Running head: Emergence of words

The emergence of words: Attentional learning in form and meaning

Terry Regier
University of Chicago

To appear in *Cognitive Science*.

Abstract

Children improve at word-learning during the second year of life – sometimes dramatically. This fact has suggested a change in mechanism, from associative learning to a more referential form of learning. This paper presents an associative exemplar-based model that accounts for the improvement without a change in mechanism. It provides a unified account of children's growing abilities to: (i) learn a new word given only one or a few training trials ("fast mapping"); (ii) acquire words that differ only slightly in phonological form; (iii) generalize word meanings preferentially along particular dimensions, such as object shape (the "shape bias"); and (iv) learn second labels for already-named objects, despite a persisting resistance to doing so ("mutual exclusivity"). The model explains these improvements in terms of increased attention to relevant aspects of form and meaning, which reduces memory interference. The interaction of associations and reference in word-learning is discussed.

Keywords: word-learning, fast mapping, phonological detail, shape bias, mutual exclusivity, associative learning, selective attention, exemplar-based, connectionist, memory interference, reference, vocabulary spurt, lexical development, language acquisition.

How do children learn words? One possibility is that they rely in part on simple associative learning (e.g. Pavlov, 1927), of the sort found in non-human animals. Children might link the sound "ball" to the idea of a ball using the same general principles that allow a dog to associate a bell with an upcoming meal.

This cannot be all there is to it. In learning words, children rely on many skills beyond association, including social awareness of reference (Baldwin, Markman, Bill, Desjardins, and Irwin, 1996), inference about events that are not present (Gleitman, 1990), knowledge of syntax (Fisher, Gleitman, & Gleitman, 1991), and pragmatic inference (Clark, 2004), among others (Bloom, 2000). Yet associative learning *can* account for some aspects of word-learning (e.g. Merriman, 1999; Plunkett, Sinha, Møller, & Strandsby, 1992; Smith, 2000). What is not yet clear is just how much of word-learning this simple idea can explain.

One influential proposal is that associative learning accounts for only the initial stages of word-learning, in which words are acquired fairly slowly. Then, on this view, sometime during the second year of life, the child has a conceptual insight into the symbolic, referential nature of words (Lock, 1980; McShane, 1979). This insight qualitatively changes the nature of the word-learning mechanism, allowing the child to learn words more quickly and effectively.

Several improvements in word-learning occur at around this age, and when viewed together, they do suggest a mechanistic change of some sort. I shall argue, however, that these changes can be accounted for without recourse to a referential insight during the second year, or any other qualitative change in mechanism. Instead, they can be explained in a unified fashion by an associative model which gradually learns to attend to communicatively relevant aspects of the world (Kruschke, 1992; Smith, 1989). These learned attentional shifts reduce memory interference, and improve word-learning. The present work expands on existing attentional proposals (e.g. Merriman, 1999; Smith, 2000; Smith, Jones, Landau, Gershkoff-Stowe, and Samuelson, 2002) in bringing together these parallel changes in the child's treatment of both form and meaning (Hespos & Spelke, 2004).

This proposal is compatible with the recent suggestion that children have an *earlier* insight – at around 12 months – into the social bases of reference and communication. That insight may initiate word-learning (Tomasello, 1999), and set in motion gradual attentional shifts. After several months, attention may have shifted enough to bring about the improved word-learning we see in children partway through the second year of life.

Changes in Word-Learning

During the second year, the child's word-learning behavior changes in at least four respects: ease of learning, honing of linguistic form, honing of linguistic meaning, and the learning of second labels.

Ease of learning

As children first begin to produce words, their acquisition of new words is slow and errorful. Between 12 and 16 months of age, children tend to learn new words at the rate of 2 or 3 per week (Fenson, Dale, Reznick, Bates, Thal, & Pethick, 1994; Gershkoff-Stowe & Smith, 1997). Later, around 18 to 22 months of age – often when the child has about 50 words in productive vocabulary – the acquisition of new words accelerates, sometimes sharply (a "vocabulary spurt", but see also Ganger & Brent, 2004, who show that the increase is usually more gradual than has been assumed). Many children at this age acquire 8 or 9 words per week (Fenson et al., 1994), and reports show some children learning as many as 44 new words in a week (Dromi, 1987).

Experimental studies in the laboratory provide further evidence for a shift from slow to quick word-learning. 13- to 15-month-olds can acquire a word-object linkage in comprehension based on 9-12 training trials (Schafer & Plunkett, 1998; Woodward, Markman, & Fitzsimmons, 1994). By the time children are 2 to 3 years of age, 1-3 training trials are sufficient (Behrend, Scofield & Kleinknecht, 2001; Carey, 1978). This rapid, efficient word-learning is known as "fast mapping". Children may retain their knowledge of such a newly-learned word for up to a month (Markson & Bloom, 1997).

The topic of fast mapping has attracted considerable attention (Behrend et al., 2001; Bloom, 2000; Bloom & Markson, 2001; Markson & Bloom, 1997; Waxman & Booth, 2000; see also Dickinson, 1984; Dollaghan, 1987; Heibeck & Markman, 1984). This interest stems in part from the idea that fast mapping may reflect a specialized mechanism for word-learning – part of an overall human predisposition for language. Another possibility, however, is that fast mapping may arise from more general processes of learning and memory. This is suggested by children's and adults' ability to similarly learn linguistically-presented *facts* about objects given only a few exposures, and to retain that information for up to a month (Markson & Bloom, 1997; although visually-presented information is not retained as well) – and by the finding that at least one dog is able to fast-map words (Kaminski, Call, & Fischer, 2004).

Honing of linguistic form

Some word forms differ only very slightly. For example, the pronunciations of the words "bit" and "pit" differ only in the initial consonant – yet they carry very different meanings. Such *minimal pairs* are widespread in human language, so children must eventually learn them if they are to communicate effectively. The ability to learn such pairs of words seems to emerge gradually. Young children, at 14 months of age, are able to link two phonologically *dis*similar sounds (e.g. "lif" and "neem") to different objects (Stager & Werker, 1997; Werker, Cohen, Lloyd, Casasola, & Stager, 1998). But these young children sometimes encounter difficulty linking two *similar* sounds (e.g. "bih" and "dih") to different objects (Stager & Werker, 1997; Werker, Fennell, Corcoran, & Stager, 2002; but see also Swingley & Aslin, 2002 for evidence that children this age nonetheless have fairly detailed phonological representations). In contrast, older children, at 18-23 months, reliably learn and discriminate similar-sounding names for objects (Werker et al., 2002; see also Bailey & Plunkett, 2002; Swingley & Aslin, 2000). The overall picture is one of a gradual "honing" of linguistic form: children move from being only somewhat sensitive to minor differences between word forms when linking them to objects, to being acutely sensitive to them. It is as if they gradually learn that some features of word form can signal a large difference in meaning, as in a minimal pair – and they then attend preferentially to these features.

Honing of meaning

As the child gradually determines which aspects of word form are relevant, an analogous process may take place for word *meaning*. Early in development, children are sometimes conservative in generalizing a newly learned object name to other referents. For example, in one of Woodward, Markman, and Fitzsimmons' (1994) studies, young (13-month-old) children failed to generalize a newly-learned word to another object of the same shape as the original, but a different color. In contrast, older English-speaking children systematically and robustly generalize novel names for artifact-like objects to other referents of the same shape as the original, and tend to ignore, for naming purposes, differences along other dimensions such as color and size (e.g. 17-month-olds: Smith et al., 2002; 2-3-year-olds: Landau, Smith, and Jones, 1998; Smith, 2000; but see also Booth

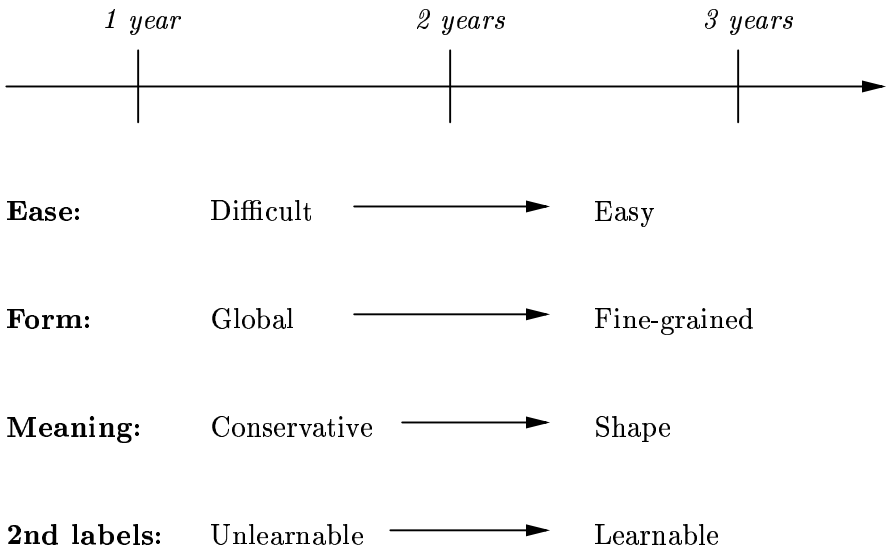


Figure 1: Four trends in early word-learning.

& Waxman, 2002). Smith et al. (2002) show that this "shape bias" is strengthened by learning words for object categories defined by shape, and that it assists the learning of more such words. Since artifacts are often named on the basis of shape in English, the shift toward a shape bias is analogous to children's treatment of word form: in both cases, there is a trend of increasing attention to communicatively relevant details.[1]

Second labels

Children tend to assume that if an object has one name, it will not have another. This bias against multiple labels is called the *mutual exclusivity* bias (Markman, 1989; Merriman & Bowman, 1989; cf. the principle of contrast, Clark, 1987). One manifestation of this bias is that young children sometimes fail at learning second labels for objects. Liittschwager and Markman (1994) found that 16-month-olds, who can learn a new word for an as-yet-unnamed object, have trouble learning a new word for an already-named object – that is, a second label. In contrast, they found that older children, 24-month-olds, are capable of learning both novel words (first labels) and second labels (see also Mervis, Golinkoff, & Bertrand, 1994). Older children nonethless continue to prefer mappings in which an object has only one label (Markman & Wachtel, 1988; Merriman & Bowman, 1989).

Overview

When these roughly simultaneous changes – in ease of learning, form-honing, meaning-honing, and second-label-learning – are viewed together, as in Figure 1, they suggest some sort of change in mechanism, sometime before the second birthday. This might at first seem to support the proposal that there is a referential insight at this age, taking the child from associative learning to another

[1] Children may actually attend to an artifact's *intended purpose*, rather than its shape per se, when generalizing a name for it (Diesendruck, Markson, & Bloom, 2003); object shape just happens to be a fairly good perceptual cue to this more conceptual content. Still, the general idea of learning to attend to a communicatively relevant aspect of a referent is unaffected by whether that aspect is shape or intended purpose – or other aspects in other languages, e.g. the substance of which an object is made, in Yucatec Mayan (Lucy, 1992).

more effective form of word-learning. However, I shall argue that these parallel developmental trajectories can all be accounted for through associative learning, without any qualitative change in mechanism. On this account, there is a single underlying reason for these four phenomena – but it is not a conceptual insight. These points are argued through the LEX ("Lexicon as EXemplars") computational model.

There are a number of existing models of word-learning (Cottrell & Plunkett, 1994; Elman, Bates, Johnson, Karmiloff-Smith, Parisi, & Plunkett, 1996; Farkas & Li, 2001; Gasser & Smith, 1998; Gupta & MacWhinney, 1997; Li & Farkas, 2002; MacWhinney, 1987; Merriman, 1999; Miikkulainen, 1997; Niyogi, 2002; Plaut, 1999; Plunkett, Sinha, Møller, & Strandsby, 1992; Regier, 1996; Roy & Pentland, 2002; Schafer & Mareschal, 2001; Siskind, 1992, 1996; Tenenbaum & Xu, 2000; Thompson & Mooney, 2003; Yu, Ballard, and Aslin, 2003). Several of these models account for one or more of the phenomena under consideration, e.g. fast mapping (Gupta & MacWhinney, 1987; Niyogi, 2002; Siskind, 1996; Tenenbaum & Xu, 2000; Thompson & Mooney, 2003), changes in sensitivity to phonological cues (Schafer & Mareschal, 2001), the difficulty of learning second labels (Cottrell & Plunkett, 1994; MacWhinney, 1987; Merriman, 1999, Siskind, 1996; Thompson & Mooney, 2003), and the shape bias (Merriman, 1999). However, the LEX model, which builds on this earlier work in word-learning, and related work in categorization, is to my knowledge the first computational model that provides a unified account of all four of the developmental phenomena outlined above. In doing so, it embodies the idea that learning words causes children to improve at learning yet more words (Smith, 2000; Smith et al., 2002; see also Brent 1996; Regier, 2003; Siskind, 1996). This dynamic affects both form and meaning, and may help to explain the accelerating development of the lexicon.

In what follows, I first review existing models of word-learning that are of relevance to this work. I then present the LEX model, initially in general conceptual terms, and then formally. I then demonstrate that the model accounts for the developmental trends above, and present predictions that it makes for future empirical test. Finally, I discuss the ramifications of this account.

Models of Word Learning

There are three general classes of word-learning models that are directly relevant to this work. These are subsymbolic models, competition models, and inferential models.

Subsymbolic models. Symbolic behavior may emerge from subsymbolic mechanisms. Connectionist models are frequently used to make this general point, by modeling language behavior without any explicit representation of linguistic categories or symbols (Elman, 1993; Plunkett & Marchman, 1991; Rohde & Plaut, 1999; Rumelhart & McClelland, 1986). A particularly relevant example is the associative word-learning model of Plunkett et al. (1992; see also Cottrell & Plunkett, 1994; Schafer & Mareschal, 2001). This multi-layer connectionist network was trained, under back-propagation (Rumelhart, Hinton, & Williams, 1986) and related gradient descent algorithms, to associate labels with images, without any intervening symbolic representations. Given a representation of both a label and its associated image as input, the network was trained to produce as output a copy of both the label and the image. In the course of training, the model exhibited a number of effects that characterize children's word-learning: word comprehension preceding production, generalization to prototypical exemplars of the meanings being learned, and some over- and under-generalization errors. Most significantly, for current purposes, the model exhibits a vocabulary spurt, without a change in mechanism. In addition, the subsymbolic model of Schafer and Mareschal (2001) accounts for a shift in children's sensitivity to phonological cues that is closely related to the shift

investigated here (Stager & Werker, 1997). However, neither model accounts for the specific cluster of improvements in word-learning treated here. This class of models (multilayer perceptrons) is also unfortunately susceptible to *catastrophic interference:* when learning two patterns sequentially, learning the second can largely eliminate the model's memory of the first (McCloskey & Cohen, 1989). Thus, this class of models fails to capture children's ability to remember a newly-learned word in the face of up to a month's subsequent exposure to other words (Markson & Bloom, 1997). On balance, these models offer important insights into the possible emergence of words from subsymbolic representations, but also leave other important issues unresolved.

