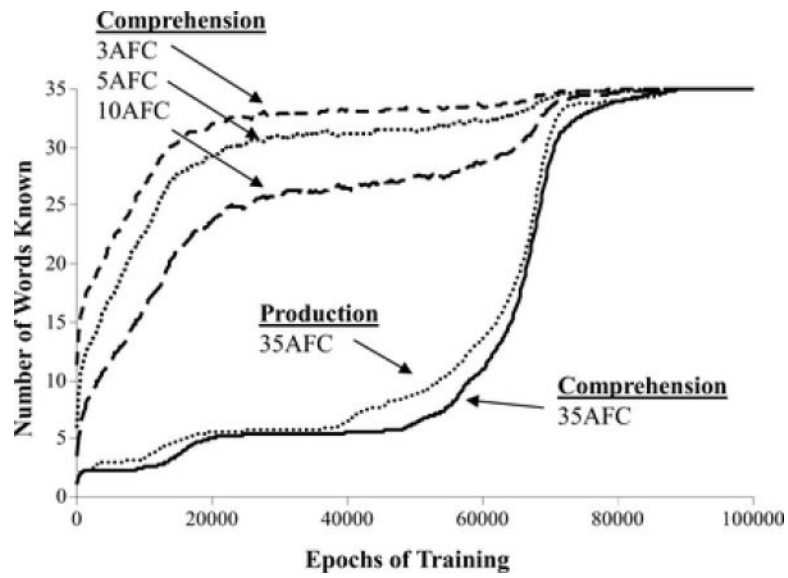
Figure 5.
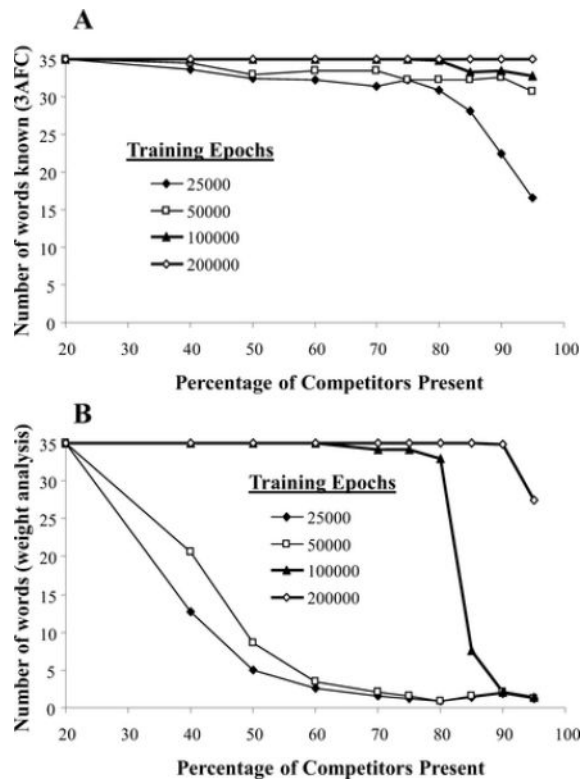Number of words known by several comprehension and production measures over the course of training.

