Learning Nouns with Domain-General Associative Learning Mechanisms

George Kachergis
gkacherg@indiana.edu
Department of Psychological & Brain Science / Cognitive Science Program
Bloomington, IN 47405 USA

Abstract

Associative learning has been meticulously studied in many species, and diverse effects have been explained using a handful of basic assumptions and mechanisms. Human language acquisition proceeds remarkably quickly and is of great interest, but is arguably more difficult to capture under the microscope. Nonetheless, empirical investigations have led researchers to theorize a variety of language learning principles and constraints. While there may indeed be language-specific learning mechanisms that are distinct from more universal associative learning mechanisms, we seek to explain some basic principles of language acquisition using domain-general mechanisms. Using an experiment and a model, we show how the principles of mutual exclusivity—an assumption of 1-to-1 word-object mappings, contrast, and other constraints related to fast mapping may stem from attention mechanisms attributed to associative learning effects such as blocking and highlighting, but directed by competing biases for familiar and unfamiliar pairs instead of surprise.

Keywords: statistical learning; language acquisition; cross-situational learning; associative learning; attention

Introduction

All organisms learn, but only humans master human languages. Since many neural structures and basic learning mechanisms are conserved across species, it bears asking how much of human language learning can be explained with domain-general mechanisms, without appealing to innate (i.e., evolved) linguistic knowledge, exemplified by the work of Noam Chomsky, or domain-specific principles and constraints, whether innate or developed early in life (e.g., Markman, 1992).

One essential part of language learning is learning word-object mappings—nouns. Two border collies have been shown to learn hundreds of nouns over years of training (Pilley & Reid, 2011; Kaminski, Call, & Fischer, 2004). Of course, this feat pales in comparison to human language learning: infants begin producing words at 1 year, and by the end of high school have command of 60,000 words, conservatively (Bloom, 2000). However, both dogs and infants have been shown to fast map: given a new word, they will choose a new object over an object with a known label, and retain the mapping weeks later (see Bloom, 2000). Fast mapping is a powerful ability for word learning, but is it based on domain-general or domain-specific learning mechanisms?

One approach to studying language acquisition views word learning as a problem of induction with an enormous hypothesis space, and proposes a number of constraints to restrict the space (Markman, 1992). In this view, infants generate hypotheses that are consistent with this set of constraints and principles. The present paper is concerned with a subset of these principles that relate to how people map new words to objects.

Mutual exclusivity (ME) is the assumption that every object has only one name (Markman & Wachtel, 1988). A fill-the-lexical gap bias, which causes children to want to find a name for an object with no known name, has also been proposed (Clark, 1983) and argued (Merriman and Bowman, 1989). When given a set of familiar and unfamiliar objects, it has been shown that 28-month-olds assume that a new label maps to an unfamiliar object (e.g., Mervis & Bertrand, 1994). Similarly, the principle of contrast states that an infant given a new word will seek to attach it to an unlabeled object (Clark, 1983). Fill-the-gap, ME, and contrast make many of the same predictions made by the more general novel name-nameless category principle (N3C), which states that novel labels map to novel objects (Golinkoff, Mervis, & Hirsh-Pasek, 1994).

It is not our goal to explore the overlapping and nuanced ways that these various principles interact. Indeed, we hope to avoid this confusing plurality of explanations by showing that many of the behaviors ascribed to these theories can be explained by domain-general learning mechanisms uncovered by studies of associative learning. Nor are we the first to suggest that human language acquisition—as fast and yet difficult as it is—can be explained with domain-general learning mechanisms: Smith (2000) argued as much, and much recent work in statistical learning (described below) is motivated by this premise. Recent work has even found that children show a 1-to-1 bias in domains other than language: voices to faces (Moher, Feigenson, & Halberda, 2010) and actions to objects (Childers & Tomasello, 2003). However, few direct analogies have been drawn between the models and paradigms of word learning and associative learning, but see Ramscar et al. (2010). After introducing some associative learning paradigms and linking them to word learning, we discuss how universal attentional biases may account for many behaviors observed across domains. Finally, we report a new empirical word learning study using an associative learning highlighting design, and explain the results with a word-learning model that has competing attentional biases for familiarity and uncertainty.