Competition models. Several existing models of word-learning are built around the concept of lexical competition. A salient example is the competition model (Bates & MacWhinney, 1989; MacWhinney, 1987; MacWhinney, 1989). Given a featural representation of a referent – e.g. +round, +bounces, -animate for a toy ball – the competition model determines the activation in favor of each possible lexical choice – e.g. "ball", "box", "dog", etc. – by summing the associative strengths of the cues favoring that choice. In this case, the summed activation in support of "ball" will be high, while that supporting the other two choices will be low. At the core of the model, these lexical choices compete with each other, based on their activation levels. This is captured through normalization:

$$p_i = \frac{a_i}{\sum_j a_j + noise} \tag{1}$$

where a_i is the activation of word i, p_i is its probability of being produced, j indexes over all words, and *noise* denotes noise in the system, and prevents division by zero (MacWhinney, 1989).

Merriman (1999) has advanced a more elaborated competition model, CALLED (Competition, Attention, and Learned LExical Descriptions). Like MacWhinney's model, on which it builds, CALLED is in part conceptually rather than formally specified, but its formalized core is the same: Equation 1. Merriman (1999) shows that this formula accounts for a variety of mutual exclusivity effects – explaining the difficulty of learning multiple labels. CALLED also incorporates the idea of selective attention to particular dimensions of experience, such that some aspects of the referential world come to carry more weight than others, for the purposes of naming. Merriman (1999), like Linda Smith and colleagues (e.g. Smith, Jones, & Landau, 1996; Smith, 2000; Smith et al., 2002), argues that selective attention can account for the growth of the shape bias: if the dimension of shape receives increasing amounts of attention over development, it will become increasingly influential in naming decisions. However, CALLED contains no representation of phonological form, only of meaning – so it cannot account for apparent changes in phonological sensitivity. The model is also potentially vulnerable to interference from subsequent learning. Like subsymbolic models of word-learning, CALLED serves as an important starting-point for the present work, highlighting important principles but also leaving important questions open.

There is a family of more completely formalized models that capture the interaction of competition and attention. These are models of category learning in adults, rather than of word-learning in children. It is possible that these two sorts of phenomena, in populations of different ages, may be accounted for by the same set of general principles.

Several aspects of human categorization behavior are well-described by models that store specific *exemplars,* or instances, of categories (Nosofsky, 1986; see also Smith, 1989). An especially relevant example is Kruschke's (1992) model of category-learning, ALCOVE. In this model, each exemplar occupies a point position in a multi-dimensional psychological space, and is represented by an exemplar node encoding that position. A presented stimulus then activates each existing exemplar node to the extent that it is psychologically near that exemplar. Psychological nearness is modu-

lated by the amount of selective attention paid to particular dimensions of the input space. Thus, the underlying psychological space may be thought of as being "stretched" along highly attended dimensions - such that if the input differs from a stored exemplar along attended dimensions, it will be psychologically distant from that exemplar, and will activate it only very weakly. Analogously, the space may be thought of as "compressed" along unattended dimensions - such that if the input differs from a stored exemplar only along unattended dimensions, it will be psychologically close to the exemplar, and will activate it strongly. Exemplar nodes then project their activation along weighted associative connections to category nodes, and the choice of output category for the presented stimulus is determined by a competitive rule analogous to Equation 1. Training is through gradient descent in error, over both associative weights, and weights encoding selective attention. Thus, ALCOVE embodies, in a formalized associative model, the ideas of competition and selective attention to dimensions – in addition to the central idea of exemplar-based categorization. Significantly, this model, like humans, does not suffer from catastrophic forgetting: once it has learned that a particular exemplar is a member of some category, this learning is well-preserved despite subsequent training on other associations. This fact is relevant to word-learning since children can retain a new word's meaning despite much subsequent exposure to language.

For current purposes, the most significant point concerning this class of models is that they suggest that some general principles of category learning may also subserve word-learning.

Inferential models. There are several models of word-learning that are structured around *inference rules*. The model of Siskind (1996) is particularly relevant (see also Thompson & Mooney, 2003). Siskind's model accepts as input a series of multi-word utterances, each paired with a set of possible meanings for the utterance as a whole; each meaning is represented as a logical expression. The model's task is to determine the meaning of each word in the language. This is accomplished through a set of inference rules that embody general constraints on the problem, such as that the words of an utterance contribute non-overlapping portions of utterance meaning. This constraint, a version of the principle of contrast (Clark, 1987), allows the model to capture children's progression from multi-trial word-learning to one-trial learning. As the model learns more about the meanings of some words, that partial knowledge constrains the possible meanings of other words in an utterance – resulting in faster learning. The model is capable of learning a homonymous lexicon, from noisy input, in the presence of referential uncertainty.

Recently, several researchers have cast word-learning as Bayesian inference. Tenenbaum and Xu (2000) present a Bayesian model of fast mapping with appropriate generalization to other exemplars of the same kind. The core of their model is Bayesian combination of evidence, with an intuitively reasonable prior and likelihood. The prior was chosen such that if two clusters of referents were clearly distinct in extension, they were *a priori* more likely to have distinct names. The likelihood captured the idea that exemplars of word-referent pairings were sampled at random – such that after many positive instances of dogs being labeled "dog", the learner may be reasonably certain that cats are not also labeled "dog" – for if they were, such a labeling should have been encountered. This model accounts well for adults' generalization patterns in an artificial word-learning task (see also Niyogi, 2002 for a related application of the same general principles). One limitation of these models, and Siskind's, however, is that they do not include a representation of word form. This means that in these models, the similarity of two word forms cannot affect learning – although as we have seen, it does affect learning in children.

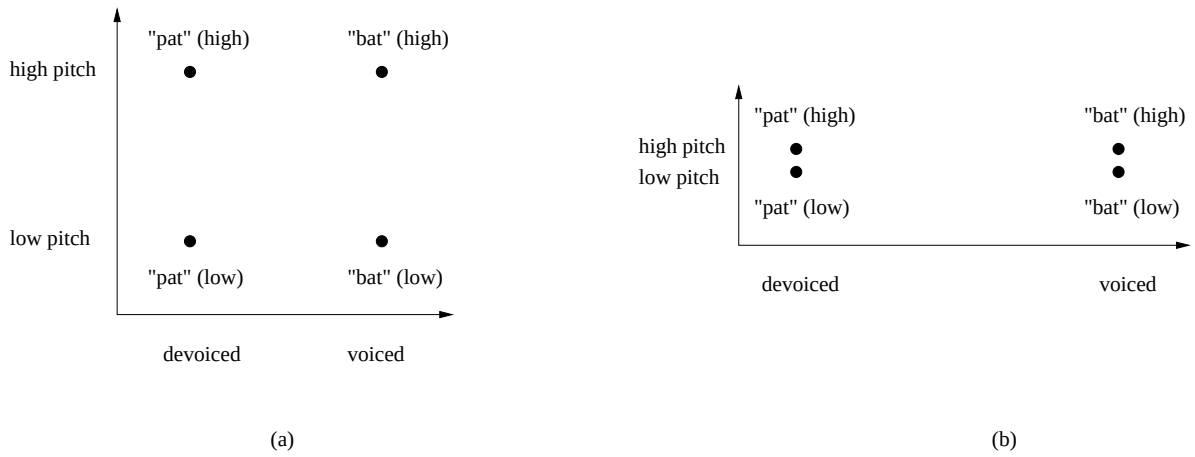


Figure 2: Words as clusters of exemplars

The Lexicon as EXemplars (LEX)

The LEX model brings together two separate strands of this earlier work: emergent symbols (Cottrell and Plunkett, 1994; Plunkett et al., 1992; Schafer & Mareschal, 2001), and competition with selective attention to dimensions (Merriman, 1999; Smith et al., 2002), particularly as formalized in exemplar-based models of categorization (Kruschke, 1992; Nosofsky, 1986). At the heart of this marriage is the idea that word forms and word meanings may both be usefully thought of as *clusters of exemplars*.

Consider, for example, the four exemplars of word form displayed in Figure 2(a):

the word "pat" spoken with high pitch,
the word "pat" spoken with low pitch,
the word "bat" spoken with high pitch, and
the word "bat" spoken with low pitch.

The two exemplars of the word "pat" differ in pitch, just as the two exemplars of "bat" do. Similarly, the two high-pitch exemplars ("pat" (high) and "bat" (high)) also differ by one feature: voicing of the initial consonant – just as the two low-pitch exemplars do. At the outset, in this idealization, the dimensions of voicing and pitch are equally weighted, so these exemplars do not cluster. However, the dimension of voicing is *communicatively significant* in English - that is, one must attend to voicing in order to correctly predict meaning - while the dimension of pitch is not (although it is in tone languages). Thus, voicing will gradually receive more attention, stretching the voicing dimension, and pulling the two "bat" exemplars away from the "pat" exemplars. At the same time, pitch will receive *less* attention, compressing the pitch dimension. The result is shown in Figure 2(b): the two "pat" exemplars are near each other, as are the two "bat" exemplars. And these two groupings, each corresponding to an English word, are distant from each other. In this manner, words, as discrete unitary symbols, may emerge gradually as tightening clusters of exemplars (Pierrehumbert, 2001; Johnson, 1997).

An analogous grouping process takes place over meanings - such that meaning exemplars that differ along significant dimensions, such as shape for object nouns in English, are pulled apart,

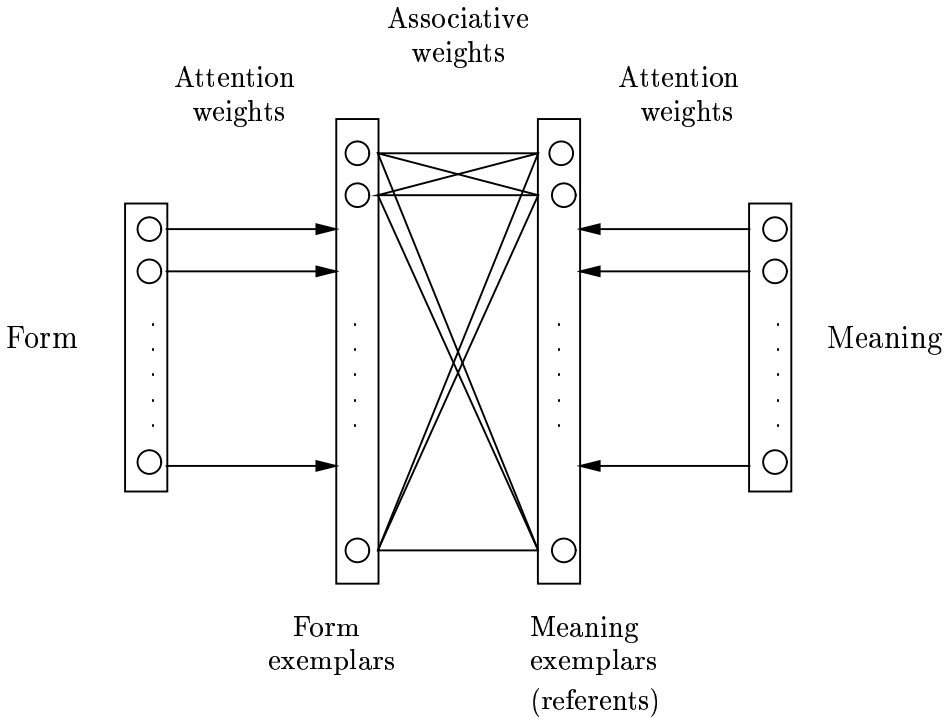


Figure 3: The LEX model of word-learning

---

while others are pulled together. This forms clusters of meanings, analogous to the clusters of word forms. On this view, feature selection – the discrimination of relevant from irrelevant features (e.g. Dy & Brodley, 2004; Theodoridis & Koutroumbas, 1999) – is a core element of word-learning.

The essence of the current proposal is that individual exemplars of form are associated with individual exemplars of meaning – then, as both form and meaning exemplars group into clusters, the result is a set of associations between categories (clusters) of form and categories (clusters) of meaning.

Overview

The LEX model is shown in Figure 3. It is a bidirectional associative memory: given a word form, the model produces a probability distribution over associated referents (i.e., exemplars of meaning); analogously, given a referent, it produces a probability distribution over associated exemplars of form. These two-way associations are mediated by a single set of associative links, connecting the two hidden layers of the model. The hidden layers contain nodes that represent already-encountered exemplars that have been stored - one hidden layer for form exemplars, and one for meaning exemplars, or referents. Form exemplar nodes and meaning exemplar nodes are directly associated one-to-one – thus, the model contains no unitary representation that corresponds either to a word or to its meaning.

In addition to the associative weights, there are also weights encoding selective attention to each dimension of form, and each dimension of meaning - these stretch and compress the underlying psychological spaces, as described above. Potential forms and potential meanings are assumed to

have been separated from each other,[2] but the model must learn which dimensions of form, and of meaning, are communicatively significant.

The associative weights and attention weights are the free parameters of the model.

LEX adopts the *macro*structure of Plunkett et al.'s (1992) model, in that either word form or word meaning may be supplied as input, and either may be retrieved as output. The two models also illustrate the same very general point: words and their meanings may be paired through subsymbolic associations – that is, associations among elements that are finer-grained than either words or their meanings. But LEX's *micro*structure is much more similar to Kruschke's (1992) ALCOVE model: the equations and concepts governing LEX's operation have been adapted from ALCOVE, and from ALCOVE's predecessor, Nosofsky's (1986) categorization model. LEX extends and generalizes this earlier work, and the work of MacWhinney (1989) and Merriman (1999), by applying the same attentional learning processes to form and meaning, in parallel.

For the purposes of this article, I shall consider a model to be "associative" if it is built on general principles of strengthening connections between mental objects on the basis of spatio-temporal contiguity and similarity (James, 1890). A classic associative model is the Rescorla-Wagner (1972) learning rule, which accounts for Pavlovian conditioning phenomena such as overshadowing, extinction, and blocking (Kamin, 1969). Connectionist networks that operate under gradient descent in error are variants of the Rescorla-Wagner model (Gluck & Bower, 1988; Sutton & Barto, 1981). Since LEX and several of its predecessors (e.g. Kruschke, 1992; Plunkett et al., 1992) are connectionist networks of this sort, they are descendants of this well-known associative model – and clearly "associative" themselves.

Formal presentation

For purposes of exposition, assume we are given form as input, and wish to receive meaning as output. In actuality, the model is symmetrical, so the same process operates in the other direction as well.

