Cognitive Science (2010) 1–49

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2009.01092.x

The Effects of Feature-Label-Order and Their Implications for Symbolic Learning

Michael Ramscar, Daniel Yarlett, Melody Dye, Katie Denny, Kirsten Thorpe

Department of Psychology, Stanford University

Received 29 October 2008; received in revised form 4 November 2009; accepted 9 November 2009

Abstract

Symbols enable people to organize and communicate about the world. However, the ways in which symbolic knowledge is learned and then represented in the mind are poorly understood. We present a formal analysis of symbolic learning—in particular, word learning—in terms of prediction and cue competition, and we consider two possible ways in which symbols might be learned: by learning to predict a label from the features of objects and events in the world, and by learning to predict features from a label. This analysis predicts significant differences in symbolic learning depending on the sequencing of objects and labels. We report a computational simulation and two human experiments that confirm these differences, revealing the existence of Feature-Label-Ordering effects in learning. Discrimination learning is facilitated when objects predict labels, but *not* when labels predict objects. Our results and analysis suggest that the semantic categories people use to understand and communicate about the world can only be learned if labels are predicted from objects. We discuss the implications of this for our understanding of the nature of language and symbolic thought, and in particular, for theories of reference.

Keywords: Language; Learning; Representation; Concepts; Computational modeling; Prediction

Symbolic thought and symbolic communication are defining human characteristics. Yet despite the benefits symbols bring in allowing us to organize, communicate about, manipulate, and master the world, our understanding of symbols and symbolic knowledge is poor. Centuries of pondering the nature of symbolic representation, in terms of concepts and categories and words and their meanings, has yielded more puzzles than answers (Murphy, 2002; Wittgenstein, 1953). Our impoverished understanding of symbolic learning, and especially how words and their meanings are learned, represented, and used, contrasts starkly with the progress made in other areas, where computational models of learning processes

Correspondence should be sent to Michael Ramscar, Department of Psychology, Stanford University, 450 Serra Mall, Stanford, CA 94305. E-mail: ramscar@stanford.edu

have been developed (e.g., Gallistel & Gibbon, 2000; Rescorla & Wagner, 1972) and related to the neuroanatomical structures in which these learning mechanisms are realized (e.g., Hollerman & Schultz, 1998).

In what follows, we present an analysis of symbolic learning—and in particular, word learning—in terms of error-driven learning, which forms the basis of most formal learning models (e.g., Barlow, 2001; Gallistel & Gibbon, 2000; Pearce & Hall, 1980; Rescorla & Wagner, 1972; Rumelhart, Hinton, & McClelland, 1986). This analysis predicts significant differences in learning depending on the ways in which the relationship between symbols and their meanings is established; that is, depending on the way that symbols are related to the aspects of the world they typically label, such as objects, events, etc.

Formally, in learning, two relations are possible between a symbol and, say, a set of objects labeled by that symbol: learning to predict the label from the objects, or learning to predict the objects from the label. Crucially, discrimination learning is facilitated when objects predict labels, but not when labels predict objects. This is due to differences in cue competition. When objects predict their labels, the various features of those objects compete for relevance, which results in the features that are most predictive (or ''definitive'') of each label—and which discriminate the meanings of labels from one another—being highlighted in learning. On the other hand, when labels predict objects, the sparse features of labels inhibit competitive discrimination learning and impair symbolic learning.

The results of a computational simulation and a study of adults learning artificial categories confirm these differences, as does a study of children learning color words. In each of these studies a Feature-Label-Ordering (FLO) effect in learning is clearly evident; discrimination learning is facilitated when objects predict labels, but *not* when labels predict objects. Even when learners are apparently given exactly the same information, manipulating the order in which objects and labels are encountered has a dramatic effect on learning (because as we will show, the information available to a learner is critically affected by the ordering of those objects and their labels). The studies we describe here, along with other work we review, pose serious questions for traditional theories of language based on reference. We review these challenges and suggest that our findings offer support for an alternative approach in which language is seen as a fundamentally predictive process.

1. Symbolic representation and the problem of reference

People use symbols (such as words, signs, or pictures) and arrangements of symbols to communicate about the world. In seeking to understand how learning makes this possible, we do not presuppose that symbolic thought is necessarily the same thing as ''symbolic computation,'' where symbolic computation is equated with a particular algorithmic—usually procedural—approach to computer programming. There are numerous ways in which symbolic thought might be implemented in the mind, and the model adopted in the ''symbolic approach'' to cognitive science is but one of these (Haugeland, 1985, makes a similar point regarding ''Good Old Fashioned Artificial Intelligence'' and other approaches to artificial intelligence).

''Symbolic'' approaches to thought and language typically characterize mental representations in terms of rules that define relationships between classes of entities (such as ''if X then Y''). In the classic statement of this approach, Fodor and Pylyshyn (1988) argue that there is a ''*combinatorial syntax for mental representation*, in which (a) there is a distinction between structurally atomic and structurally molecular representations; (b) structurally molecular representations have syntactic constituents that are themselves either structurally molecular or are structurally atomic; and (c) the semantic content of a (molecular) representation is a function of the semantic contents of its syntactic parts, together with its constituent structure'' (p. 12).