Associative Learning

Associative learning paradigms typically present one or more perceptual cues (e.g., objects, sounds), learners make a response (e.g., a button press), and feedback is given (e.g., food, a shock). When one cue q_1 is paired with outcome o on each trial, the resulting q_1-o association is stronger than q_1-o when two simultaneous cues $\{q_1, q_2\}$ predict o during training; thus, q_2 is said to *overshadow* q_1 (Pavlov, 1927). A

reasonable way to explain overshadowing is that attention is split between the two cues, and thus the associations q_1-o and q_2-o grow more slowly than when q_1 appears alone. Attention is also used to explain the *blocking* effect (Kamin, 1968), which can be induced using a design with two training stages. In the early stage, cue q_1 is repeatedly paired with outcome o, and in the late stage q_1 and q_2 appear jointly preceding o. The association between q_2 and o is found to be much weaker than when only the late stage occurs. Thus q_2 has been blocked by q_1's earlier association with o—much like mutual exclusivity (ME) states that learners will not map a second label (q_2) to a known object (o). Learning models, updating knowledge trial-to-trial, account for blocking using selective attention to q_1: since q_1 already predicts o, there is no need to strengthen q_2-o (e.g., Rescorla & Wagner, 1972; Pearce & Hall, 1980). Is blocking found in word-learning experiments? Can ME be thought of as blocking? As it happens, two cross-situational word-learning studies can be seen to address these questions.

Cross-situational Word Learning

A key challenge in early word learning is to deal with the referential uncertainty intrinsic to complex scenes and utterances. Recent research has focused on how regularities in the co-occurrence of words and objects in the world can significantly reduce referential ambiguity across situations. Statistical word learning relies on two assumptions: 1) that spoken words are often relevant to the current situation, and 2) that learners can remember to some degree the co-occurrence of multiple words and objects in a scene. Thus, as the same words and objects are observed in different situations across time, people can learn the correct word-object mappings.

In adult cross-situational learning studies (e.g., Yu & Smith, 2007), participants are asked to learn the meaning of alien words from a series of training trials, each of which contains a few spoken words and a few objects. Although each word refers to a particular onscreen object, the intended referent is not indicated in any way, leaving meanings ambiguous on individual trials. Ichinco, Frank, and Saxe (2009) used a cross-situational word-learning task in which learners are first exposed to 1-to-1 pairings on a series of trials with four word-object pairs per trial. In the late stage, after people had presumably learned some of the mappings, a fifth object (or word, in another condition) began to consistently co-occur with one of the early words (or objects). The result was little learning of the association between old word (or object) and new object (or word) association, consistent with ME. However, this design can be seen to closely match a blocking design (see Table 1), with a few notable differences.

First, it is unclear whether words should be construed as cues and objects as outcomes, or the reverse—an issue we will return to. Second, a cross-situational trial has multiple outcomes, unlike associative learning paradigms. Finally, no trial-to-trial feedback is given, but the learner may generate it on the basis of the preceding training trials. We contend that none of these differences are a fundamental problem with seeing cross-situational learning as associative learning. Indeed, if anything the learning problem in the real world is more like cross-situational learning: with a multitude of stimuli that may simultaneously serve as either cues or outcomes for as many other stimuli, learners attempt to associate correlated stimuli.

Training Stage	Ichinco et al., 2009	Kamin, 1968
Early	$\{w_1, w_x, w_y, w_z\}$-$\{o_1, o_x, o_y, o_z\}$	q_1-o
Late	$\{w_1, w_x, w_x, w_z\}$-$\{o_1, o_2, o_x, o_y, o_z\}$	$\{q_1, q_2\}$-o

Table 1: Comparison of the blocking paradigm (right) with a cross-situational word learning paradigm (left). In both paradigms, the late-stage stimulus (q_2 / o_2) is blocked from becoming associated with the outcome (o / w_1), despite consistent co-occurrence in the late stage.