Consider a multidimensional space of possible word forms. Each instance of a word form (and thus each form input pattern, and each stored form exemplar) is represented as a point in this space. Assume several exemplars of form and several of meaning have been encountered and stored in the model; there is a node corresponding to each such stored exemplar. An input form pattern strongly activates those stored form exemplars that are near it in space, and weakly activates those that are distant from it. Distance is modulated by the amount of attention paid to each dimension of the space. Specifically, the psychological distance between the input and stored form exemplar i is:

$$d_i = \sqrt{\sum_j s_j (inp_j - ex_{i,j})^2} \tag{2}$$

Here j indexes dimensions of form, s_j is the selective attention to form dimension j, inp_j is the value (generally 0 or 1) of the form input along dimension j, and $ex_{i,j}$ is the value of stored form exemplar i along dimension j. Distance d_i will be 0 when the input matches exemplar i along all dimensions j for which $s_j > 0$; the remaining dimensions do not affect distance. The activation of form exemplar node i is then a Gaussian function of this psychological distance:

$$a_i = exp(-d_i^2), \tag{3}$$

[2] See the General Discussion for discussion of how this may be accomplished.

and these form exemplars i activate their associated *meaning* exemplars k:

$$net_k = \sum_{i \in F} w_{ki} a_i \tag{4}$$

Here, net_k is the net input to meaning exemplar node k, F is the set of all form exemplar nodes, i indexes over these nodes, and w_{ki} is the associative weight on the link between form exemplar node i and meaning exemplar node k. We normalize to obtain a probability distribution over stored meaning exemplars:

$$p_k = \frac{net_k}{\sum_{l \in M} net_l + noise}. \tag{5}$$

where p_k is the probability of producing meaning exemplar k, M is the set of all meaning exemplar nodes, and l indexes over these nodes. The *noise* constant denotes the level of noise in the system, and prevents division by zero when $net_l = 0, \forall l$. Thus, given a form as input, the model produces a probability distribution over meaning exemplars (referents). Analogously, when operating in the other direction, given a referent as input, we obtain a probability distribution over form exemplars. Equations 2 through 5 are adapted from the categorization models of Kruschke (1992) and Nosofsky (1986).

When the current form input is perceived as novel, a new form exemplar node is allocated, corresponding to that input. Specifically, for each presentation of input, we define:

$$a_{new} = 1 - max_{i \in F}(a_i) \tag{6}$$

where F again denotes the set of all form exemplar nodes. This quantity captures novelty since it is activated to the extent that known form exemplars are not. A new form exemplar node is then allocated, positioned at the current input, with probability:

$$p_{new} = \frac{a_{new}}{\sum_{i \in F} a_i + a_{new}}. \tag{7}$$

All weights on connections touching this new node are initialized at zero.

Training

The model is trained under gradient descent in error, on a training set of word forms paired with their referents, both instantiated as bit vectors. A pattern over communicatively significant dimensions of form is predictive of patterns over significant dimensions of meaning, and vice versa.[3] Other dimensions of both form and meaning are insignificant – that is, patterns across them are not predictive of anything. The model must learn to allocate attention to significant dimensions, and away from insignificant ones.

We consider the training of the model when it is given word form as input, and produces a probability distribution over word referents at the output. We calculate error for this output by comparing it with the referent supplied by the teacher. A natural approach would be to compare each meaning exemplar node with the supplied referent, and consider the exemplar correct if it matches the teacher-supplied referent along significant dimensions; insignificant dimensions would

[3]Note that any given significant dimension of form will not be predictive of meaning (or vice versa) – it is only a *pattern* over significant dimensions that is predictive. This corresponds to the fact that a phoneme such as /b/ is not predictive of meaning, while a pattern of phonemes such as /bat/ is.

be ignored. We could then sum the probabilities of all correct meaning exemplars, and error would be one minus this sum:

$$E_m^{ideal} = 1.0 - \sum_{k \in C} p_k \tag{8}$$

Here C is the set of all correct meaning exemplars, k indexes over these nodes, and p_k is the output probability of meaning exemplar k as given by Equation 5. Unfortunately, this idealized method of determining error is not psychologically plausible, since it requires that the teacher designate each known exemplar as either correct or incorrect – and the child is not given such detailed information. The child can only assume that the supplied referent is correct, and the correctness or incorrectness of other exemplars must be inferred from that. This forms the basis for an error function that *estimates* the idealized one above:

$$E_m = 1.0 - \sum_{k \in M} g_k p_k \tag{9}$$

Here the sum is over *all* meaning exemplar nodes k; p_k is defined as above; and g_k is a Gaussian function of the psychological distance, in meaning space, between meaning exemplar k and the target referent supplied by the teacher. This value g_k is determined by analogy with Equations 2 and 3; since the distance is in meaning space, it is modulated by meaning attention weights. g_k is 1.0 for exemplars located exactly at the position specified by the supplied referent, and is effectively 0.0 for exemplars very distant from it. Thus, the overall quantity $\sum_{k \in M} g_k p_k$ is the sum of probabilities of all meaning exemplars, weighted by their psychological nearness to the teacher-supplied referent. This quantity is an estimate of the overall probability of a correct response. The estimate is based on the assumption that since the supplied referent is assumed to be a correct response, nearby exemplars are also likely to be correct. This assumption will become increasingly valid as the meaning space stretches and shrinks appropriately during training: the space will distort so as to bring exemplars of the same category near each other, and to pull apart exemplars of different categories, as in Figure 2. To this end, the attention weights for meaning, which govern this stretching and shrinking, are driven only by error in *form*, while the attention weights for form are driven by the above error in meaning (see Appendix A). This dependence of selective attention in one part of the model on error in the other follows the treatment of these quantities in Kruschke's (1992) category-learning model.

Analogous computations yield E_f, the error in predicting word *form*, given a referent. The total error in the model is form error plus meaning error:

$$E = E_f + E_m. \tag{10}$$

At each presentation of a form-referent pair to the model, only one weight is trained – capturing the idea that one exemplar of form is being associated with one exemplar of meaning (a referent). Given a form-referent pair from the training set, the model selects a form exemplar node f to represent that form, and a meaning exemplar node m to represent the referent. If a new form exemplar node has just been allocated, it will be the one selected to represent the form. Otherwise, the node is sampled from the set of all form exemplar nodes, with the probability of selecting node i proportional to the activation of that node a_i. The meaning exemplar node m is selected analogously. The weight between these two nodes is trained under gradient descent in the error quantity E, as described in Appendix A, which also describes the training of attention weights.

Testing of a learned word is described in Appendix B.

The Memory Interference Principle

There is a single central principle that provides this model with much of its explanatory power: *Learning of a novel word is most effective when memory interference is minimized.* "Memory interference" is taken to mean a situation in which, prior to learning, a given input causes exemplars at the output to be activated – that is, there are exemplars j for which $net_j > 0$. This principle will be referred to as the "memory interference principle".[4] Appendix C demonstrates formally that this principle holds in the present model. I take this principle, with its reliance on selective attention in both form and meaning, to be the central concept that this work advances.

This principle explains three of the four developmental shifts we examined earlier: those concerning the ease of learning, the similarity of word form, and the learning of second labels. The remaining developmental shift, the growth of meaning biases, such as the shape bias, is not directly explained by this principle – although it is explained by shifts in attention.

This section briefly outlines – in general conceptual terms – how the memory interference principle accounts for existing data.

1. *Ease of learning a novel word.* Early in learning, the form and meaning spaces will be relatively compact, since learning begins with only minimal attention paid to any given dimension. These spaces will also already contain some exemplars, corresponding to the child's prelinguistic experience. Under these circumstances, a new exemplar corresponding to a novel word will – despite its novelty – lie fairly near other exemplars in the form and meaning spaces. The nearness in form will cause neighboring form exemplars to be activated, and they will activate their associated meaning exemplars, yielding memory interference. The same process also operates in the other direction. Because of this interference, early word learning will be difficult. Later, the form and meaning spaces will warp (i.e. stretch and compress) appropriately, with attention allocated to significant dimensions and away from insignificant ones. This means that there will be no other exemplars near the novel word, in either form or meaning space – since the novel word by definition differs from other words along significant (now stretched) dimensions. Under these circumstances, there will be little memory interference, and therefore stronger learning. As more words are learned, they will begin crowding the form and meaning spaces, reintroducing some interference (Gershkoff-Stowe & Smith, 1997) – but this will then immediately drive further warping of the spaces, keeping the interference at a minimum.

2. *Similarity of form.* If the word to be learned is similar in form to an already-learned word, it will partially activate that word form, which will result in memory interference, and difficulty in learning. This is especially true early in learning, when the form space is compact, causing the new word to be psychologically even more similar to the already-learned word. Later, the warping of the form space will stretch apart exemplars that differ along significant dimensions of form, so the new word will only minimally activate existing wordforms. This will lessen the learning difficulty induced by similarity in linguistic form. This accounts for children's initial inability to learn formally similar words, and their later ability to do so.

3. *Second labels.* The role of memory interference is particularly clear for second labels. Imagine learning a novel word for an object that already has a known name. When that object is

[4] I am grateful to William Merriman for drawing my attention to memory interference as an explanation for word-learning phenomena.

presented as (meaning) input, it will activate the associated known name – yielding memory interference in learning the novel word. This will degrade learning, by the memory interference principle. This learning will be especially vulnerable early on, before the psychological spaces have warped appropriately, since the interference due to the first label is combined with interference due to an initially compact space. Later, with appropriately warped spaces, there will still be a disadvantage for second labels, but the difficulty will not be compounded by a compact space. This accounts for children's initial inability to learn second labels for objects, their persisting preference to avoid multiple labels for a single object, and their eventual ability to learn second labels, despite this preference.

Thus, this principle seems to account for these different yet simultaneous changes in children's word-learning. We turn next to simulations, to test this proposal computationally.

Simulation 1: Ease of Learning

The purpose of the first simulation was to determine whether LEX could account for children's progression from word-learning that requires multiple exposures, to one-trial learning, or fast mapping.

The simulation proceeded in two stages. LEX was first trained on a fixed training set of form-referent pairings. This was intended to capture the child's gradual acquisition of a lexicon. In the course of this acquisition process, exemplars were encountered and stored, and the form and meaning spaces stretched and shrank appropriately. The model was then tested on its ability to learn a *novel* word, one not included in the training set. This test was conducted at each epoch of training – this once-per-epoch testing is meant to correspond to testing children at different stages of development, to examine their word-learning abilities across time.

Training set. An artificial training set was used. The dimensions of the form and meaning spaces represented artificial phonological or semantic features (that were either significant or not), rather than natural features such as voicing or shape. This allowed a qualitative test of the ideas proposed here. A more complete test would require a naturalistic dataset, with dimensions corresponding to measurable phonological and semantic features.

The training set consisted of 50 form-referent pairs. The forms and referents were represented by either 1 or 0 on each dimension, indicating the presence or absence, respectively, of the feature corresponding to that dimension. There were 50 dimensions for form: 25 were significant (such that a pattern over these dimensions was always predictive of the referent) and 25 were insignificant (the values along these dimensions were set randomly at each presentation, such that they were never predictive). Similarly, there were 50 dimensions for meaning: 25 significant (such that a pattern over these dimensions was always predictive of the form), and 25 insignificant (not predictive). All 50 form patterns were distinct along significant dimensions, as were all 50 referent patterns. The specific form and referent patterns used were generated randomly, subject to the constraint that patterns over the significant dimensions of form be predictive of significant dimensions of meaning, and vice versa.

Test set. The test set consisted of three form-referent pairs: one target (a novel form-referent pair, on which the model was to be trained), and two distractors. The target pattern was constructed to be distant from all training patterns along significant dimensions of both form and meaning. The meaning dimensions for the two distractors were randomly generated; the form dimensions were the same as for the target. Insignificant dimensions for all three test patterns were specified as 0.

Procedure: training. The model was initialized with 15 random associations, representing the child's

pre-linguistic experience. Specifically, there were 15 form exemplars and 15 meaning exemplars, each positioned at a random corner of the [0,1] hypercubes defined by the dimensions of form, and of meaning. These random form and meaning exemplars were interconnected with associative weights set to random values in the interval [0,2]. All attention weights in the model were initialized at 0.3. The associative weight learning rate and attention weight learning rate were both set at 0.6. The *noise* constant from Equation 5 was set at 0.1. At each presentation of a pattern to the model, insignificant dimensions in the input were set randomly to either 0 or 1, such that they carried no predictive ability. This was meant to correspond to the child's encountering elements of both word form and referent that would ultimately have to be ignored for effective naming. The model was trained on the training set for 20 epochs. Both associative and attention weights were constrained to be ≥ 0. After each epoch of training, the resulting model state (including number and positions of form and meaning exemplars, and associative and attention weights) was stored, for retrieval during testing.

Procedure: testing. Once training on the training set had been completed, the model's ability to learn a novel word was tested. For each consecutive epoch, the model was initialized using the previously stored state that resulted from training at that epoch. The model was then trained for one exposure on the target pattern in the test set. The associative weight learning rate was set at 0.002, and the attention weight learning rate at 0.0, both substantially *lower* than the values used for training. This was done to attempt a conservative test of one-trial learning. After training for one trial, the model was tested: it was provided with the target form, and was required to select the target object from among the two distractors. This test was intended to correspond to a word-learning experiment, in which the child is asked, for example, "Where's the mido? Can you tell me which one is the mido?" after having been taught a new word, and being presented with its referent among distractors. The probability of correct choice was determined using Equation 26 (Appendix B). Chance performance was $\frac{1}{3}$. Production probability was also determined, using Equation 24 (Appendix B). This procedure was repeated at each epoch of training.

Two additional variants of this testing regime were pursued. The first examined whether the model would improve its performance given multiple training trials. To that end, the same test as that described above was performed, but with 10 training exposures rather than one.

The second variant of testing was intended to probe the resilience of one-trial learning. This test was also identical to the standard one-trial test described originally, except that immediately after one exposure to the new form-referent pattern, the model was trained for an additional epoch on the full training set. This exposed the training on the novel word to possible interference from subsequent training on other words. To make this a conservative test, the associative weight learning rate for this interference training was set to a value that should provide for strong interference: 2500.0, as compared with 0.002 for the one trial of training on the novel word. Also in the interests of a conservative test, the attention weight learning rate was set to 0, so that the form and meaning spaces would not be further stretched – since that could help preserve the newly-learned word.

Comprehension and production measures were obtained for these two variants of the standard test as well.