Figure 6.
Number of words known as a function of the referential ambiguity and time. A) Measured in a 3AFC task. B) Measured via weight analysis.
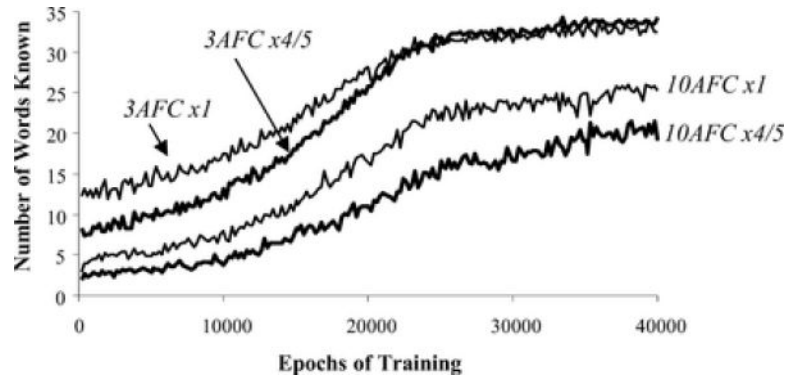
Figure 7.
Number of words known as a function of training when all words were equal frequency.
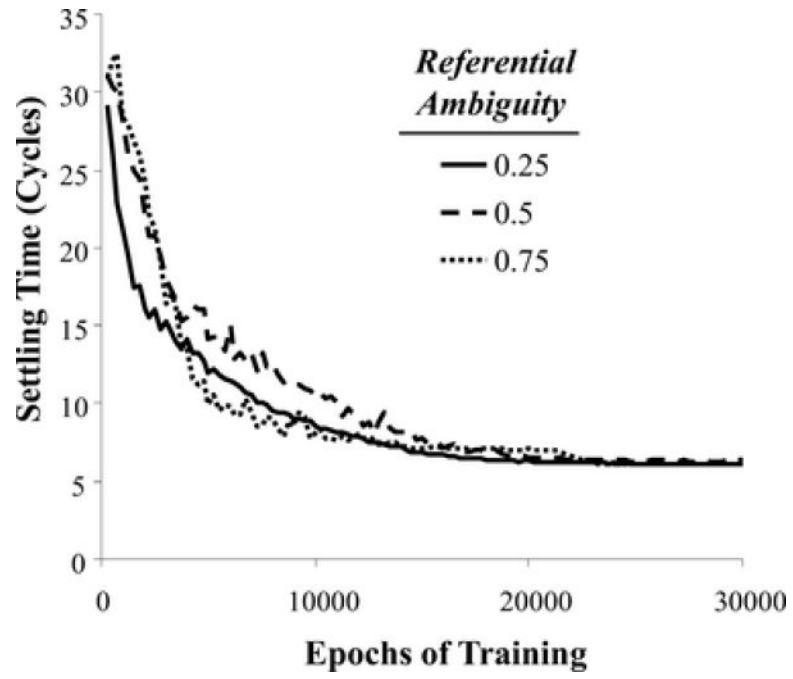
Figure 8.
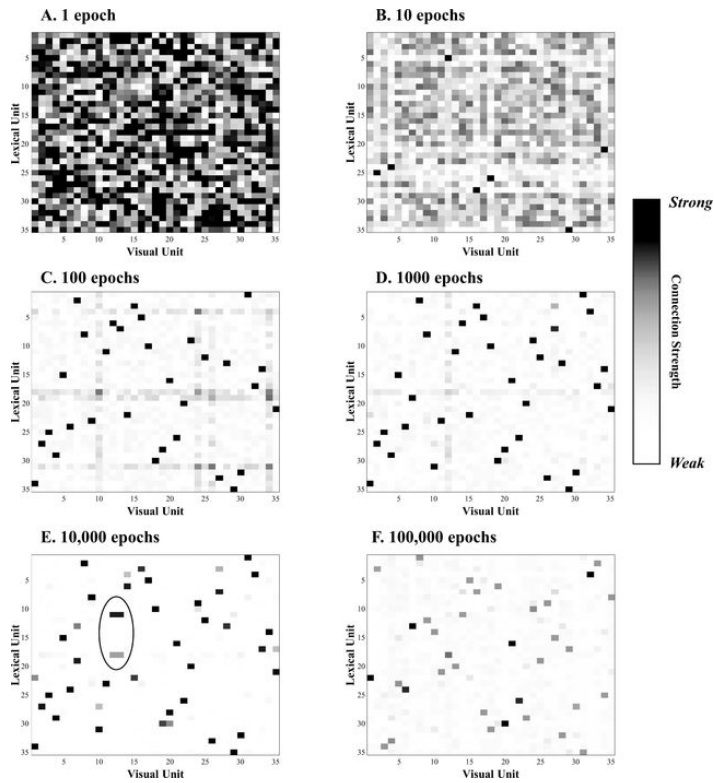Settling time over the course of training, as a function of referential ambiguity.