Thus, learners in the Ichinco et al. study may not learn the extra association (w_1-o_2) because attention remains focused on strengthening the still-present early-trained association (w_1-o_1). This attentional account is equivalent to the popular account for blocking, and is corroborated by an earlier result that defies ME: Yurovsky and Yu (2008) used a two-stage cross-situational design much like Ichinco et al., but in the late stage when adding a new stimulus to an existing association, removed the old object (or word). Faced with a word (w_1) they have associated with o_1, but now seeing o_2 without o_1 repeatedly, people learned the association, but also retained w_1-o_1 at test. Yurovsky & Yu's learners cast about for a new associate, unblocked by the presence of an old associate to attend to—unlike in Ichinco et al.'s study. In summary, by establishing an analogy of cross-situational learning as a complex associative learning paradigm, we found that two cross-situational studies can be explained with a domain-general selective attention mechanism, without recourse to a language-specific constraint such as ME. To further examine the role of attention in cross-situational learning, we do a word learning experiment using a design that in associative learning yields the interesting order effect of highlighting.

Experiment: Highlighting

Like blocking, highlighting is a learning order effect that has been attributed to selective attention (Medin & Edelson, 1988; Kruschke, 1996). In an early stage of training, a cues *PE* (Perfect Early) and *I* (Imperfect) jointly appear on each trial, followed by outcome *E* (Early). In a late stage, cue *I* appears with *PL* (Perfect Late), followed by outcome *L*. Thus, *I* imperfectly predicts both outcomes, having first predicted *E*, and later *L*. On the other hand, *PE* perfectly predicts *E*, and symmetrically, *PL* perfectly predicts *L*. As depicted in Figure 1, learners show an order effect: *PE* and *I* both become associated with *E* in the early stage, and then *PL* becomes more strongly linked with *L* while *I*-*PL* languishes. This is presumably because attention is shifted away from *I*, since it already predicts *E* in the early stage.

Formerly known as the inverse-base rate effect (note that *I* is twice as frequent as *PE* or *PL*), Kruschke (2009) presented a study with balanced frequency of the early and late training stages and still found highlighting, lending further credence to the attention account.

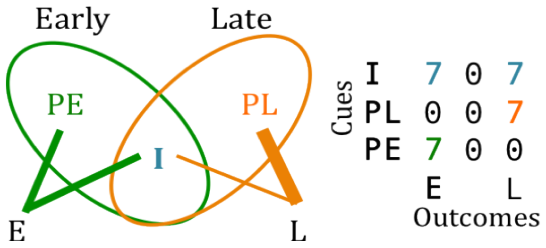

Figure 1: The co-occurrences of cues and outcomes in the highlighting design (right), and the estimated strength of associations between each cue and outcome (left), shown by the thickness of the lines.

Following a similar design, we use the Experiment to ask 1) whether highlighting occurs in a cross-situational framework with no explicit feedback on each trial, and 2) if words are cues and objects are outcomes, vice versa, or if they are interchangeable. As shown in Figure 2, this is done by making the cues in a highlighting design correspond to either words or objects, resulting in 2 words (cues) and 1 object (outcome) per trial, or 2 objects (cues) and 1 word (outcome) displayed per trial. Seeing a highlighting effect in one condition and not the other may suggest one correspondence over the other, whereas highlighting in both conditions suggests that words and objects can act as either cues or outcomes. Finally, finding no highlighting would suggest that domain-specific mechanisms may be at work in word learning.