Results and Discussion. Figure 4 shows the error obtained on the training set, as a function of time. After an initial period of relatively high error, the model converges to much lower error by the 6th epoch. As a part of this process, the attention weights differentiated themselves, as shown in Figure 5. Here, (a) shows attention weights for form, while (b) shows attention weights for meaning. In both cases, the significant (predictive) dimensions, shown in solid lines, gradually

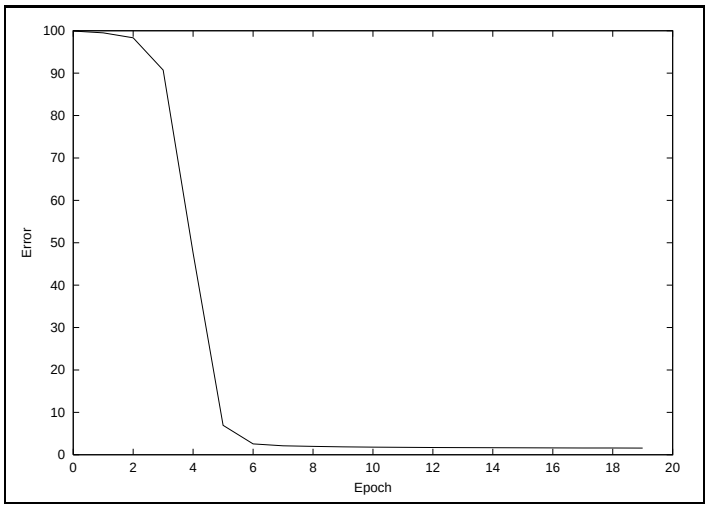
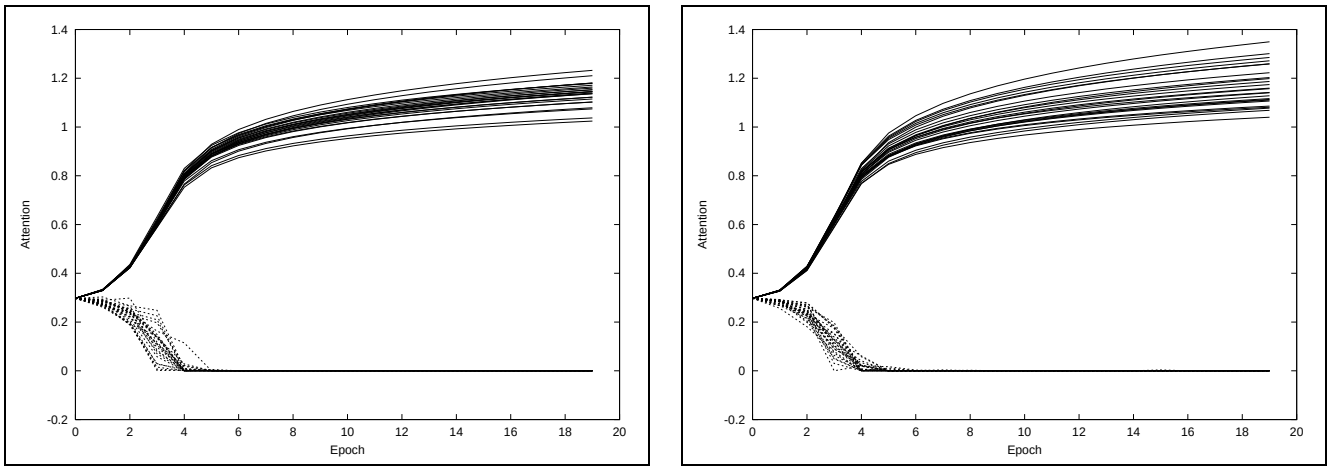
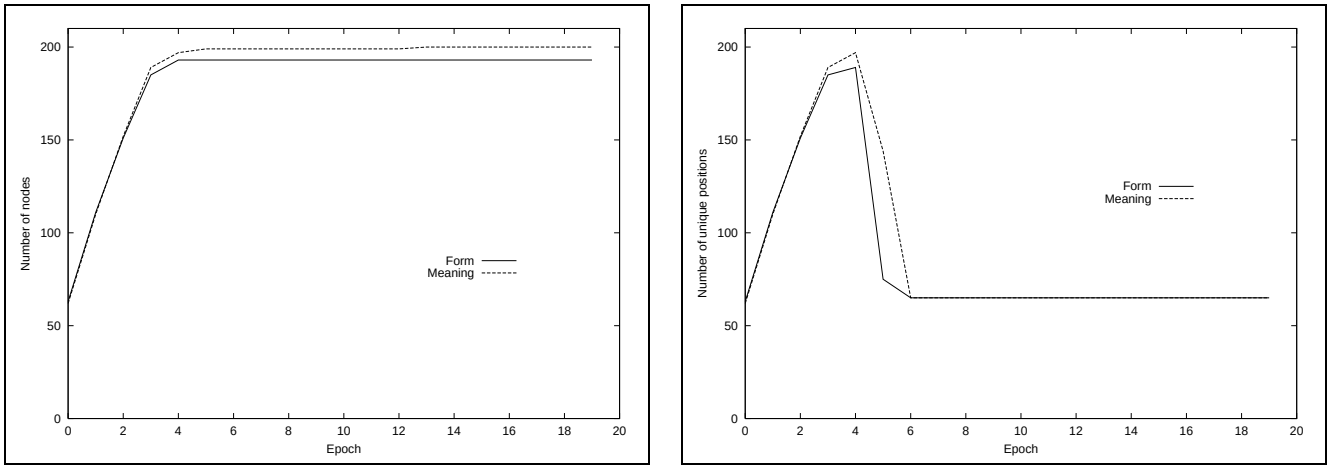


Figure 4: Error on the training set, over time



(a) (b)

Figure 5: Attention weights over time: (a) form, (b) meaning. Significant dimensions (solid lines) receive more attention with time, while insignificant dimensions (dashed lines) receive less.



(a) (b)

Figure 6: (a) Total number of exemplar nodes over time; (b) number of nodes at distinct positions.

received greater amounts of attention, while insignificant (non-predictive) dimensions, shown in dashed lines, eventually decayed to very near zero.

During training, exemplars were probabilistically added to the model. Figure 6(a) shows the number of form and meaning exemplar nodes in the model as a function of epoch number. Since insignificant dimensions of the input were set randomly during training, there were often novel variants of the training set patterns; this is reflected in the steady increase in the number of exemplar nodes. However, once the attention weights along insignificant dimensions fell to zero, between epochs 3 and 5, the number of exemplar nodes plateaued. This happened since no variant of a pattern in the training set, obtained by setting insignificant dimensions of that pattern randomly, was perceived as novel: it would differ from existing exemplars only along insignificant dimensions, which no longer contribute to psychological distance. This collapsing of insignificant dimensions is also reflected in Figure 6(b), which shows the number of form and meaning nodes at *distinct positions* in psychological space (assuming a tolerance of distance 0.1 for determining whether two nodes are coincident). As the insignificant dimensions contract, many exemplars ultimately coincide in space, yielding a reduction in the number of distinct points represented. This illustrates the very general idea of clusters of exemplars shrinking to points in psychological space.

The results of one-trial learning of a novel word are shown in Figure 7. The solid line shows comprehension probability: the probability of correctly selecting the referent of the novel word, given one trial of learning, and then given the novel form. This probability starts only just above chance levels, but eventually rises well above chance. This progression into 1-trial learning captures the behavior of 1-2 year old children. We do not have evidence of 1-trial learning in young one-year-olds, but shortly thereafter children become capable of word-learning given only 1-3 exposures (Behrend et al., 2001; Carey, 1978). The dashed line shows the probability of correct referent choice in the 10-trial test – this demonstrates that at the earliest stages of learning, when the model cannot effectively learn a novel word given only one trial, it *can* learn such a word given ten trials. This success at word-learning given multiple trials at very early stages of development matches the behavior of very young one-year-olds, who can learn novel words under such conditions (Schafer & Plunkett, 1998; Woodward et al., 1994). Finally, the dotted-and-dashed line in the figure shows the

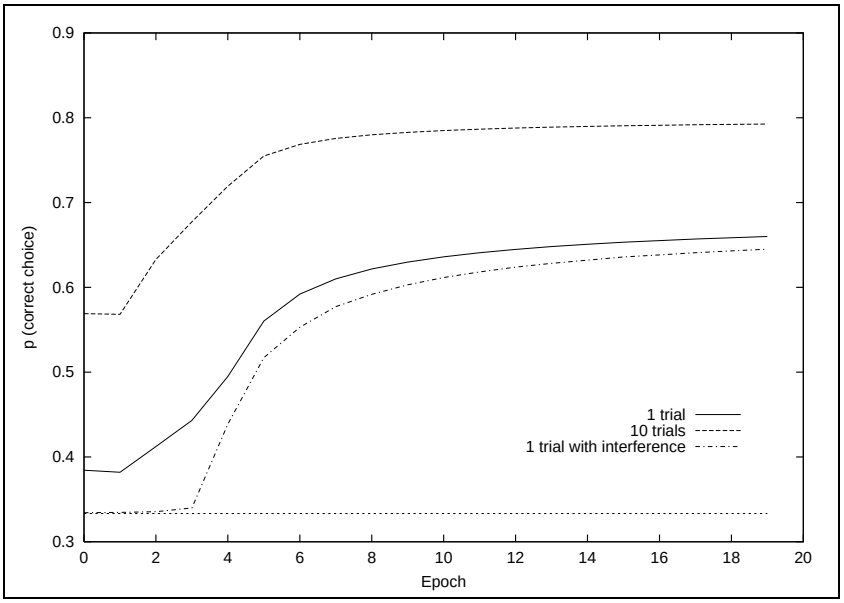


Figure 7: Learning a novel word. The solid line indicates the probability of correctly selecting the referent of the novel word, after 1 trial of learning, over time. The dashed line indicates this probability given 10 trials of learning, rather than just one. The dotted-and-dashed line indicates the probability of correct choice after 1 trial of learning, followed by interference training. The dotted line indicates chance performance.

---

probability of correct referent choice when the model is trained on the novel word once, and then given interference training that has the potential to disrupt one-trial learning. The training on the novel word is quite resistant to subsequent exposure to other form-meaning pairs, as appears to be true of children (Carey, 1978; Markson & Bloom, 1997).

Why does LEX behave in this fashion? The central reason for improved one-trial learning is the memory interference principle: learning becomes more effective as memory interference decreases – which occurs as significant dimensions of form and meaning receive substantial amounts of attention. This increased attention pulls apart exemplars that differ along those dimensions, and that therefore should be kept apart. The relation between memory interference and learning is illustrated in Figure 8. The solid line is a measure of memory interference; specifically it is the denominator of Equation 5 for form[5], plus the analogous denominator for meaning. Since these denominators hold the total output activation for all exemplars, given associations from an input, their sum is a reasonable measure of memory activation. The dashed line is the weight change given one trial of learning. As shown in Appendix C, and illustrated here, learning improves as memory interference decreases. This demonstration suggests that the child's entry into one-trial word-learning may similarly reflect a gradual clearing of memory interference.

The model is resilient to interference because of its use of localized exemplar representations; LEX inherits this resilience from Kruschke's (1992) category-learning model ALCOVE. Once a novel word is learned, that learning establishes an associative link of some strength between an exemplar of form and an exemplar of meaning. That new link will be affected by subsequent training only if the later training selects those exemplars – which is unlikely for patterns that lie far

[5]This is the quantity referred to as *denom* in the Appendices.

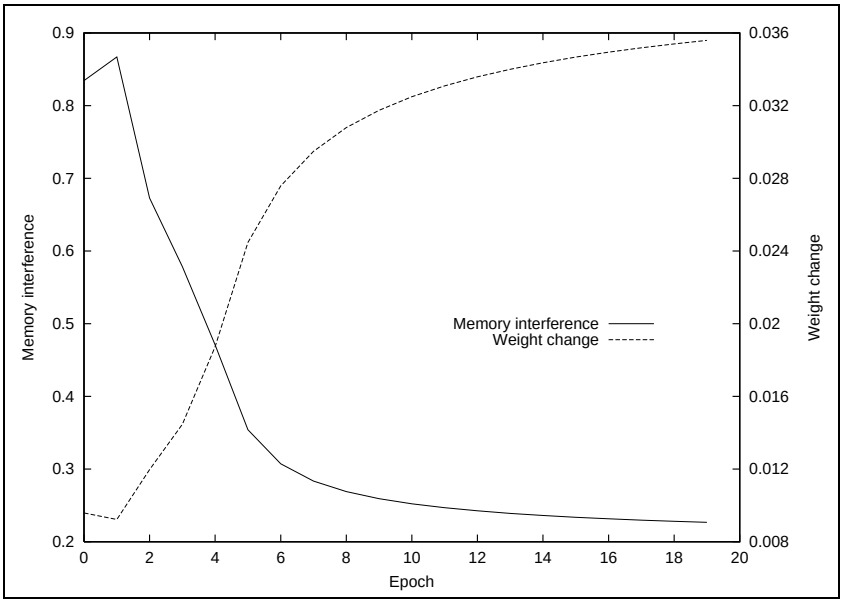


Figure 8: Memory interference and learning. Memory interference, measured by the denominator sum (see text), decreases with time. This decrease in interference causes the associative weight change from a single trial of learning to increase with time.

---

from the newly-learned word in form and meaning spaces. Thus, LEX almost completely avoids the problem of catastrophic interference, in which learning of one pattern is eliminated by subsequent learning of other patterns (McCloskey & Cohen, 1989). LEX does exhibit catastrophic interference on the rare occasions when the exemplars of the newly-learned word are selected during subsequent training. In 10 replications of the simulations in this paper, one datapoint in one replication, out of 200 datapoints in all replications, showed catastrophic interference.[6] In contrast, multi-layer perceptrons – a widely-used form of connectionist model – regularly suffer from catastrophic interference.

Ten trials of learning are more effective than one trial for standard reasons of error-driven associative learning: since there is residual error after one trial, further trials will strengthen the associative connection in trying to eliminate that error.

The model also exhibits a comprehension/production asymmetry, as shown in Figure 9. Here, the solid line is the same as in Figure 7, representing comprehension ability after one trial of learning. The dashed line represents the probability of correct production after one trial of learning. Such a comprehension/production asymmetry, favoring comprehension, is often noted in children's lexical development. LEX's account of this phenomenon is similar to that of Huttenlocher (1974): the task of comprehension, as tested here and in many word-learning experiments, requires only *recognition* – the child hears the word, and must select from among several presented objects. In production, in contrast, there is no external set of (e.g. sound) stimuli to "recognize" – the name must be *recalled.* In the model, comprehension and production probabilities are computed in a manner that reflects this asymmetry: the presence of a target object and two distractors is central to the computation of comprehension probability (Equation 26, Appendix B), while there are no such external supports in the computation of production probability (Equation 24, Appendix B).

[6]These replications are at http://www.psych.uchicago.edu/~tpregier/words/.

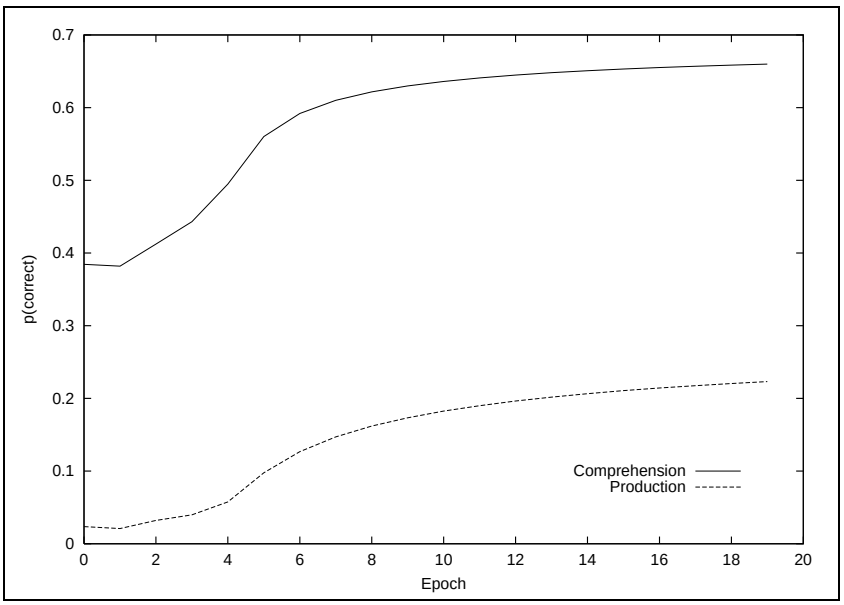


Figure 9: Comprehension/production asymmetry: The solid line represents comprehension following a single trial of learning, while the dashed line represents production.