Figure 9.
A visual representation of the visual→lexical weight matrix over the course of development. On the X-axis is each of the 35 objects, on the Y-axis is a subset of the lexical units (including all of the ones that are ultimately "used". The connection between them is represented by the darkness of the patch.
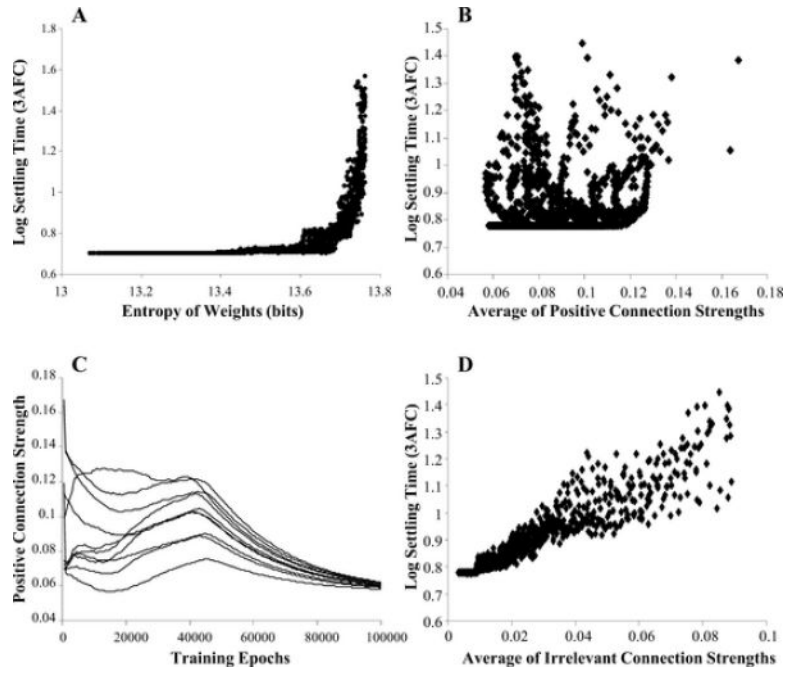
Figure 10.
Four measures of the weight matrix related to settling time. A) RT as a function of
Shannon's entropy. B) RT as a function of the average of the positive (correct) connections
linking each word/object to its correct lexical unit. C) The strength of the largest positive
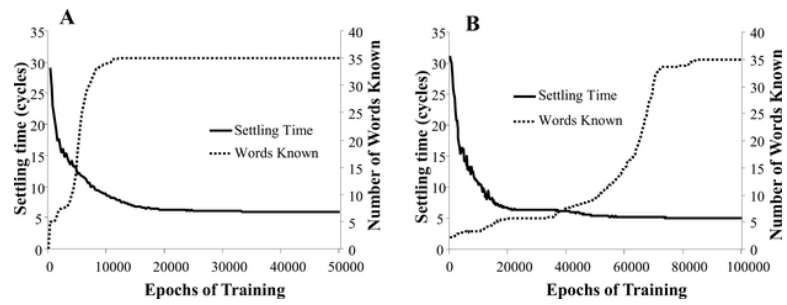connection over training. D) RT as a function of the average of the unused connections.