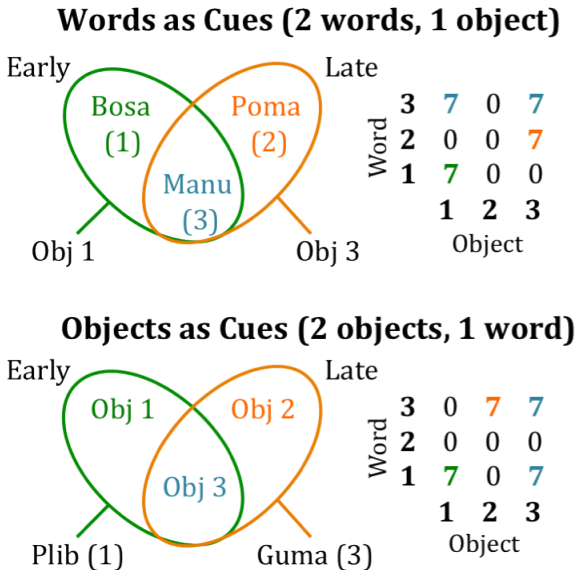

Figure 2: Highlighting designs in the Experiment (left), with word x object co-occurrence matrices (right). In the words as cues condition, 2 words and 1 object were given on each trial (top), while 2 objects and 1 word were given in the objects as cues condition (bottom).

Subjects

Participants were 67 undergraduates at Indiana University who received course credit for participating. None had previously participated in cross-situational experiments.

Stimuli & Procedure

Twelve pseudowords and 12 objects were randomly drawn from larger sets of stimuli, randomly paired, and split between the two conditions. The pseudowords (words) are phonotactically-probable in English (e.g., "bosa"), and were spoken by a monotone, synthetic female voice. The objects were photographs and drawings of uncommon objects (e.g., sculptures, specialty tools). Each training trial in the **words as cues** condition consisted of one object and two spoken words, while training trials in the **objects as cues** condition had two objects visible while one word was spoken. In both conditions, the object(s) remained visible for the duration of the trial. Each trial began with 2s of silence before the first 1s word was heard. In the words as cues condition, the second word was played after 1s of silence. In both conditions, the last word was followed by 3s of silence. In total, each trial in the objects as cues condition lasted 6s and trials in the words as cues condition lasted 7s.

Training for each condition consisted of 28 trials. The highlighting structures shown in Figure 2 were replicated within each condition: in words-as-cues, people heard six words and saw four objects, while in objects-as-cues, people heard four words and saw six objects. Knowledge was assessed after the completion of each condition using 6AFC testing: learners were asked to choose the best object for each of the six words. That is, we are probing the conditional probability objects, given a word. Note that in words-as-cues, two of the six objects available at test had not been seen during training, while in objects-as-cues, two words were never heard. These were not removed to keep the conditions symmetric, and in case systematic response deviations were found. Words were tested in random order. Note that the test in the words as cues condition corresponds most directly to associative learning testing: participants are given a cue (word) and asked to predict the outcome (object). In the objects as cues condition, we are actually asking learners to choose the best cue (object) when given an outcome (word). Participants completed both conditions in counterbalanced order.

Results & Discussion

Figure 3 displays the conditional probabilities of choosing each object[1], given each word, and the corresponding estimated relative strengths of each word-object association. The results in both conditions exhibit all the characteristics of highlighting: cue *I* is more strongly linked to *E* than *L*, and although *PE-E* and *PL-L* are both quite strong, *PL-L* is stronger. In the words as cues condition, object o_1 (*E*) was

[1] As noted before, there were two highlighting replications in each condition, so there were six objects available at test. Here we have collapsed the two replications for ease of presentation, and left out incorrect responses (e.g., choosing o_4, o_5, or o_6 for w_1, w_2, or w_3). The mean response probability for these cells is .08.

chosen significantly more than o_3 (*L*; .51 vs. .25) for word w_3 (*I*; $\chi^2(1,N=79) = 9.23$, $p<.01$). In the objects as cues condition, o_3 (*I*) was chosen significantly more often for w_1 (*E*) than w_3 (*L*; .28 vs. .16; $\chi^2(1,N=73) = 6.04$, $p=.01$). Thus, the early association of *I* with *E* kept *I* from becoming strongly associated with *L*—much like a mutual exclusivity constraint would keep people from associating a second word with an already-labeled object. Given words as cues, *L* (o_3) was chosen more often for *PL* (w_2) than *E* (o_1) was chosen for *PE* (o_1; .82 vs. .69), though the difference was not significant ($\chi^2(1,N=157) = 1.08$, $p=.30$). Similarly, given objects as cues, *PL* (o_2) was chosen more often for *L* (w_3) than *PE* (o_1) was chosen for *E* (w_1; .71 vs. .60), but again the difference was not significant ($\chi^2(1,N=215) = 1.68$, $p=.20$). Despite not being statistically significant[2], these conditional response rates match a highlighting result in both cases: *PL-L* is learned faster (stronger) because little attention is given to *I-L*, as cue *I* is already associated with *E*. In terms of word learning, this is much like the novel name-nameless category principle (N3C; Golinkoff et al., 1994): given a new object (or word—*PL*), it is reasonable to associate this with a new word (or object—*L*), rather than a word (or object—*PE*) with an already-known associate (*E*).