Simulation 2: Honing of linguistic form

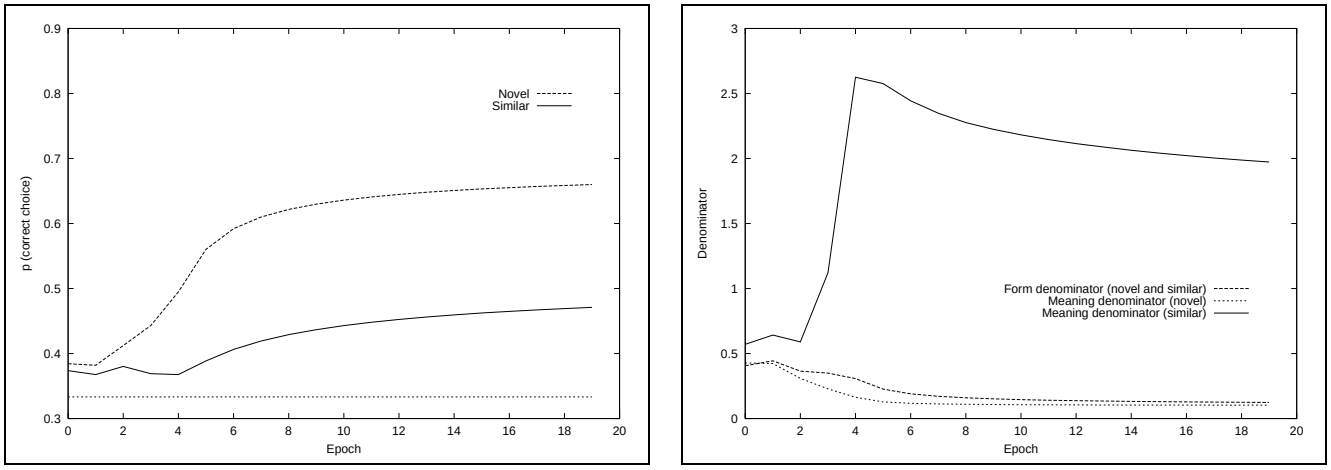
As we have seen, young children sometimes have difficulty learning a word that is similar in form to another word – but they can learn dissimilar-sounding words. In contrast, older children can learn both similar- and dissimilar-sounding words (Stager & Werker, 1997; Werker et al., 1998). The next simulation examined whether LEX could account for these phenomena. To that end, the model was trained on a novel word that was similar in form to an existing word; its ability to learn this similar word was compared to its ability to learn the entirely novel word from the previous simulation.

This simulation built on the previous one. No new training on the training set occurred in this simulation – instead, the tests in this simulation were based on the previously stored states from each epoch of the training run in Simulation 1.

Test set. The test set was almost identical to that used in Simulation 1, but with one difference. Whereas before the new word to be taught to the model was distant in both form and meaning from items in the training set, the target word in the current test set differed by only 1 significant bit in form from one of the training patterns. The referent for this word was, as before, constructed to be distant from those in the training set. The target word in this simulation will be referred to as the "similar word", while the (dissimilar) target from the previous simulation will be referred to as the "novel" word.

Procedure. The procedure was analogous to the one-trial testing phase of Simulation 1, and was repeated at each epoch of training on the full training set. The model was trained for one exposure on the target pattern in the test set, after which comprehension probability was determined.

Results and Discussion. The results are shown in Figure 10(a). The dashed line indicates, for reference, the comprehension probability for the novel word from Simulation 1. The solid line indicates the comprehension probability for the similar word in the current simulation: there is a



(a) (b)

Figure 10: Similarity of form. (a) The solid line indicates the probability, after one trial of learning, of correctly choosing the target referent of a novel word that is similar in form to a known word. The dashed line indicates this choice probability when the novel word is not similar to any known words (from Simulation 1). (b) Memory interference for similar and novel words.

---

clear disadvantage for the similar word, relative to the novel word. Importantly for our purposes, this disadvantage keeps the comprehension probability for the similar word low, and near chance, early in learning. Later in learning, despite the persisting disadvantage for the similar word, its comprehension probability rises, eventually rising well above chance, which is $\frac{1}{3}$. This simulation mirrors Stager and Werker's (1997) finding that very young children cannot learn similar-sounding words, although they can learn dissimilar-sounding words. It also captures the finding that slightly older children are capable of learning such confusingly similar-sounding words (Werker et al., 2002).

These results, like those of the previous simulation, are attributable to the memory interference principle. As discussed earlier, a similarity in form will cause memory interference: when the new word form is presented, it will partially activate the existing word form to which it is similar, which will in turn activate its associated meaning nodes. This scenario is illustrated in Figure 10(b). Here, memory interference is compared across this simulation and the previous one – comparing similar and dissimilar words. The solid line shows the meaning denominator for the similar word – this quantity is substantially higher in this case than for the dissimilar word, throughout learning, consistent with the above reasoning. (The form denominators are identical for the two words: the referent is the same in the two cases, and therefore induces the same distribution of activation over form exemplars.) This strong memory interference impedes learning. For both words, the memory interference does eventually decay as the spaces stretch, more clearly separating confusingly similar representations of form. Further stretching would nearly eliminate all memory interference, if it pulled the form exemplars far enough apart to prevent activation of neighbors. Thus, eventually, similar words could in principle be linked to objects almost as easily as dissimilar words.

These results suggest that children may improve at linking similar words to their referents because of growing selective attention to significant phonological dimensions, and a resultant decrease in memory interference.

Simulation 3: Growth of meaning biases

As we have seen, very young children sometimes tend to generalize newly-learned object names conservatively (Woodward et al., 1994), while older children generalize English object names on the basis of object shape (Smith, 2000). The next simulation examined whether LEX could account for the growth of meaning biases such as the shape bias. To this end, the model was trained on a novel form-meaning pairing – the same one used in Simulation 1 – and was then tested to see whether it would generalize the new word to referents other than the one on which it had been trained. The specific question was whether a change in *insignificant* dimensions of the referent would affect generalization. Since object categories are often defined by shape in English, rather than size or color, this investigation is analogous to teaching a child that a large red ball is called a "ball", and then seeing if the child will apply the word to a small white ball – that is, to a referent differing from the original along the insignificant dimensions of size and color.

This simulation, like the previous one, builds on the central training run conducted in Simulation 1.

Test set. The test set was almost identical to that used in Simulation 1. The only difference was that in addition to the novel word to be learned, a meaning-variant of that word was also specified. This variant had the same form as the novel word, and the same significant dimensions of meaning – but the insignificant dimensions of meaning were all set to the opposite of the values they held in the novel pattern (all 0s were set to 1, and vice versa). This was intended to test the possible application of the newly-learned word to an object that differed from the original exemplar along insignificant dimensions – in the case of object nouns, such dimensions might be color, size, etc.

Procedure. The procedure was similar to the investigation of one-trial learning in Simulation 1. For each previously stored model state, corresponding to consecutive epochs of training on the full training set, the model was first initialized using that stored state. It was then trained for one exposure on the novel word, and tested for comprehension on the *variant* of that word, to determine whether it would generalize the novel word to a referent that varied from the original along insignificant dimensions. This was repeated at each epoch of training on the full training set.

Results and Discussion. The results are displayed in Figure 11, in which the solid line represents the comprehension probability for the variant, while the dashed line represents the comprehension probability for the novel word from Simulation 1, reproduced here for reference. LEX is originally quite conservative in generalizing the new word to an exemplar that differs from the original along insignificant dimensions. Eventually, however, the model comes to generalize strongly from the novel word to the variant – corresponding to children's apparent initial conservatism, and eventual strong generalization along appropriate dimensions.

This behavior results from the changing selective attention to different dimensions of meaning. As shown in Figure 5(b), the attention to significant and insignificant dimensions of meaning is originally roughly equal. As long as the insignificant dimensions receive a substantial amount of attention, the model will fail to generalize completely to a variant that differs from the original along those dimensions. Eventually, however, the insignificant dimensions receive effectively no attention – which means that differences along these dimensions become irrelevant for naming, such that the novel word is strongly generalized to the variant of the original referent at this time. In this manner, a meaning bias can be learned through gradual attentional shifts (Smith et al., 2002).

The developmental data suggest that the shape bias – an instance of such a meaning bias – eventually becomes quite strong. Landau, Smith, and Jones (1988) found that some 2-3-year

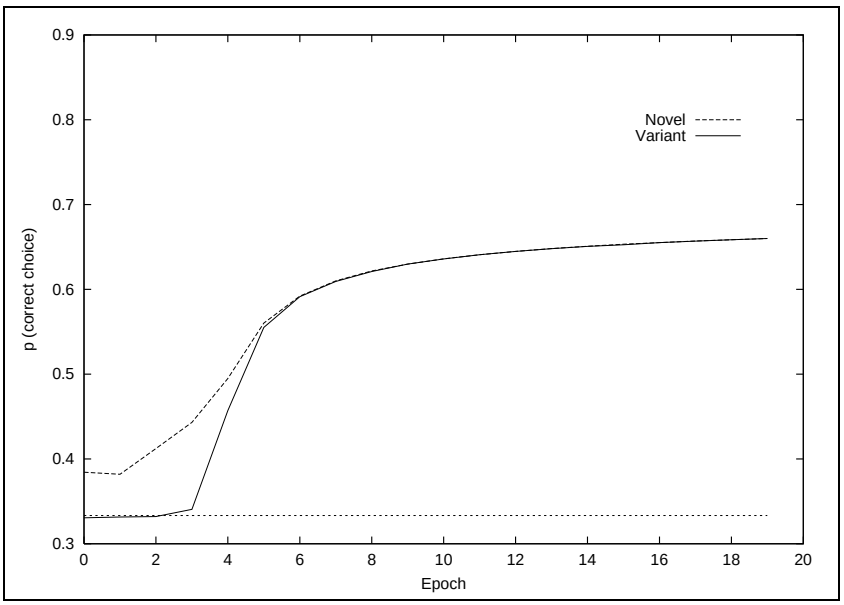


Figure 11: Growth of a meaning bias. The solid line indicates the probability, after one trial of learning, of choosing a referent that differs from the original along insignificant dimensions. The dashed line indicates the probability of choosing the original referent itself (from Simulation 1).

---

children would generalize newly-learned object names to objects many times the size of the original, or differing from it substantially in material – in one test, children generalized a name from a wooden object to a replica that preserved shape but was made of chicken wire. These constitute large changes along "insignificant" dimensions. Thus, it is important for current purposes that this simulation involved a variant referent that was deliberately made as different as possible from the original along insignificant dimensions. Despite this, the model generalized the newly-learned word to this variant.

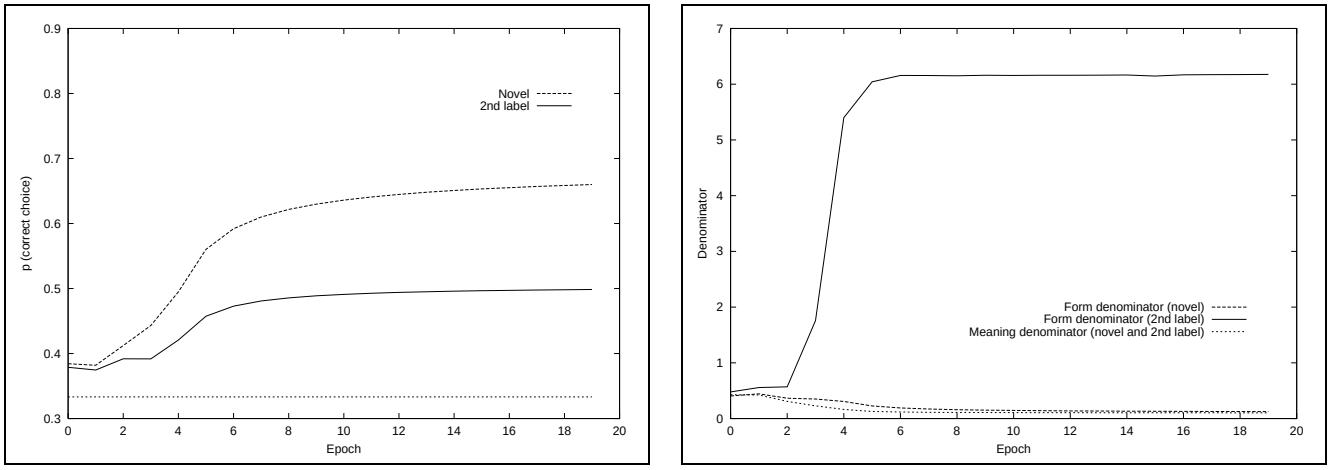
Such an account has frequently been promoted by others. Smith (2000; see also Landau et al., 1998; Smith et al., 2002) argues that the shape bias may stem from selective attention to dimensions during associative learning, and cites Kruschke's (1992) ALCOVE model – on which LEX builds – as an instance of this general principle. Merriman's (1999) word-learning model CALLED accounts for the shape bias through selective attention to dimensions – and also accounts for mutual exclusivity effects. LEX incorporates these existing proposals, and unifies them with an account of other word-learning phenomena.

Simulation 4: Second labels

As we have seen, children resist learning another name for an already-named object. This is particularly true of young 1-year-olds, who cannot learn second labels for objects (Liittschwager & Markman, 1994). 2-year-olds, in contrast, are capable of learning second labels (Liittschwager & Markman, 1994; Mervis et al., 1994), although they tend to prefer mappings in which an object has only one name. This simulation examined whether LEX could account for these findings.

The central training run for this simulation was, once again, that of Simulation 1.

Test set. The test set was similar to that used in Simulation 1, but with one difference. In Simulation 1, the referent of the target word to be learned by the model was different from those of all words



(a) (b)

Figure 12: Second labels. (a) The solid line indicates the probability, after one trial of learning, of correctly choosing the target referent of a novel word that is a second label for an object. The dashed line indicates this choice probability when the novel word is not a second label (from Simulation 1). (b) Memory interference for second label and novel word (first label).

---

in the training set. In contrast, in this simulation, the target word had the *same* referent as one of those words. Specifically, the meaning dimensions of the target pattern were set to be identical to those of one of the training patterns to which the model had been exposed during training. The form dimensions of the target were the same as in Simulation 1: distant from word forms in the training set.

Procedure. The procedure was analogous to the one-trial testing phase of Simulation 1: the model was trained for one exposure on the target pattern in the test set, after which comprehension probability was determined. This was repeated at each epoch of training on the full training set.