Figure 11.
A) Settling time as a function of training on a log-log scale. B) Settling time (left axis) and Words Known (right axis) for model trained under 25% ambiguity. C) Same data points for one trained under 75 % ambiguity.
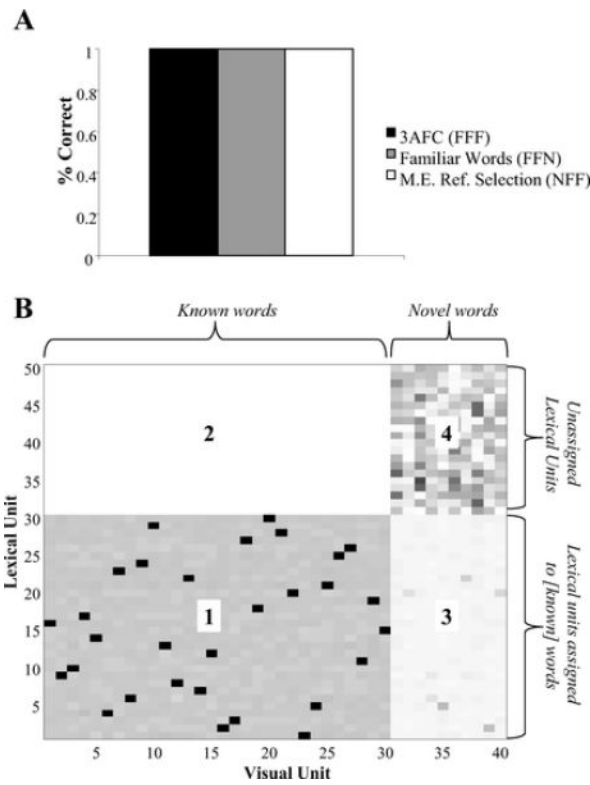
Figure 12.
A) Performance on the 3AFC, familiar word, and M.E. referent selection trials. B) Representative weight matrix. Darker patches indicate stronger connections.
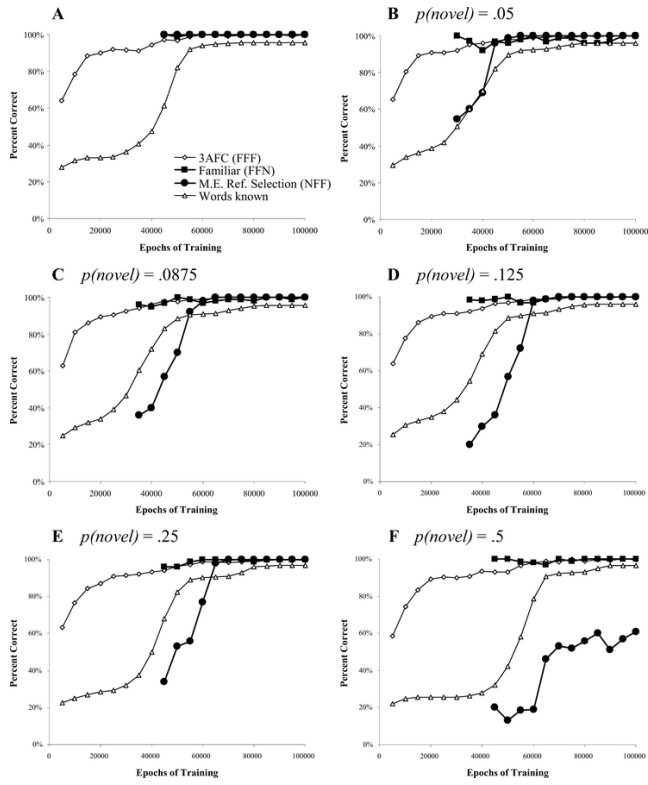
Figure 13.
Performance on familiar and M.E. referent selection trials over development (filled symbols). Also shown are the # of words known (percentage of 35) measured by both the 3AFC and weight analysis (open symbols). Panels represent different likelihoods of the novel words appearing as visual foils.
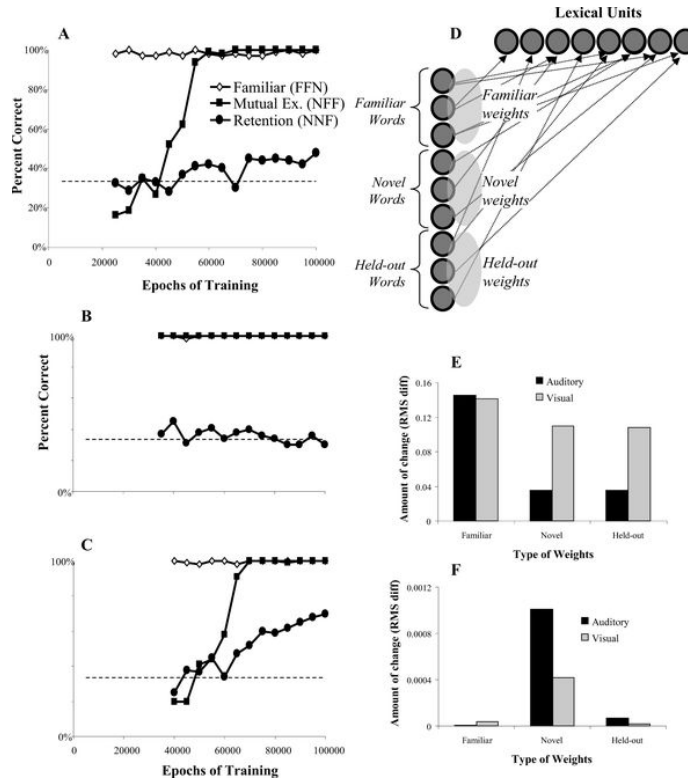
Figure 14.
Performance on mutual exclusivity and retention. A) Performance on familiar word, M.E. referent selection, and retention trials when novel items appeared with a likelihood .0875 B) Same, but novel items appeared with p=0. C) p=.375. C) Diagram of the weight matrix indicating which components were measured. E) RMS change in each type of weights during learning. F) RMS change during the five mutual exclusivity trials.