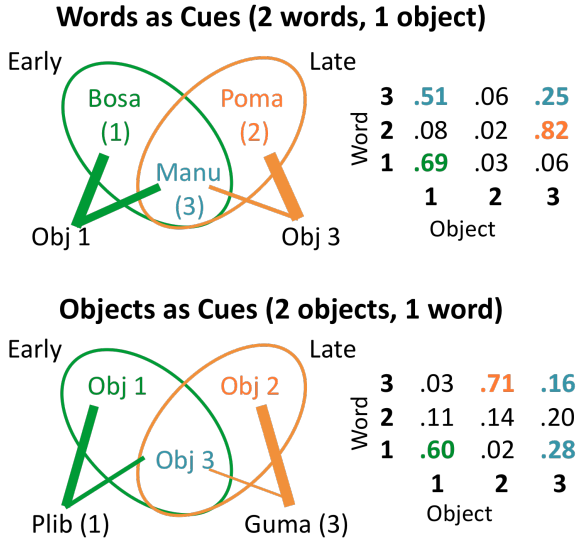

Figure 3: Collated response probabilities ($p(o|w)$) for the two conditions in the Experiment (right). Both conditions show evidence of highlighting, with estimated association strengths shown by thickness of cue-outcome links (left).

In summary, the Experiment shows that highlighting can take place in a cross-situational word learning context, both with objects as cues and words as outcomes, and vice versa. The selective attention account of highlighting holds that the early association of *PE* and *I* with *E* reduces attention to the later co-occurrence of *I* with *L*, thereby leaving *PL-L* to gain more attention (i.e., strength). We contend that this domain-general account explains word-learning behavior not only in this Experiment, but in many situations that have motivated verbal theories of language-specific constraints. In the next section, we introduce a version of a recent associative model of word learning that shows what sort of attentional biases can account for highlighting—and word learning.

Model

Familiarity and novelty are among the simplest ways to be aware of one's knowledge state about stimuli, and both biases have been observed in infants—inferred from their influence on attention (for an overview, see Hunter & Ames, 1988). Kachergis, Yu, and Shiffrin (2012) introduced an associative model with these biases, and showed that it accounts for fast mapping in adults, as well as gradual relaxation of ME with further training. The model assumes that word-object pairings on each trial compete for attention (i.e., associative strength). Attention is preferentially given to word-object pairings that are already associated by previous co-occurrence. Such a mechanism naturally exhibits blocking, since after the early association of q_1 with *o*, it will continue to strengthen q_1-*o* in the late stage, barely attending q_2-*o*. However, the model's bias for familiar pairings competes with a bias to attend to stimuli with no strong associate (e.g., a novel stimulus). This bias can help explain behaviors covered by language-learning principles such as contrast and N3C. We describe the model below, and show how it accounts for highlighting using competing attention for familiar pairings and uncertain stimuli.

Formally, let *M* be an *m* word × *m* object association matrix that is arbitrarily large (here, $m=100$). Cell $M_{w,o}$ is the strength of association between word *w* and object *o*. Strengths are subject to forgetting (i.e., general decay) but are augmented by viewing the particular stimuli. Before the first trial, *M* has no information: each cell is set to $1/m$. On each training trial *t*, a subset *S* of words and objects appear.