Results and Discussion. The results of this test are shown in Figure 12(a). The dashed line indicates, for reference, the comprehension probability for the (first label) novel word from Simulation 1. The solid line indicates comprehension probability for the second label in the current simulation – there is a disadvantage for the second label, relative to a first label. This disadvantage keeps the comprehension probability for the second label near chance at the beginning of learning. Later, despite the persisting disadvantage for the second label, the comprehension probability rises well above chance performance, shown by the dotted line in the figure. Thus, this simulation qualitatively captures the findings of Liittschwager and Markman (1994) and Mervis et al. (1994): throughout development, second labels are harder to learn than first labels – and although very young children cannot learn second labels, older children can.

As with several of the earlier simulations, the present results are traceable to the memory interference principle. When the second label and its referent are presented to the model, the referent will match existing meaning exemplar nodes along significant dimensions – and will activate them. These will in turn activate their associated form nodes, resulting in memory interference, and impaired learning. Figure 12(b) illustrates this. Here, memory interference in this simulation is compared with that in Simulation 1, in which a novel word was paired with its referent. The quantities displayed are the form and meaning denominators – the same measures of memory

activation that were examined earlier – obtained for this simulation and for Simulation 1. As expected, throughout learning, the form denominator for the second label is higher than that for the novel word from Simulation 1. (The meaning denominators for the second label and the novel word are identical: the two patterns have the same form, which therefore activates the same pattern over meaning exemplars (referents).) The strong memory interference weakens the learning of form-referent mapping in the case of the second label. LEX derives its ability to account for this effect from the interaction of competition and attention embodied in Merriman's (1999) word-learning model, CALLED – on which LEX builds.

Predictions

LEX makes two predictions, which are open to empirical test.

The first prediction is that the four phenomena of interest – fast-mapping, learning of similar forms, second-label-learning, and the shape bias – should change in synchrony over time, within a given individual. This follows since they are all mediated by changes in the same underlying representation: selective attention. This prediction assumes that attention to form and to meaning will change roughly in parallel, as in the simulations. However, even if attention to form and meaning change at different rates, fast-mapping, which relies on shifts in attention in *both* form and meaning, should show some correlation with each of the other phenomena, each of which depends primarily on attention shifts in either form, or meaning.

The second prediction concerns similarity in form and meaning. Since weight updates are affected by both form and meaning, two words that are similar in form *and* have similar referents should be maximally difficult to learn and keep distinct in memory; two words that are similar in only one or the other should be of intermediate difficulty; and two words that are dissimilar in both form and referent should be relatively easily learned. This idea suggests that the known difficulty of linking two similar-sounding words might be alleviated by making the *referents* of the two words more dissimilar – to lessen any semantic contribution to memory interference. While this prediction has not yet been tested, there is some indirect evidence that may suggest support. Martin, Gagnon, Schwartz, Dell and Saffran (1996) examined the errors produced by adult subjects, both normal and aphasic, in a picture naming task. They focused in particular on semantic errors, in which the subject intends to produce one word, but instead produces a semantically related one. They found that in both groups of subjects, these semantic errors preserved phonological characteristics of the target at rates greater than chance. Thus, when there is both semantic and phonological similarity between two words, subjects seem to be particularly susceptible to errors. A more direct test of this prediction would manipulate both formal and semantic similarity in a word-learning task, with young children.

General Discussion

1- to 2-year old children undergo a qualitative change in word-learning: they learn words more easily, become sensitive to small differences in word form, correctly generalize newly-learned words to novel exemplars, and become able to learn second labels. It has been suggested that improvements in word-learning at this age may result from a conceptual insight into the referential nature of words. The LEX model suggests a different account. On this model, the improvements may result not from an abrupt insight, but rather from an associative learner gradually determining which aspects of the world are relevant for communication, and attending preferentially to those dimensions.

In proposing this account, the present work suggests *continuity* of three sorts. First, it suggests

continuity in mechanism across time – for the improvements in word-learning are explained by continuous changes within a single mechanism (Plunkett et al., 1992; Schafer & Mareschal, 2001; Siskind, 1996). Second, it suggests continuity across phenomena. It unifies the four word-learning phenomena we have discussed, suggesting that they ultimately spring from the same underlying source: shifts in selective attention, in both form and meaning.

The third form of continuity is that these word-learning phenomena may be linked not just to each other, but also to learning behavior that is not directed specifically at words. The LEX model's internal structures are adapted from existing models of category-learning in adults (Kruschke, 1992; Nosofsky, 1986). Moreover, its component parts appear to be shared with non-human animals: error-driven learning (Rescorla & Wagner, 1972), selective attention to dimensions (Mackintosh, 1965), and representations corresponding to cue combinations (here, exemplars) (Pearce, 1987; Pearce, 1994). In this sense, LEX, building on earlier work by MacWhinney (1989), Merriman (1999), Plunkett et al. (1992) and Smith (2000), grounds some aspects of word-learning in more general cognitive processes.

This account touches a number of current issues in word-learning.

Fast mapping: what is the mechanism?

Fast mapping has attracted attention in part because of a suspicion that it might reflect a dedicated word-learning mechanism unique to humans. However, several recent findings challenge this idea. Markson and Bloom (1997) showed that children and adults can fast-map new *facts* about objects (e.g. "my uncle gave this to me"). The newly-learned facts were retained for up to a month, paralleling the long-term retention of fast-mapped words (but visually presented information was not retained as well; see also Behrend, Scofield, & Kleinknecht, 2001; Waxman & Booth, 2000 on differences between word-learning and fact-learning). And recently, Kaminski et al. (2004) showed that a dog is able to fast-map human-uttered words to their referents. These findings suggest that fast-mapping is specific neither to words, nor to humans. Instead, it seems to rely on general learning processes of some sort.

But of what sort? The present work suggests one possibility: the rapidity and resilience of fast mapping may be a natural result of associative learning, once adequate attention is paid to relevant features of form and meaning. This suggests that we might find similarly rapid and resilient learning in other species – and to some extent we do, e.g. one-trial spatial learning in young chicks with retention up to 24 hours (Regolin & Rose, 1999); one-trial learned food avoidance in squirrel monkeys and marmosets (Laska & Metzker, 1998) and even in slugs (Sahley, Gelperin, & Rudy, 1981). At the same time, there are species differences in the *kinds* of stimuli that can be associated in one trial (Laska & Metzker, 1998). Critically, only humans (and now apparently at least one dog) are known to fast map words to their referents. Why should this be? A plausible answer is that humans (and perhaps some dogs) appear to have social abilities and motivations that other animals lack – specifically, the ability to understand communicative intent, and the motivation to share experiences with others (Tomasello, Carpenter, Call, Behne, & Moll, in press; see also Hare, Brown, Williamson, & Tomasello, 2002 on social cognition in dogs). These species-*specific* social abilities may serve to highlight both word and referent – allowing the two to be rapidly and firmly associated through species-*general* associative learning. This line of argument is not meant to imply that human word-learning is reducible to the sort of word-referent linking that has been found in dogs; merely that the specific phenomenon of fast-mapping, defined as rapid and resilient linking of a form with a referent, may appear in these two species for similar reasons. We return below to the relation of associations and social cognition in word-learning.

There are other computational explanations of the progression from multi-trial to one-trial word-learning. Siskind (1996) proposes that this progression results from children gradually acquiring enough partial knowledge about the meanings of some words to constrain the meanings of other words. It is not yet clear whether the progression results from partial knowledge, from gradual shifts in attention, from some combination of the two, or from something else.

Attention to dimensions of meaning.

The idea that word-learning can be explained in associative and attentional terms has often been advanced by Linda Smith and colleagues, particularly in connection with the shape bias (e.g. Landau et al., 1988, Smith et al., 1996; Smith, 2000; Smith et al., 2002). Specifically, they have argued that as children learn that shape often predicts object name, more attention is allocated to shape, resulting in a shape bias, and enhanced learning of more object names. The LEX model, building on the work of Merriman (1999), instantiates these ideas, and unifies them with word-learning phenomena that concern word form as well as meaning – suggesting a parallel between the learning of form and the learning of meaning (see also Hespos & Spelke, 2004 for other evidence of such a parallel).

LEX's account of meaning biases has a limitation, however. In LEX, the amount of attention paid to a dimension of meaning changes only rather gradually, through learning. In language, in contrast, attention shifts rapidly from one aspect of a referent to another. For instance, in the simple noun phrase "the red ball", the three words single out three aspects of the same referent: its definiteness, its color, and its shape. A given meaning bias, such as the shape bias (or "intended purpose" bias, Diesendruck et al., 2003), appears to be associated with syntactic frames suggesting object names, such as "Look at the ___", while other syntactic frames suggest other aspects of meaning. A more complete account of meaning biases and their development would require the ability to modulate these biases based on the current syntactic frame (Fisher et al., 1991), and to rapidly shift attention accordingly (see Niyogi, 2002, for a Bayesian model that adjusts meaning biases depending on syntactic frame). This attentional shifting should also be under the control of conceptual knowledge about the object being named, since such knowledge has been shown to modulate the shape bias (Booth & Waxman, 2002).

Word form and selective attention.

The account proposed here is consistent with a number of existing results concerning word form, beyond those already discussed. Woodward and Hoyne (1999) tested children's ability to learn the correspondence between a novel sound, made by the experimenter on a noisemaker, and an object. They found that 13-month-olds could learn the connection: they reliably chose the target object in response to the sound. However older children, 20-month-olds, failed on the same task – suggesting that these older children had clear expectations as to what form a word should take, and did not entertain the noise as a possible word. Children of both ages were successful at learning novel words that were spoken in the usual manner. This developmental trend, from initial success in learning a non-canonical word, to later failure, is compatible with gradual shifts in selective attention, as in LEX. In the model, all dimensions of word form originally receive equal, but weak, attention. Under these circumstances, a non-canonical word – such as the sound made on the noisemaker – should be learnable since there is some attention paid to the dimensions of form along which it is specified. Later, however, the child will come to allocate attention away from these dimensions, in favor of the dimensions that are habitually used for communication, such as voicing and place of articulation. This will hamper the learning of a new non-canonical word, specified along unattended dimensions. This account predicts that those children of a given age

who are good at learning canonical words should be poor at learning non-canonical words, and vice versa. This follows since the two phenomena are, on this view, linked by a single psychological representation: selective attention to dimensions of word form.

There has been debate recently over the possibility that young children's underlying lexical representations may lack phonological detail. LEX exemplifies the opposing view, in that the overall phonological representation of a word consists of a cluster of fine-grained phonological exemplars, preserving all phonological detail encountered in utterances of that word. There is some evidence consistent with this aspect of the model. Swingley and Aslin (2002) find that 14-month-olds – who Stager and Werker (1997) found could not learn similar words – nonetheless do possess detailed phonological representations. They show that these children can detect minor mispronunciations of known words (e.g. "vaby" instead of "baby"; see also Swingley & Aslin, 2000 and Bailey & Plunkett, 2002 for analogous findings with slightly older children, and Ballem & Plunkett, in press, for evidence that even 14-month-olds can detect mispronunciations of novel as well as known words). In reconciling these findings with those of Stager and Werker (1997), Swingley and Aslin (2002) argue that the underlying representations of words may be phonologically detailed – but if two words are similar, that similarity may cause memory interference between the words' well-detailed representations, and this interference may then impair learning (see also Bailey & Plunkett, 2002 for a similar suggestion). This is essentially the same as the LEX model's account of this phenomenon.

Another finding that potentially highlights the role of exemplar representations in the lexicon concerns subsymbolic regularities in the mapping between linguistic form and meaning. Jurafsky, Bell, and Girand (2002) studied the pronunciation, in a spoken language corpus, of a number of English function words. They tested in particular whether different semantic senses of these words received different pronunciations – and found that in some cases, they did. For instance, they found that the word "of" tends to lose its final /v/ significantly more often when the word is used in its partitive sense – picking out a subpart of a larger quantity, e.g. "one of them", "cup of coffee" – than when it is used in its genitive sense – e.g. "friend of mine". This effect was not attributable merely to the predictability of this word in the context of neighboring words, but rather seemed to indicate a genuine underlying difference in pronunciation for different word senses. The authors suggest that an exemplar-based model of the lexicon may accommodate these results (Pierrehumbert, 2001; see also Johnson, 1997). On such a model, the two different word senses would be associated with two distinct subclusters of forms: one corresponding to the reduced form, and another to the non-reduced form. No model that assumed a single symbolic representation for a word would be able to account for such a finding, since the form-meaning regularity is at a level of representation more fine-grained than the word. Goldinger (1998) provided further support for an exemplar-based view of the lexicon. He found that an exemplar-based model of phonological representation in the lexicon provided a good overall fit to speech production data from a word-shadowing task. He found in particular that shadowers spontaneously imitated details of the acoustic patterns they heard. This fine-grained "echoing" of word utterances is compatible with exemplar-based theories of the lexicon, but incompatible with theories based on coarser-grained phonological representations.

LEX assumes that one-year-olds' attentional changes in form are driven by meaning. This may or may not be the case. Language-specific phonological organization begins well before children start attaching meanings to words (Werker & Tees, 1984), and the phonological changes in one-year-olds may be an extension of this earlier semantics-free learning – unlike the meaning-driven learning in LEX. However, even if this is the case, the model's larger message might still hold. Children's improvements in word-learning may still be due to a reduction in memory interference, due to increasing attention to relevant dimensions of form – even if the attentional shifts are driven

by bottom-up phonological processes, rather than meaning.

Limitations.

The LEX model has a number of limitations, beyond those mentioned above. One very general limitation is that it provides only a *qualitative* explanation of children's improvements in word-learning. A more quantitative, realistic, account would need to address several issues. The use of fixed-length bit vectors of artificial form and meaning features is convenient but unrealistic: these do not easily accommodate word forms of arbitrary length, nor do they easily accommodate word meanings based on conceptual frames involving the meanings of other words (Fillmore, 1975; Fillmore, 1982). Another shortcoming is that the simulations rely on a training set of only 50 words. A larger training set would be more appropriate since the simulations are intended to model children who *begin* to improve in word-learning when they have roughly 50 words in production – and who then continue to learn more words. Another simplification is that the simulations begin with attention equally distributed across all dimensions of form and meaning. Yet children learn a good deal about the sounds of their native language before beginning to speak, and infants as young as 12-13 months already have expectations about what sorts of meanings words will assume (Waxman & Markow, 1995).

It is also a serious simplification to assume, as LEX does, that form and referent are unambiguously specified. In actuality, children must often segment both form and referent from perceptual streams (although see Brent & Siskind, 2001). And in many cases, the word's referent is not even present when the word is uttered (Gleitman, 1990). To accommodate these issues, LEX would have to be supplemented with a mechanism that would segment word forms out of a speech stream, and also mentally supply possible referents, together with some means for determining which is likely to be intended. Some of these issues, which lie outside the scope of LEX, are addressed by Roy and Pentland (2002), Siskind (1996), and Yu et al. (2003).