Computationally, this approach ultimately requires that type/token relationships for classes of structures be defined. For example, defining what constitutes an atomic X or a Y enables instances of Xs and Ys to be bound to the appropriate part of a molecular structure, such as ''if X then Y,'' allowing the structure to describe a relationship in the world. Importantly, however, if the definitions of classes are themselves symbolic (i.e., molecular), this in turn imposes a requirement that *all* symbols in the definitions be defined (i.e., if X is defined as ''all Xs have Z,'' one needs to define Z).

Unless classes are defined, a representational ''Russian doll'' problem arises, because defining symbols with other symbols is inherently regressive. If ''dog'' is a token of the type ''noun,'' ''spaniel'' is a token of ''dog,'' and ''Fido'' is a token of ''spaniel,'' the relationships between ''Fido,'' ''dog,'' and ''spaniel'' cannot be explained by, implemented, or generalized from ''a dog is a noun'' or ''a spaniel is a dog'' without an account of *what makes* Fido a spaniel, and spaniels dogs—*as opposed to something else*. Similarly, saying, ''a sentence is grammatical if it is syntactically correct,'' explains little unless one defines which things in the world are and are not members of the classes ''sentence,'' ''grammatical,'' and ''syntactically correct.''

The problems do not end there. If symbolic representations are conceived of as ''compositional'' (such that sentences in natural language have structural meanings that are derived from the structure of sentence, which—in turn—affects the specific meanings of words out of which the sentence is composed; see e.g., Fodor, 1998), one needs an account of how relevant individual tokens of meaning are extracted from descriptions that only mention types. For example, one needs to be able to say which aspects of the meanings of ''cat,'' ''sat,'' and ''mat'' are relevant to the meaning of ''the cat sat on the mat.'' This requires a further account of how one goes from a class label (''cat'') to a specific individual or instance (a particular cat) in a particular context.

No satisfactory solution to these problems is provided by any existing symbolic approach (Fodor, 1998, 2000; Murphy, 2002). Indeed, there are good reasons to believe—in principle—that these problems cannot be solved. The kinds of things that people represent and think about symbolically do not fall into discrete classes, or categories, of Xs and Ys (Wittgenstein, 1953); symbolic categories do not possess discrete boundaries (i.e., there *are no* fixed criteria for establishing whether an entity is an X or a Y); and entities are often assigned to multiple symbolic classes (i.e., they are sometimes Xs; sometimes Ys). As a result of these and many other factors, symbolic type/token relationships appear to be inherently underdetermined (see e.g., Fodor, 1998; Quine, 1960; Wittgenstein, 1953). This is a

serious problem for all current symbolic approaches (Fodor, 1998), and it has prompted theorists to conclude that while there *must be* a solution, it is innate and largely inscrutable (i.e., it is there, but we do not know what it is; Chomsky, 2000; Fodor, 1983, 1998).

Alternative approaches to characterizing thought and language—especially those that take an associative (or connectionist) approach to mental representation—are often termed ''subsymbolic,'' to distinguish them from ''symbolic'' models (e.g., Fodor & Pylyshyn, 1988; Rumelhart & McClelland, 1986). However, to the extent that we think of thought as being symbolic (and it seems natural to do so, especially with regard to language) and to the extent that associative, connectionist, and ''symbolic'' approaches all seek to explain the nature of thinking, differentiating ''symbolic'' and ''subsymbolic'' approaches to representation without a clear idea of what symbolic thought actually *is* runs the risk of missing the point altogether.

Far more important than any ''symbolic''/''subsymbolic'' distinction is the assumption made by cognitive theories of all persuasions that symbolic thought is *referential*; that is, that symbols both represent and point to meanings, so that symbols and their meanings share a *bidirectional* relationship. Symbols are typically seen as abstract representations that either *exemplify* (stand for) or *refer* (point) to their meanings (*referents*; these meanings are often considered to be defined by reference to things in the world; for example, the symbol ''dog'' is considered to be defined by reference to a class of things in the world, *dogs*). The problems with this approach are largely the same as for type/token definitions, and they have been laid out exhaustively (see e.g., Fodor, 1998; Goodman, 1972; Murphy, 2002; Quine, 1960; Wittgenstein, 1953).

While reference presupposes that the relationship between symbols and meanings is bidirectional, this assumption is at odds with the idea that symbols actually are *abstract* representations, because abstraction is not a bidirectional process. Abstraction involves reducing the information content of a representation, such that only information relevant to a particular purpose is retained (Hume, 1740; Rosch, 1978). As such, abstraction is an inherently directed process: one can abstract *from* a larger body of information *to* an abstract representation of it, but one cannot reverse the process, because discarded (as opposed to compressed) information cannot be recovered. For example, while one might sensibly read an article and summarize it in an abstract, the idea of ''reverse abstraction'' supposes that one can get detailed methods and results information from the abstract of a research article that one has never read.