Association strengths are allowed to decay, and on each new trial a fixed amount of associative weight, χ, is distributed among the associations between words and objects, and added to the strengths. The rule used to distribute χ (i.e., attention) balances a preference for attending to unknown stimuli with a preference for strengthening already-strong associations. When a word and referent are repeated, extra attention (i.e., χ) is given to this pair—a bias for prior knowledge. Pairs of stimuli with no or weak associates also attract attention, whereas pairings between uncertain objects and known words, or vice versa, do not attract much attention. To capture stimulus uncertainty, we allocate strength using entropy (*H*), a measure of uncertainty that is 0 when the outcome of a variable is certain (e.g., a word appears with one object, and has never appeared with any other object), and maximal ($log_2 n$) when all of the *n* possible object (or word) associations are equally likely (e.g., when a stimulus has not been observed before, or if a stimulus were to appear with

[2] Testing the relative strength of *PE-E* and *PL-L* would ideally be done with a trial that presents both cue *PE* and *PL*, and asks learners which outcome is preferred. However, a test of this sort is difficult to do in a paradigm with spoken words, and we instead chose to match previous word learning paradigms for consistency.

every other stimulus equally). In the model, on each trial the entropy of each word (and object) is calculated from the normalized row (column) vector of associations for that word (object), $p(M_{w,\cdot})$, as follows:

$$H(M_{w,\cdot}) = -\sum_{i=1}^{n} p(M_{w,i}) \cdot \log(p(M_{w,i}))$$

The update rule for adjusting and allocating strengths for the stimuli presented on a trial is:

$$M_{w,o} = \alpha M_{w,o} + \frac{\chi \cdot e^{\lambda \cdot (H(w)+H(o))} \cdot M_{w,o}}{\sum_{w \in S} \sum_{o \in S} e^{\lambda \cdot (H(w)+H(o))} \cdot M_{w,o}}$$

In this equation, α is a parameter governing forgetting, χ is the weight being distributed, and λ is a scaling parameter governing differential weighting of uncertainty and prior knowledge (familiarity). As λ increases, the weight of uncertainty (i.e., the exponentiated entropy term, which includes both the word and object's association entropies) increases relative to familiarity. The denominator normalizes the numerator so that exactly χ associative weight is distributed among the potential associations on the trial. For stimuli not on a trial, only forgetting operates. After training, for each word the model's choice probabilities on k alternative objects is determined by the softmax choice rule (Bridle, 1990):

$$p(o|w) = \frac{exp(\phi M_{w,o})}{\sum_k exp(\phi M_{w,o_k})}$$

where ϕ is a scaling parameter that determines the level of discrimination the model shows: ϕ values above 1 amplify small differences in association weights.

The model was trained on the same 28 trials of word-object co-occurrences experienced by participants in the two conditions, and the four parameters were fit to minimize the discrepancy between the model's predicted response rates and the 36 human choice proportions for each condition. In the words as cues condition, the best-fitting parameters (χ=.11, λ=.46, α=1, ϕ=6.16) achieved an R^2 of .984 (MSE=9.5e-4). In the objects as cues condition, the best-fitting parameters (χ=.12, λ=.37, α=1, ϕ=6.16) achieved an R^2 of .884 (MSE=.0044). Both fits are quite good, and the best-fitting parameters are close in value. With α=1, forgetting was not operating; perhaps memory is not taxed by such a small number of words and objects—cross-situational studies typically have more than a dozen pairs. With ϕ=6.16, the model showed good discrimination at test.

Shown in Figure 4, the model's response proportions are close to the data and fit qualitatively well, showing the highlighting effect in both conditions. How does the model do this? In the first stage of words-as-cues, when w_1 (*PE*) and w_3 (*I*) co-occur with o_1 (*E*), attention is split between the associations w_1-o_1 and w_3-o_1, and the uncertainty about all three stimuli drops as knowledge grows. Moving to the second stage, when w_2 (*PL*) and w_3 (*I*) appear with o_3 (*L*), w_2-o_3 demands more attention than w_3-o_3 because w_3 has lower uncertainty from early training, while w_2 is novel and has no associates. During the second stage, w_2-o_3 thus gets more attention than w_3-o_3, and becomes relatively stronger than the early w_1-o_1 association. Thus, using competing biases for uncertain stimuli and familiar associations, the model mimics the highlighting effect shown by people.