A final limitation concerns the generalization of newly-learned words. If children are shown an object, and are told once that it is a "dax", they infer that "dax" may be a somewhat broad category, possibly encompassing objects of intermediate similarity to the original. But if they are taught that *three* very similar objects are all daxes, they infer that "dax" is a narrow category, encompassing the three objects they have seen, but not other objects of intermediate similarity to them. This finding is easily explained by Tenenbaum and Xu's (2000) Bayesian model, but not as easily by LEX. LEX also does not allow multiple conceptual perspectives on a single object (Clark, 1997) – although children do adopt multiple perspectives in naming. To accommodate both these findings, LEX would need to be extended to allow rapid attentional shifting across dimensions of meaning – something for which we have already seen a need, in discussing meaning biases such as the shape bias.

Despite these limitations, LEX does suggest a central idea: that learned shifts in attention in both form and meaning may reduce memory interference, and improve learning. There may be other instantiations of this general notion – instantiations that explain phenomena that LEX cannot, as well as those it can. If this work can help lead to such a model, it will have served its purpose.

Associations and reference.

Earlier work (Lock, 1980; McShane, 1979) proposed that the first stages of word-learning may be a matter of mere associative learning, but that the subsequent improvement in word-learning reflects a conceptual insight into the referential nature of words. In contrast, the central message of the current work is that four parallel early improvements in word-learning can be explained in a

unified manner through associative learning.

For reasons sketched at the beginning of this paper, it seems clear that associative learning can be only a partial explanation of word-learning. There are two further reasons, reinforcing this conclusion. The first, already touched on briefly above, is that associative learning appears in many other species (and in prelinguistic infants: Barr, Vieira, & Rovee-Collier, 2002) – but word-learning does not. Tomasello and Akhtar (2000) raise exactly this objection, arguing that word-learning cannot be associative, and that it relies instead on social sensitivities that are largely unique to humans – including the ability to understand referential intent, and the motivation to share experiences with others (see also Tomasello et al., in press). The second reason is that language *is* inescapably referential. A word such as "dog" acts much as a referential pointing gesture might, as a means for a speaker to coordinate attention with a hearer, such that they jointly attend to a dog, or to dogs generally. If the word were merely a bidirectional association between the sound "dog" and its meaning, of the sort captured by LEX, there would be no sense in which "dog" stands for a kind of animal – rather than the kind of animal standing for the sound "dog". Thus, the *directionality of reference* is absent on a strictly associationist view. So if associative learning accounts for some aspects of word-learning (as argued here) but not others, and if social abilities are also centrally involved (as appears to be the case) – how might these two forces work together?

One possibility is suggested by LEX's structure. The model assumes that form and referent have been segregated from each other, so these two entities may then be associated. But it is not clear how form and referent are separated in the mind of a child – after all, a word's utterance and its referent are both simply events or states in the surrounding world. An appealing possibility is that the child may rely on social cues for this fundamental distinction. Specifically, the child may take the object of the interlocutor's *attention* as a potential referent (Yu et al., 2003). There is evidence supporting this idea. Baldwin et al. (1996) found that 15-20 month olds, who were attending to a novel object, could learn a novel word for that object only when the word was uttered by a speaker who was also attending to the object – and the child could see that the speaker was attending. This suggests that children follow the attention of their interlocutors, to socially highlight potential referents, picking them out for association with word forms. By 12 months of age if not earlier, children also show some understanding of the *intentions* of others (Phillips, Wellman, & Spelke, 2002; Woodward, Sommerville, & Guajardo, 2001) – thus, children could take the intentional actions of others to be the set of potential word forms. This would include verbal utterances, gestures (for sign language), as well as incidental actions such as scratching one's ear. The child would have to learn to focus on the communicatively relevant material within this class.

While speculative, this proposal does suggest a way in which social cues and associations might work together in word-learning. Social cues may highlight potential forms and referents, and the directionality of reference may be derived from the direction of the speaker's attention: toward the intended referent. Associative learning may then link form to referent in memory. Critically, no sudden referential insight is needed to explain the acceleration of word-learning in the second year of life. One possibility, however, is that a related insight occurs earlier, and plays a different role. Children begin detecting language-relevant social cues at around 12 months, and this "social awakening" may initiate the process of word-learning (Tomasello, 1999). Once this process has started, it may gradually bring about learned shifts of attention, resulting in a clear improvement

in word-learning several months later – for the reasons described here.

Acknowledgments

The ideas behind this work grew out of discussions with William Merriman, Amanda Woodward, Linda Smith, and Michael Gasser. Programming and testing of the model were done in part by David Burkett, Rachael Cabasaan, Bryce Corrigan, John O'Leary, and Mingyu Zheng. Thanks to Susanne Gahl, Susan Goldin-Meadow and David Burkett for helpful comments on an earlier draft of this paper, and to Susan Carey, Eve Clark, Brian MacWhinney, Kim Plunkett, and two anonymous reviewers for their comments. This work was supported by grant DC03384 from the National Institutes of Health.

Address correspondence to: Terry Regier, Department of Psychology, University of Chicago, 5848 S. University Ave., Chicago, IL 60637, phone: 773-702-0918, email: t-regier@uchicago.edu

I dedicate this paper to my 7-month-old son Paul, in fond anticipation of the emergence of his words.

References

Bailey, T. M. & Plunkett, K. (2002). Phonological specificity in early words. *Cognitive Development*, 17:1265–1282.

Baldwin, D., Markman, E., Bill, B., Desjardins, R., & Irwin, J. (1996). Infants' reliance on a social criterion for establishing word-object relations. *Child Development*, 67:3135–3153.

Ballem, K. & Plunkett, K. Phonological specificity in children at 1;2. *Journal of Child Language*. In press.

Barr, R., Vieira, A., & Rovee-Collier, C. (2002). Bidirectional priming in infants. *Memory and Cognition*, 30:246–255.

Bates, E. & MacWhinney, B. (1989). Functionalism and the competition model. In MacWhinney, B. & Bates, E., editors, *The Crosslinguistic Study of Sentence Processing*, pages 3–73, Cambridge. Cambridge University Press.

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

Bloom, P. (2000). *How Children Learn the Meanings of Words*. MIT Press, Cambridge, MA.

Bloom, P. & Markson, L. (2001). Are there principles that apply only to the acquisition of words? A reply to Waxman and Booth. *Cognition*, 78(1):89–90.

Booth, A. E. & Waxman, S. R. (2002). Word learning is 'smart': Evidence that conceptual information affects preschoolers' extension of novel words. *Cognition*, 84:B11–B22.

Brent, M. & Siskind, J. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, 81:B33–B44.

Brent, M. R. (1996). Advances in the computational study of language acquisition. *Cognition*, 61:1–38.

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

Clark, E. (1987). The principle of contrast: A constraint on language acquisition. In MacWhinney, B., editor, *Mechanisms of Language Acquisition*, pages 1–33, Hillsdale, NJ. Lawrence Erlbaum.

Clark, E. (1997). Conceptual perspective and lexical choice in acquisition. *Cognition*, 64:1–37.

Clark, E. (2004). Pragmatics and language acquisition. In Horn, L. R. & Ward, G., editors, *Handbook of pragmatics*, pages 562–577, Oxford. Blackwell.

Cottrell, G. & Plunkett, K. (1994). Acquiring the mapping from meaning to sounds. *Connection Science*, 6(4):379–412.

Dickinson, D. K. (1984). First impressions: Children's knowledge of words gained from a single exposure. *Applied Psycholinguistics*, 5(4):359–373.

Diesendruck, G., Markson, L., & Bloom, P. (2003). Children's reliance on creator's intent in extending names for artifacts. *Psychological Science*, 14:164–168.

Dollaghan, C. A. (1987). Fast mapping in normal and language-impaired children. *Journal of Speech and Hearing Disorders*, 52(3):218–222.

Dromi, E. (1987). *Early Lexical Devlopment*. Cambridge University Press, New York.

Dy, J. G. & Brodley, C. E. (2004). Feature selection for unsupervised learning. *Journal of Machine Learning Research*, 5:845–889.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, 48:71–99.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking Innateness: A Connectionist Perspective on Development*. MIT Press, Cambridge, MA.

Farkas, I. & Li, P. (2001). A self-organizing neural network model of the acquisition of word meaning. In Altmann, E., editor, *Proceedings of the Fourth International Conference on Cognitive Modeling*, pages 67–72, Mahwah, NJ. Lawrence Erlbaum Associates.

Fenson, L., Dale, P., Reznick, J., Bates, E., Thal, D., & Pethick, S. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, 59(5).

Fillmore, C. (1975). An alternative to checklist theories of meaning. In et al., C. C., editor, *Proceedings of the First Annual Meeting of the Berkeley Linguistics Society*, pages 123–131, Berkeley. Berkeley Linguistics Society.

Fillmore, C. (1982). Frame semantics. In of Korea, L. S., editor, *Linguistics in the Morning Calm*, pages 111–138. Hanshin, Seoul.

Fisher, C., Gleitman, L., & Gleitman, H. (1991). On the semantic content of subcategorization frames. *Cognitive Psychology*, 23:331–392.

Ganger, J. & Brent, M. (2004). Reexamining the vocabulary spurt. *Developmental Psychology*, 40(4):621–632.

Gasser, M. & Smith, L. B. (1998). Learning nouns and adjectives: A connectionist account. *Language and Cognitive Processes*, 13(2-3):269–306.

Gershkoff-Stowe, L. & Smith, L. B. (1997). A curvilinear trend in naming errors as a function of early vocabulary growth. *Cognitive Psychology*, 34:37–71.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, 1:3–55.

Gluck, M. & Bower, G. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, 117(3):227–247.

Goldinger, S. D. (1998). Echoes of echoes? an episodic theory of lexical access. *Psychological Review*, 105(2):251–279.

Gupta, P. & MacWhinney, B. (1997). Vocabulary acquisition and verbal short-term memory: Computational and neural bases. *Brain and Language*, 59(2):267–333.

Hare, B., Brown, M., Williamson, C., & Tomasello, M. (2002). The domestication of social cognition in dogs. *Science*, 298:1634–1636.

Heibeck, T. & Markman, E. (1987). Word learning in children: An examination of fast mapping. *Child Development*, 58(4):1021–1034.

Hespos, S. & Spelke, E. (2004). Conceptual precursors to language. *Nature*, 430:453–456.

Huttenlocher, J. (1974). The origins of language comprehension. In Solso, R. L., editor, *Theories in Cognitive Psychology*, pages 331–368, Hillsdale, NJ. Lawrence Erlbaum Associates.

James, W. (1890). Association. In *Psychology (Briefer course)*, chapter XVI, pages 253–279. Holt, New York.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In Johnson, K. & Mullennix, J. W., editors, *Talker Variability in Speech Processing*, pages 145–165. Academic Press, San Diego.

Jurafsky, D., Bell, A., & Girand, C. (2002). The role of the lemma in form variation. In Warner, N. & Gussenhoven, C., editors, *Papers in Laboratory Phonology 7*, pages 1–34. Mouton de Gruyter, Berlin/New York.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In Campbell, B. A. & Church, R. M., editors, *Punishment and aversive behavior*, pages 279–296, New York. Appleton-Century-Crofts.

Kaminski, J., Call, J., & Fischer, J. (2004). Word learning in a domestic dog: Evidence for "fast mapping". *Science*, 304:1682–1683.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99:22–44.

Landau, B., Smith, L. B., & Jones, S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, 3:299–321.

Landau, B., Smith, L. B., & Jones, S. (1998). Object shape, object function, and object name. *Journal of Memory and Language*, 38(1):1–27.

Laska, M. & Metzker, K. (1998). Food avoidance learning in squirrel monkeys and common marmosets. *Learning and Memory*, 5(3):193–203.

Li, P. & Farkas, I. (2002). Modeling the development of lexicon with DevLex: A self-organizing neural network model of lexical acquisition. In Gray, W. & Schunn, C., editors, *Proceedings of the Twenty-fourth Annual Meeting of the Cognitive Science Society.*

Liittschwager, J. & Markman, E. (1994). Sixteen and 24-month-olds' use of mutual exclusivity as a default assumption in second label learning. *Developmental Psychology*, 30:955–968.

Lock, A. (1980). *The Guided Reinvention of Language.* Academic Press, London.

Lucy, J. (1992). *Grammatical categories and cognition: A case study of the linguistic relativity hypothesis.* Cambridge University Press, Cambridge.

Mackintosh, N. J. (1965). Selective attention in animal discrimination learning. *Psychological Bulletin*, 64(2):124–150.

MacWhinney, B. (1987). The competition model. In MacWhinney, B., editor, *Mechanisms of Language Acquisition*, pages 249–308, Hillsdale, NJ. Lawrence Erlbaum.

MacWhinney, B. (1989). Competition and lexical categorization. In Corrigan, R., Eckman, F., & Noonan, M., editors, *Linguistic Categorization*, number 61 in Current Issues in Linguistic Theory, pages 195–242, New York. John Benjamins.

Markman, E. & Wachtel, G. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, 20:121–157.

Markson, L. & Bloom, P. (1997). Evidence against a dedicated system for word-learning in children. *Nature*, 385:813–815.

Martin, N., Gagnon, D., Schwartz, M., Dell, G., & Saffran, E. (1996). Phonological facilitation of semantic errors in normal and aphasic speakers. *Language and Cognitive Processes*, 11:257–282.

McCloskey, M. & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In Bower, G. H., editor, *The Psychology of Learning and Motivation*, volume 24, pages 109–165. Academic Press, San Diego.

McShane, J. (1979). The development of naming. *Linguistics*, 17:879–905.

Merriman, W. (1999). Competition, attention, and young children's lexical processing. In MacWhinney, B., editor, *The Emergence of Language*, pages 331–358. Lawrence Erlbaum Associates, Mahwah, NJ.

Merriman, W. & Bowman, L. (1989). The mutual exclusivity bias in children's word learning. *Monographs of the Society for Research in Child Development*, 54(3-4).

Mervis, C. B., Golinkoff, R. M., & Bertrand, J. (1994). Two-year-olds readily learn multiple labels for the same basic-level category. *Child Development*, 65:1163–1177.

Miikkulainen, R. (1997). Dyslexic and category-specific aphasic impairments in a self-organizing feature map model of the lexicon. *Brain and Language*, 59(2):334–366.

Niyogi, S. (2002). Bayesian learning at the syntax-semantics interface. In Gray, W. & Schunn, C., editors, *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society*, pages 697–702, Mahwah, NJ. Lawrence Erlbaum Associates.

Nosofsky, R. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115(1):39–57.

Pavlov, I. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex*. Oxford University Press, London.

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review*, 94(1):61–73.

Pearce, J. M. (1994). Similarity and discrimination: A selective review and a connectionist model. *Psychological Review*, 101(4):587–607.

Phillips, A., Wellman, H., & Spelke, E. (2002). Infants' ability to connect gaze and emotional expression to intentional action. *Cognition*, 85:53–78.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In Bybee, J. & Hopper, P., editors, *Frequency and the emergence of linguistic structure*, pages 137–158. John Benjamins, Amsterdam.

Plaut, D. C. (1999). A connectionist approach to word reading and acquired dyslexia: Extension to sequential processing. *Cognitive Science*, 23(4):543–568.

Plunkett, K. & Marchman, V. (1988). U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition*, 28:73–193.