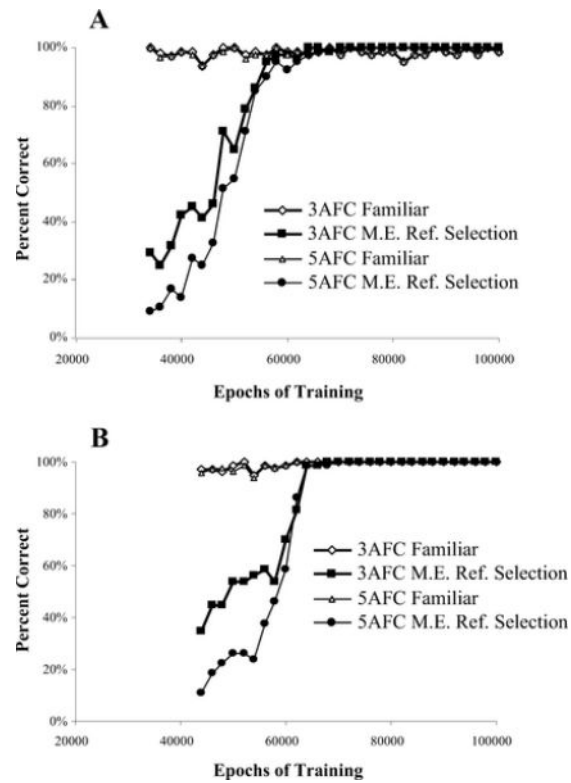
Figure 15.
Performance on 3- and 5AFC familiar and fast mapping tasks over training. A) Novel objects appeared with a likelihood of p=.0875 during training. B) p=.025.
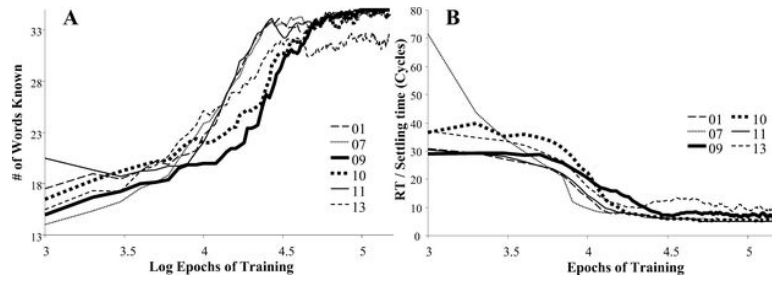
Figure 16.
A representative sampling of individual models. A) Number of words known. B) Settling time.
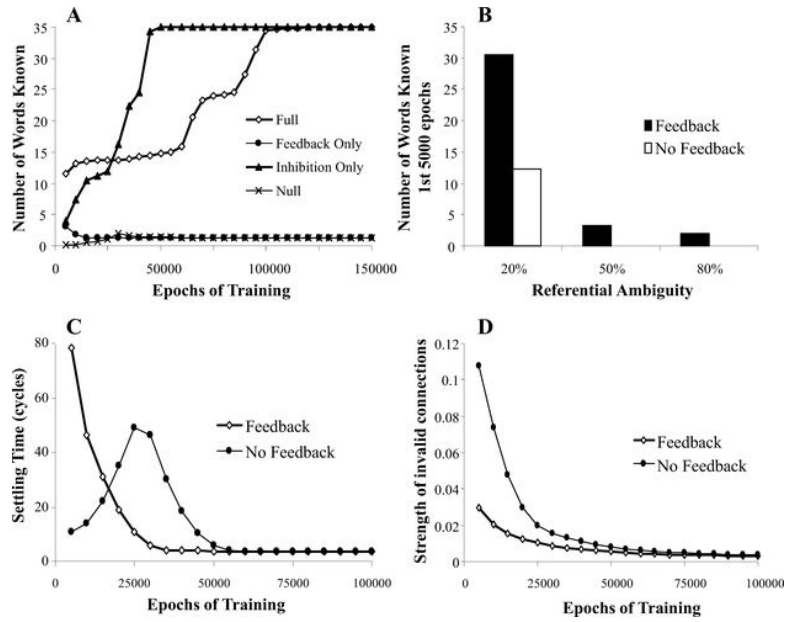
Figure 17.
Model performance in Simulation 3.4. A) Number of words known (via weight analysis) over development for the full model, and those lacking lexical inhibition or feedback. B) Number of words known in first 5000 epochs. C) Average settling time on a single word as a function lexical feedback and training. D) Strength of invalid connections over development for models with and without feedback. E) Same data as panel A further broken down by the presence of inhibition in input layers.
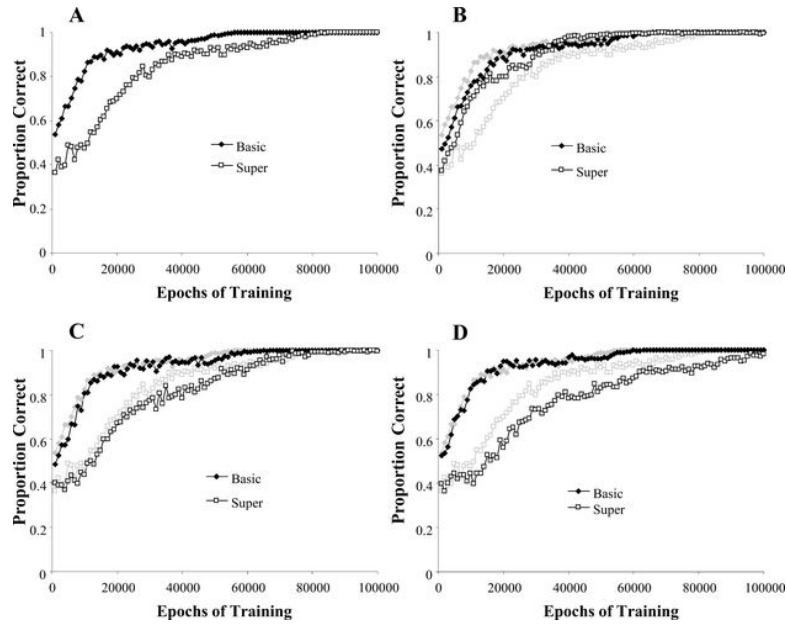
Figure 18.