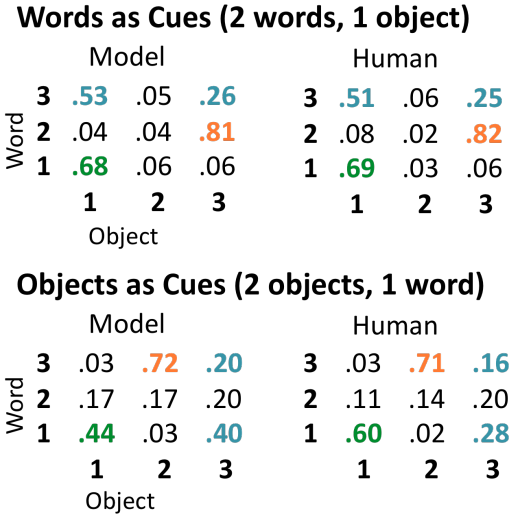

Figure 4: Human response probabilities (right) and model response probabilities (left) closely match in the words as cues condition (top), and match well with objects as cues, showing highlighting in both cases.

Intriguingly, the model shows an asymmetry between conditions that is less striking in humans. With objects as cues, when given w_1 (*E*), the model shows less bias towards o_1 (*PE*) than humans do: people choose o_1 twice as often as o_3 (*I*), whereas the model chooses o_1 only a bit more than o_3. Humans may show a stronger bias for o_1 (*PE*) because they have retrospectively decreased the association between o_3 and w_1 once o_3 began appearing with w_3. Another possibility is that people use uncertainty at test: o_1 (*PE*) has lower entropy than o_3 (*I*) since it only occurred with w_1. With words as cues, both objects have equal entropy. This asymmetry deserves future study, and may yet leave room for language-specific constraints.

General Discussion

We have presented an analogy between cross-situational word learning and associative learning, shown how a study of the former (Ichinco et al., 2009) is a blocking design, and suggested the result is straightforwardly explained with the same domain-general attention mechanism. As evidence that attention creates order effects in word learning, we found highlighting—an associative learning effect ascribed to attention (Medin & Edelson, 1988; Kruschke, 1996)—in a cross-situational word learning experiment. Moreover, we showed that an associative word-learning model with competing attentional biases for familiarity and uncertainty (Kachergis, Yu, & Shiffrin, 2012) accounts for these results.

By linking word learning to associative learning, as suggested by Smith (2000), we may find that the plurality of overlapping language-specific constraints (e.g., ME, N3C, contrast, and fill-the-gap) are unnecessary to explain many

language learning behaviors. Instead, we predict that a more parsimonious explanation will emerge, built upon a foundation of domain-general mechanisms. Language-specific principles and constraints may yet exist, but we should first see how far more universal mechanisms take us.

Moreover, note that this bridge between domains is two-way: the present study used what was originally a word-learning model to explain highlighting. Although our model's attentional account is similar to the account given by other learning models (for an overview see Kruschke, 2011), other models do not use competing uncertainty and familiarity biases to shift attention. Instead, many models use a measure of prediction error to determine the rate of association change (e.g., Rescorla & Wagner, 1972; Pearce & Hall, 1980). In language, objects produce words in speakers ("Watch out—snake!"), but words predict objects for listeners. For language learners, we have shown that both directions of training produce a highlighting effect, captured by our model's symmetric associations and simple biases without generating predictions. These mechanisms, based on some of the simplest cues of knowledge state, may also fare well in other associative learning paradigms—in and out of a word-learning context.

Thus, future work in both domains can benefit from an exchange of ideas to uncover commonalities and differences, and to flesh out and refine verbal theories. We hope that others will find it enlightening to explore the link between associative learning, language acquisition, and other domains.