Plunkett, K., Sinha, C., Møller, M., & Strandsby, O. (1992). Symbol grounding or the emergence of symbols? vocabulary growth in children and a connectionist net. *Connection Science*, 4:293–312.

Regier, T. (1996). *The Human Semantic Potential: Spatial Language and Constrained Connectionism*. MIT Press, Cambridge, MA.

Regier, T. (2003). Emergent constraints on word-learning: a computational review. *Trends in Cognitive Sciences*, 7:263–268.

Regolin, L. & Rose, S. (1999). Long-term memory for a spatial task in young chicks. *Animal Behaviour*, 57(6):1185–1191.

Rescorla, R. A. & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In Black, A. H. & Prokasy, W. F., editors, *Classical Conditioning II: Current Research and Theory*, New York. Appleton-Century-Crofts.

Rohde, D. & Plaut, D. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, 72(1):67–109.

Roy, D. K. & Pentland, A. P. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, 26(1):113–146.

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning internal representations by error propagation. In Rumelhart, D., McClelland, J., & the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, volume 1*, pages 318–362. MIT Press, Cambridge, MA.

Rumelhart, D. & McClelland, J. (1986). On learning the past tenses of English verbs. In McClelland, J., Rumelhart, D., & the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, volume 2*, pages 216–271. MIT Press, Cambridge, MA.

Sahley, C., Gelperin, A., & Rudy, J. W. (1981). One-trial associative learning modifies food odor preferences of a terrestrial mollusc. *Proceedings of the National Academy of Sciences of the USA*, 78(1):640–642.

Schafer, G. & Mareschal, D. (2001). Modeling infant speech sound discrimination using simple associative networks. *Infancy*, 2(1):7–28.

Schafer, G. & Plunkett, K. (1998). Rapid word learning by fifteen-month-olds under tightly controlled conditions. *Child Development*, 69(2):309–320.

Siskind, J. (1992). *Naive Physics, Event Perception, Lexical Semantics and Language Acquisition.* PhD thesis, Massachusetts Institute of Technology.

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

Smith, L. B. (1989). A model of perceptual classification in children and adults. *Psychological Review*, 96:125–144.

Smith, L. B. (2000). Learning how to learn words: An associative crane. In Golinkoff, R., Hirsh-Pasek, K., Bloom, L., Smith, L., Woodward, A., Akhtar, N., Tomasello, M., & Hollich, G., editors, *Becoming a Word Learner: A Debate on Lexical Acquisition*, pages 51–80. Oxford University Press, New York.

Smith, L. B., Jones, S., & Landau, B. (1996). Naming in young children: A dumb attentional mechanism? *Cognition*, 60:143–171.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, 13(1):13–19.

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

Sutton, R. S. & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 88:135–170.

Swingley, D. & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, 76(2):147–166.

Swingley, D. & Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science*, 13(5):480–484.

Tenenbaum, J. & Xu, F. (2000). Word learning as Bayesian inference. In Gleitman, L. & Joshi, A., editors, *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*, pages 517–522, Mahwah, NJ. Lawrence Erlbaum Associates.

Theodoridis, S. & Koutroumbas, K. (1999). *Pattern Recognition*. Academic Press, San Diego, CA.

Thompson, C. & Mooney, R. (2003). Acquiring word-meaning mappings for natural language interfaces. *Journal of Artificial Intelligence Research*, 18:1–44.

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Harvard University Press, Cambridge, MA.

Tomasello, M. & Akhtar, N. (2000). Five questions for any theory of word learning. In Golinkoff, R. M., Hirsh-Pasek, K., Bloom, L., Smith, L. B., Woodward, A. L., Akhtar, N., Tomasello, M., & Hollich, G., editors, *Becoming a Word Learner: A Debate on Lexical Acquisition*, pages 179–186. Oxford University Press, New York.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*. In press.

Waxman, S. R. & Booth, A. E. (2000). Principles that are invoked in the acquisition of words, but not facts. *Cognition*, 77(2):B33–B43.

Waxman, S. R. & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology*, 29(3):257–302.

Werker, J., Cohen, L., Lloyd, V., Casasola, M., & Stager, C. (1998). Acquisition of word-object associations by 14-month-old infants. *Developmental Psychology*, 34(6):1289–1309.

Werker, J., Fennell, C., Corcoran, K., & Stager, C. (2002). Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. *Infancy*, 3(1):1–30.

Werker, J. F. & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7(1):49–63.

Woodward, A., Sommerville, J., & Guajardo, J. (2001). How infants make sense of intentional action. In Malle, B. & Moses, L., editors, *Intentions and intentionality: Foundations of social cognition*, pages 149–169. MIT Press, Cambridge, MA.

Woodward, A. L. & Hoyne, K. L. (1999). Infants' learning about words and sounds in relation to objects. *Child Development*, 70(1):65–77.

Woodward, A. L., Markman, E. M., & Fitzsimmons, C. M. (1994). Rapid word learning in 13- and 18-month-olds. *Developmental Psychology*, 30:553–566.

Yu, C., Ballard, D., & Aslin, R. (2003). The role of embodied intention in early lexical acquisition. In Alterman, R. & Kirsh, D., editors, *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society*, Mahwah, NJ. Lawrence Erlbaum Associates.

Appendix A. Gradient descent.

The model is trained under gradient descent in the error quantity E. That is, we adjust each free parameter w according to:

$$\Delta w = -\eta \frac{\partial E}{\partial w} \tag{11}$$

η, the learning rate, is a small positive constant.

The core of gradient descent is the computation of the gradient itself, $\frac{\partial E}{\partial w}$. We employ two methods of gradient computation, for the two sorts of weights in the model. For attention weights we numerically approximate the derivative – this method has the advantages of simplicity and flexibility. For associative weights, we compute the derivative analytically (see Kruschke, 1992, for analytic derivatives over exemplar-based structures such as those employed here), since the number of associative weights makes approximation prohibitively time-expensive. We treat these two techniques in turn.

Approximation of gradient - for attention weights

Attention weights for form are driven by meaning error E_m, while attention weights for meaning are driven by form error E_f. Here, we consider the approximation of the gradient for E_m; the gradient for E_f is computed analogously.

The partial derivative $\frac{\partial E_m}{\partial w}$ is defined as:

$$\frac{\partial E_m}{\partial w} = \lim_{\delta \to 0} \frac{E_m(w+\delta) - E_m(w)}{\delta} \tag{12}$$

where $E_m(w)$ is the value of the meaning error when the free parameter in question has value w, and $E_m(w+\delta)$ is the value of that error when the free parameter has been incremented by δ, to $w+\delta$, with all other free parameters held constant. We approximate this derivative by choosing a particular small perturbation value for δ:

$$\frac{\partial E_m}{\partial w} \approx \frac{E_m(w+\delta) - E_m(w)}{\delta}. \tag{13}$$

Thus, in order to obtain the partial derivative, we compute the error twice: once with w at its current value, and once with w slightly perturbed. In our simulations, we set the perturbation value δ to be .000001.

Analytic form of gradient - for associative weights

Here we consider the analytic form for the gradient of meaning error $\frac{\partial E_m}{\partial w}$; the gradient of form error $\frac{\partial E_f}{\partial w}$ is computed analogously. Associative weights w are then driven by $\frac{\partial E}{\partial w} = \frac{\partial E_f}{\partial w} + \frac{\partial E_m}{\partial w}$.

Let w_{rs} be the associative weight on the connection between a sending (form) node s and a receiving (meaning) node r. Then:

$$\frac{\partial E_m}{\partial w_{rs}} = \frac{\partial E_m}{\partial net_r} \frac{\partial net_r}{\partial w_{rs}} = \frac{\partial E_m}{\partial net_r} a_s \tag{14}$$

where a_s is the activation of the sending node, and net_r is the net input of the receiving node. The error E_m is:

$$E_m = 1.0 - \sum_i g_i p_i \tag{15}$$

where i indexes over all meaning exemplar nodes, g_i is a Gaussian function of the psychological distance between meaning exemplar node i and the teacher-supplied referent (determined as in Equations 2 and 3, except it is now the teacher pattern, rather than the input pattern, from which distance is measured), and

$$p_i = \frac{net_i}{(\sum_j net_j) + noise} = \frac{net_i}{denom} \tag{16}$$

is the output probability for node i. Here, j indexes meaning exemplar nodes, and net_j is the net input received at meaning exemplar node j, over associative links from form nodes. We use the expression *denom* for the denominator of this formula, as it is a quantity that will recur frequently:

$$denom = (\sum_j net_j) + noise. \tag{17}$$

Given Equation 14, we need only determine $\frac{\partial E_m}{\partial net_r}$.

$$\frac{\partial E_m}{\partial net_r} = \frac{\partial [1.0 - \sum_i g_i p_i]}{\partial net_r} = \frac{\partial [-\sum_i g_i \frac{net_i}{denom}]}{\partial net_r} = [-\sum_{i \neq r} g_i \frac{\partial \frac{net_i}{denom}}{\partial net_r}] - g_r \frac{\partial \frac{net_r}{denom}}{\partial net_r} \tag{18}$$

In general, the derivative of a quotient is:

$$(\frac{f}{g})'(x) = \frac{(g(x)f'(x)) - (g'(x)f(x))}{g(x)^2}. \tag{19}$$

This means that

$$\frac{\partial \frac{net_{i \neq r}}{denom}}{\partial net_r} = \frac{(denom \times 0) - (1 \times net_i)}{denom^2} = \frac{-net_i}{denom^2}, \tag{20}$$

and similarly, that

$$\frac{\partial \frac{net_r}{denom}}{\partial net_r} = \frac{(denom \times 1) - (1 \times net_r)}{denom^2} = \frac{denom - net_r}{denom^2}. \tag{21}$$

Thus, continuing from Equation 18:

$$\begin{aligned} \frac{\partial E_m}{\partial net_r} &= [-\sum_{i \neq r} g_i \frac{\partial \frac{net_i}{denom}}{\partial net_r}] - g_r \frac{\partial \frac{net_r}{denom}}{\partial net_r} \\ &= [\sum_{i \neq r} g_i \frac{net_i}{denom^2}] - g_r \frac{denom - net_r}{denom^2} \\ &= \frac{1}{denom^2}[(\sum_{i \neq r} net_i g_i) - (g_r(denom - net_r))] \\ &= \frac{1}{denom^2}[(\sum_i net_i g_i) - (g_r denom)] \end{aligned} \tag{22}$$

This, together with Equation 14, yields:

$$\frac{\partial E_m}{\partial w_{rs}} = \frac{a_s}{denom^2}[(\sum_i net_i g_i) - (g_r denom)]. \tag{23}$$

This is the formula used in gradient descent for associative weights.

Appendix B. Testing a newly learned word.

One may test either production or comprehension of a newly learned word.

Production

A referent is presented at the model's meaning inputs, producing a probability distribution over form exemplars. The probability of selecting an appropriate form for that referent is then estimated as:

$$p(correct) = \sum_i g_i p_i. \tag{24}$$

Here i ranges over form exemplar nodes, p_i is the probability of selecting exemplar i, and g_i is a Gaussian function of the psychological distance, in form space, between exemplar i and the form supplied by the teacher (determined as in Equations 2 and 3, except it is now the teacher pattern, rather than the input pattern, from which distance is measured). This estimates the probability of a correct response, by analogy with Equation 9.

Comprehension

Testing comprehension in word-learning experiments is often done by uttering a newly-learned word for a target referent, and seeing if the child will choose that target referent from among a set of distractors. To model this testing process, let t be the target referent, and $d1$ and $d2$ the two distractor referents. We begin by pairing each of these referents with the newly-learned word form, resulting in three referent-form pairings. The goodness of fit of a particular referent-form pairing j – the evidence supporting it – is:

$$e_j = \sum_{i \in M} g_i p_i + \sum_{i \in F} g_i p_i \tag{25}$$

Here, M is the set of all meaning exemplar nodes, and F is the set of all form exemplar nodes. For $i \in M$, p_i is determined by providing the word form as input and the referent as target, and following Equations 2 through 5, while g_i is as in Equation 9. For $i \in F$, p_i and g_i are determined by providing the referent as input, taking the form as target, and following the analogous computations in the other direction. The probability of a correct response is then the probability of selecting the target referent t:

$$p(correct) = \frac{e_t + (residual/3)}{\sum_{i \in (t,d1,d2)} e_i + residual} \tag{26}$$

where $residual$ is the amount of remaining possible evidence that was not realized in any of the three form-meaning pairings:

$$residual = 2.0 - max_{j \in (t,d1,d2)} e_j. \tag{27}$$

The inclusion of the $residual$ term prevents the possibility of division by zero. It also captures the intuition that the probability of choosing the target referent should be a function not only of the relative strengths of evidence for each of the three referents, but also of the absolute strength of evidence for the target. Under this scheme, $p(correct) = 1$ only if the evidence for the distractors is zero, and the evidence for the target is maximized at 2.0.

Appendix C. Learning and memory interference.

Principle: Learning of a novel word is most effective when memory interference is minimized.

Demonstration: Learning is driven by gradient descent in error, as described in Appendix A. Learning of a novel word is most effective when the weight change Δw – on the connection between that word's form and meaning exemplar nodes – is maximized. "Memory interference" refers to a situation in which there are already exemplars activated at the output – that is, $\exists j$ such that $net_j > 0$. Thus, it suffices to show that the weight change Δw is maximized, and strongly positive, when net_j is minimized, $\forall j$.

The update rule for associative weights may be decomposed into form and meaning components:

$$\Delta w_{rs} = -\eta \frac{\partial E}{\partial w_{rs}} = -\eta \frac{\partial E_f}{\partial w_{rs}} + -\eta \frac{\partial E_m}{\partial w_{rs}} \quad = \Delta w^f_{rs} + \Delta w^m_{rs}. \tag{28}$$

We consider the case of weight change due to meaning error, Δw^m_{rs}. The analogous argument holds for weight change due to form error, Δw^f_{rs}.

$$\begin{aligned} \Delta w^m_{rs} &= -\eta \tfrac{\partial E_m}{\partial w_{rs}} \\ &= -\eta \tfrac{a_s}{denom^2}[(\textstyle\sum_i net_i g_i) - (g_r denom)] \\ &= \eta \tfrac{a_s}{denom}[g_r - \tfrac{\sum_i net_i g_i}{denom}]. \end{aligned} \tag{29}$$

following Equation 23 in Appendix A. Variables are defined as in Appendix A. In this formula:

1. η is a positive constant, and is unaffected by net_j.
2. The multiplier $\frac{a_s}{denom}$ reaches its maximum value of $\frac{a_s}{noise} > 0$ when $net_j = 0, \forall j$. As the various net_j grow, this term approaches zero.
3. g_r is positive, and unaffected by net_j.
4. The subtracted quantity $\frac{\sum_i net_i g_i}{denom}$ reaches its minimum value of $\frac{0}{noise} = 0$ when $net_j = 0, \forall j$. As the various net_j grow, this term becomes substantially greater than zero.

Therefore, the weight update as a whole is maximized when net_j is minimized at 0, $\forall j$. Thus, learning on an associative link is strongest when memory interference is minimized.