Performance on basic-level and superordinate categorization over development. In all panels, grey curves are the same data as panel A. A) In model with 25 basic-level terms and 5 basic-level categories. B) In the same model in which the probability of a super-ordinate term was increased to 0.5. C) Performance in a model with 27 basic-level and 3 super-ordinate terms. D) Performance in that same model when the likelihood of any individual super-ordinate term was the same as basic-level terms (1/30).
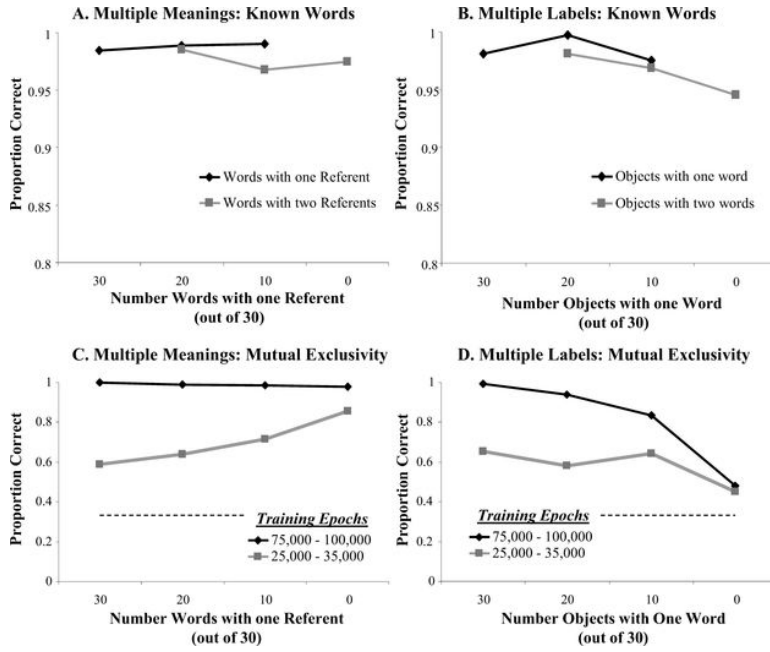
Figure 19.
A) Learning performance of the multiple-meanings model on a 10AFC task during the last 25,000 training trials as a function of the number of words with a single referent, and the type of word. B) Final learning performance in the multiple-labels model as a function of the number of objects with one word, and the type of object. C) Performance on a 3AFC fast-mapping task as a function of number of one-referent words in the multiple-meanings model. Chance is represented by the dashed line. D) Fast-mapping performance for the multiple-labels models.