Acknowledgments

Thanks to John K. Kruschke and Stephen Denton for helpful discussions, to Daniel Yurovsky for comments, and to Patrick LaFree, Jennifer Lee, and Kim Mullen for data collection.

References

Bloom, P. (2000). *How children learn the meaning of words*. Cambridge, MA: MIT Press.

Bridle, J. S. (1990). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In F. Fogelman-Soulié & J. Hérault (Eds.), *Neurocomputing: Algorithms, architectures and applications* (pp. 227-236). New York: Springer-Verlag.

Childers, J. B., & Tomasello, M. (2003). Children extend both words and non-verbal actions to novel exemplars. *Developmental Science*, *6*(2), 185–190.

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

Golinkoff, R. M., Mervis, C. V., & Hirsh-Pasek, K. (1994). Early object labels: The case for a developmental lexical principles framework. *Journal of Child Language*, *21*, 125-155.

Hunter, M. & Ames, E. (1988). A multifactor model of infant preferences for novel and familiar stimuli. In Rovee-Collier, C. & Libsitt, L. (Eds.) *Advances in Infancy Research*, *5* (pp. 69-95). Stamford, CT: Ablex.

Ichinco, D., Frank, M.C., & Saxe, R. (2009). Cross-situational word learning respects mutual exclusivity. In N. Taatgen, H. van Rijn, J. Nerbonne, & L. Schomaker (Eds.) *Proceedings of CogSci 31* (pp. 749–754).

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

Kamin, L. J. (1968). "Attention-like" processes in classical conditioning. In M.R. Jones (Ed.). *Miami Symposium on the Prediction of Behavior, 1967: Aversive Stimulation*. Coral Gables, FL: University of Miami Press (pp. 9–31).

Kaminski, J., Call, J., & Fischer, J. (2004). Word Learning in a Domestic Dog: Evidence for "Fast Mapping." *Science. 304*(5677), 1682–1683.

Kruschke, J. K. (1996). Base rates in category learning. Journal of Experimental Psychology: Learning, Memory and Cognition, 22, 3–26.

Kruschke, J. K. (2009). Highlighting: A canonical experiment. In B. Ross (Ed.), *The Psychology of Learning and Motivation*, *51*, 153–185.

Kruschke, J. K. (2011). Models of attentional learning. In: E.M. Pothos and A.J. Wills (Eds.), *Formal Approaches in Categorization*, pp.120–152. Cambridge University Press.

Markman, E. M. (1992). *Constraints on word learning: Speculations about their nature, origins and domain specificity*. In M. R. Gunnar, & M. P. Maratsos (Eds.), Modularity and constraints in language and cognition: The Minnesota symposium on child psychology (pp. 59–101). Hillsdale, NJ: Erlbaum.

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

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psych: General*, *117*, 68–85.

Mervis, C.B. & Bertrand, J. (1994). Acquisition of the Novel Name - Nameless Category (N3C) principle. *Child Development*, *65*, 1646–1662.

Moher, M., Feigenson, L. & Halberda, J. (2010). A one-to-one bias and fast mapping support preschoolers learning about faces and voices. *Cognitive Science*, 1–33.

Pavlov, I. P. (1927). *Conditioned Reflexes*. London: Oxford University Press.

Pearce, J.M. & Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*(6), 532–552.

Pilley, J. W. & Reid, A. K. (2011). Border collie comprehends object names as verbal referents. *Behavioural Processes*, 86, 184–195.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The Effects of Feature-Label-Order and their implications for symbolic learning. *Cognitive Science*, 34(7), 909-957.

Rescorla, R.A., Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black, W.F. Prokasy (Eds.) *Classical Conditioning II: Current Research and Theory*. New York: Appleton Century Crofts, pp. 64-99.

Smith, L. B. (2000). Learning how to learn words: An associative crane. In *Becoming a Word Learner*. New York: Oxford University Press.

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

Yurovsky, D. & Yu, C. (2008). Mutual exclusivity in cross-situational statistical learning. *Proceedings of CogSci 30* (pp. 715–720). Austin, TX: Cognitive Science Society.