Table 1

Summary of findings from the simulations (with reference to relevant empirical studies).

Simulation			Findings
Learning	1.1	Learning	• Model can learn under referential uncertainty. • Performance on comprehension exceeds competence due to task constraints. • "Slow-mapping" or elaboration effects without semantics. • Relevant studies: Yu & Smith (2007), Smith & Yu (2008), Capone & McGregor (2005)
	1.2	Production / comprehension	• Words can be comprehended earlier than produced. • Largely effect of task/competition environment. • Relevant studies: Huttenlocher (1974), Reznick & Goldfield (1992)
	1.3	Referential ambiguity	• Model can learn complete lexicon under high referential ambiguity. • Relevant studies: Yu & Smith (2007), Smith & Yu (2008)
	1.4	Accelerating learning	• Model shows acceleration, as long as learning task is difficult enough and learning is sampled frequently. • Relevant studies: McMurray (2007), Ganger & Brent (2004)
Situation-time processes	2.1	Familiar word recognition	• Settling time (RT) decreases over development. • Effect arises pruning connections between words and incorrect referents. • Similar pattern to power law of learning. • Acceleration in number of words unrelated to deceleration in RT. • Relevant Studies: Fernald et al (1998, 2006)
	2.2	Fast-mapping	• Model can fast-map by mutual exclusivity. • Function of both online dynamics and the weights set up by the learning/weight decay rule. • Relevant studies: Carey & Bartlett (1978) Markman & Wachtel (1988), Mervis & Bertrand (1994), Horst & Samuelson (2008)
	2.3	Fast-mapping & development	• Fast-mapping emerges out of changes in weight matrix. • Familiarity with visual objects can speed development.
Learning-Interactions	3.1	Fast-mapping & Retention	• Model fails to retain fast-mapped labels, unless visually familiar. • Only a small amount of learning occurs on any fast-mapping event. • Relevant studies: Horst & Samuelson (2008); Kucker & Samuelson (2010); Spiegel & Halberda (2011)
	3.2	Fast-mapping & Task	• Model succeeds at 3AFC fast-mapping task at earlier points than 5AFC. • Both develop over time. • Relevant studies: Mervis & Bertrand (1994), Markman et al (2003)
	3.3	Familiar words and individual differences	• Model shows stability in RT, correlations between RT and knowledge. • RT at early points in development predicts acceleration in vocabulary growth.

Simulation			Findings
			• Speed of processing is not unitary – emerges out of interactions between processing parameters, learning parameters and developmental history. • Relevant studies: Fernald et al (2006)
	3.4	Processing & Learning	• Competition is required for cross-situational learning. • Feedback slows learning but may be beneficial. • Relevant studies: Yu & Smith (2007), Smith & Yu (2008)
Scaling up	4.1	Superordinate categories	• Model can learn superordinate terms for objects in addition to basic-level. • Basic-level advantage derives from frequency, spreading of associations. • Mutual exclusivity does not block learning of second names because it is an online process, not a constraint on learning.
	4.2	One-to-One word object mappings	• Model can learn when all words have multiple meanings (e.g., polysemy). • Model can learn when all objects have multiple labels (e.g., bilingualism). • Fastmapping performance is slightly reduced by polysemy. • Fastmapping is significantly degraded when all objects have multiple labels. • Relevant studies: Byers-Heinlein & Werker (2009)
Supplement	S1	Acceleration and Word Difficulty	• Acceleration observed whenever the overall difficulty of the words are high. • Varying frequency results in longer period of apparently slow learning.
	S2	Temperature and speed of processing	• Higher temperatures appear to lead to slower processing for familiar words. • Effect derives from learning – higher temperatures offer initially faster settling and as a result, fewer weights are pruned. • Higher temperature also slows learning. • Initially faster settling causes system to commit to more erroneous interpretations.
	S3	Larger Lexica	Model can learn lexica of up to 150 words at high degrees of referential ambiguity (50% or M=75 competitors / trial)
	S4	Slow learning	• Manipulated learning rates to see effect on learning. • At normal values for typical Hebbian learning (>.01) the model fails to learn, but can learn at intermediate and low values. • Slow learning prevents model from over-committing to erroneous mappings.

Table 2

A number of parameters control the activation flow, rate of learning, and initial conditions of the model. Initial weights refers to the range of values that the connection weights were randomly initialized to (e.g., a random number between 0 and .5). Learning rate affects the amount of weight change for each leaning instance.

	Simulation			
Parameter	**1.1 Comprehension**	**1.2 Comp / Prod**	**1.3 Ambiguity**	**1.4 Acceleration**
Input Units	35	35	35	35
Lexical Units	500	500	500	500
Initial Weight Size	.5	.5	.5	.5
Learning Rate	.0005	.0005	.0005	.0005
Referential Ambiguity	.5	.5	.2 - .95	.75
Feedforward Temperature	.01	.01	.01	.01
Feedback Temperature	2	2	2	2
Stability Point	1e-12	1e-12	1e-12	1e-12
Input inhibition	1.05	1.05	1.05	1.05

Table 3

Parameters for simulations in Section 2

| Parameter | Simulation | | |
	2.1 Word Recognition	2.2 Fastmapping	2.3 Fastmapping Development
Input Units	35	40	40
Familiar Words	35	30	30
Lexical Units	500	500	500
Initial Weight Size	.5	.25	.25
Learning Rate	.0005	.0005	.0005
Referential Ambiguity	.25, .5, .75	.5	.5
Feedforward Temperature	.01	.01	.01
Feedback Temperature	2	2	2
Stability Point	1e-12	1e-12	1e-12
Input inhibition	1.05	1.05	1.05

Table 4

Parameters for simulations in Section 3.

	Simulation			
Parameter	**3.1** **F/M Retention**	**3.2** **F/M Task**	**3.3** **Longitudinal** **Word Rec.**	**3.4** **Learning /** **Competition**
Input Units	40	40	35	35
Familiar Words/Objects	30	30	35	35
Novel Words/Objects	5	10	0	0
Held Out Words/Objects	5	0	0	0
Lexical Units	500	500	500	500
Initial Weight Size	.25	.25	.5±.025	.5
Learning Rate	.0005	.0005	.0005±.00002	.0005
Referential Ambiguity	.5	.5	.65±.03	.2 .5 .8
Novel object seen	.0875	.0875, .25	n/a	n/a
Feedforward Temperature	.01	.01	.01±.0065	.01
Feedback Temperature	2	2	2±.1	2
Stability Point	1e-12	1e-12	$10^{-12\pm.25}$	1e-12
Input inhibition	1.05	1.05	1.05±.01	1.05

Table 5

Correlation of RT and accuracy (with themselves) across time-slices.

Epochs	RT	Accuracy (2AFC)
5,000 – 10,000	.77 **	.39 **
10,000 – 15,000	.83 **	.47 **
15,000 – 20,000	.90 **	.54 **
20,000 – 25,000	.94 **	.66 **
25,000 – 30,000	.94 **	.53 **
30,000 – 50,000	.96 **	.54 **
50,000 – 75,000	.92 **	.65 **
75,000 – 100,000	.95 **	.79 **

**
p<.01.

Table 6

Correlation between speed and accuracy, as well as the three components of the growth curve in overall words known (20AFC).

Timestep	Accuracy	Growth function		
		Quadratic	Linear	Intercept
5,000	.10	−.38[**]	.35[**]	−.16
10,000	.04	−.34[**]	.08	.10
15,000	−.30[**]	−.24[*]	−.07	.18
20,000	−.35[**]	−.25[*]	−.07	.16
25,000	−.48[**]	−.28[**]	−.05	.18
30,000	−.52[**]	−.28[**]	−.05	.15
50,000	−.64[**]	−.28[**]	−.06	.17
70,000	−.77[**]	−.29[**]	−.04	.16
100,000	−.85[**]	−.26[**]	−.06	.15

[**] $p < .01$;

[*] $p < .05$

Table 7

Relationship between control parameters and output measures in the network.

		Initial Weight Size	Learning rate	# Visual Competitors	Stability Point	Feed-forward Temperature	Feedback Temperature	Input Inhibition
RT	10,000	−.12	−.15	.09	−.19[+]	.79[**]	−.03	−.35[**]
	25,000	−.09	−.09	.14	−.17[+]	.82[**]	−.04	−.30[**]
	100,000	−.08	−.10	.09	−.12	.85[**]	−.07	−.26[**]
Words Known	10,000	.16	.07	−.31[**]	−.20[*]	.04	−.09	−.12
	25,000	.22[*]	.12	−.33[**]	−.19[*]	−.31[**]	.02	.13
	100,000	.09	.13	−.09	.27[**]	−.87[**]	.09	.29[**]
2AFC accuracy	10,000	−.10	.21[*]	−.32[**]	−.27[**]	.01	.02	−.11
	25,000	0	.06	−.38[**]	−.14	−.39[**]	.20[+]	.17[+]
	100,000	.09	.19[+]	−.09	.14	−.77[**]	.03	.22[*]
Growth Curve	Quad.	.02	−.04	.01	.07	−.34[**]	.06	.24[*]
	Linear	.09	.11	−.15	−.13	.09	.02	−.14
	Interc.	−.01	.03	−.09	−.02	.05	−.11	0

[**] p<.01;

[*] p<.05;

[+] p<.1

Table 8

Parameters for simulations in Section 4

Parameter	Simulation	
	4.1 Multiple Labels	**4.2 Multiple labels / referents**
Visual Units	25	30 – 15
Auditory Units	30	30 – 15
Basic Level Words	25	N/A
Superordinates	5	0
Words referring to two objects	N/A	0, 5, 10, 15
Objects with two names	N/A	0, 5, 10, 15
Novel Visual/ Auditory Units	0	5
Lexical Units	500	500
Initial Weight Size	.5	.5
Learning Rate	.0005	.0005
Referential Ambig.	.5	.5
Feedforward Temp.	.01	.01
Feedback Temp.	2	2
Stability Point	1e-12	1-e-12
Input inhibition	1.05	1.05