Given that symbols serve as abstractions in communication and thought, it seems reasonable to assume that communication and thought respect the basic principles of abstraction. In what follows, we treat symbols as abstractions in a literal sense: given that abstraction is a directed process, we assume symbolic representation and processing must be directed as well. Our approach to symbolic representation is explicitly not referential. Instead, it is *predictive*. Prediction is by its very nature directed: A prediction follows *from* the cues that lead *to* a given expectation. In what follows, we show that the relationship between symbols and the things they represent is *not* bidirectional, and that symbolic processing is a process of predicting symbols.

2. Symbolic learning

In considering how symbols are represented and used, we begin by examining how they are learned. In what follows, we conceive of learning as a process by which information is acquired about the probabilistic relationships between important regularities in the environment (such as objects or events) and the cues that allow those regularities to be predicted (Gallistel, 2001, 2003; Gallistel & Gibbon, 2000; Rescorla, 1988; Rescorla & Wagner, 1972).

Crucially, the learning process is driven by discrepancies between what is expected and what is actually observed in experience (termed *error-driven learning*). The learned predictive value of a given cue produces expectations, and any difference between the value of what is expected and what is observed produces further learning. The predictive value of a given cue is strengthened when relevant events (such as events, objects, or labels) are underpredicted by that cue and weakened when they are overpredicted (Kamin, 1969; Rescorla & Wagner, 1972). As a result, cues compete for relevance, and the outcome of this competition is shaped both by *positive evidence* about co-occurrences between cues and predicted events, and *negative evidence* about nonoccurrences of predicted events.

This process produces patterns of learning that are very different from what would be expected if learning were shaped by positive evidence alone (a common portrayal of Pavlovian conditioning, Rescorla, 1988), and there is evidence for this error-driven characterization of learning in the brain; for example, the firing patterns in dopamine neurons in monkeys' brains when learning trials are underpredicted or overpredicted closely resemble the patterns produced by error-driven learning models (Waelti, Dickinson, & Schultz, 2001).

This view of learning can be applied to symbolic thought by thinking of symbols (i.e., words) as both potentially important *cues* (predictors) and *outcomes* (things to be predicted). For example, the word ''chair'' might be predicted by, or serve to predict, the features that are associated with the things we call chairs (both when chairs and ''chair'' are present as perceptual stimuli, or when they are being thought of in mind). Word learning can thus take two forms, in which either:

(i) cues are *labels* and outcomes are *features*
(ii) cues are *features* and outcomes are *labels*.

In (i), which we term *Label-to-Feature* (*LF*) *learning*, learning is a process of acquiring information that allows the prediction of a feature or set of features given a label, whereas in (ii), which we term *Feature-to-Label* (*FL*) *learning*, learning is a process of acquiring information that allows the prediction of a label from a given feature or set of features.

Many theories of symbolic cognition emphasize the importance of structured relations *between* things in our understanding of the world (Chomsky, 1957; Fodor & Pylyshyn, 1988; Gentner, 1983, 2003; Goldstone, Medin, & Gentner, 1991; Kurtz, Gentner, & Gunn, 1999; Markman, 1999; Penn, Holyoak, & Povinelli, 2008). Despite the widespread belief that associative models are unstructured (e.g., Fodor, 1998; Fodor & Pylyshyn, 1988), the

opposite is true. Treated properly, associative models are inherently structured. Although they are often referred to as ''associative,'' all contemporary theories of learning are, as we described above, predictive. Learning discovers cue structures (O'Reilly & Rudy, 2001; Pearce, 1987, 1994) that share temporal, predictive relationships with other things (e.g., events, objects, or labels) in the environment (see also Elman, 1990). Prediction is fundamentally relational, and LF and FL learning describe the two possible ways that these relations can be structured in symbolic learning.

That associative and connectionist models can be configured so that they do not respect the predictive structure of the environment—that is, such that they model relationships between cues and outcomes that do not actually have a similar predictive relationship in the environment—is incidental to this basic point, although it almost certainly contributes to the perception of associative and connectionist models as being unstructured. Thus, for example, the influential Rumelhart and McClelland (1986) model simulates English past tense production as a process of predicting past tense forms—*walked*—from stem forms—*walk*. This is not, however, how children learn language: Children do not learn their first (native) language by memorizing verb conjugations by rote, as in a classroom, but rather they learn language in context, which leads them to expect a past tense form given a semantic context that predicts its occurrence (see Ramscar & Dye, 2009a; Ramscar & Yarlett, 2007, for application of this idea to the learning of English plural inflection that allows many ''mysteries'' relating to the way children learn English plurals to be resolved).

With regard to the structure of symbolic learning, in FL learning, the set of cues being learned from is generally larger than the set of outcomes being learned about, whereas in LF learning, the set of outcomes is generally larger than the set of cues. As we will now show, these set-size differences in the number of *cues* and *outcomes* being learned about in each of these two forms of word learning result in different levels of discrimination learning, and asymmetries in the cognitive representations learned.

FL learning can be illustrated by imagining a learner in a world containing two kinds of animals: wugs and nizzes (Fig. 1). Both share identical bodies, but wugs are red, whereas nizzes are blue. In order to communicate about wugs and nizzes, the learner must discover the relationship between their features (color and body type) and their labels. As cue competition in learning is essentially a process of revising expectations, this relationship will be easily discovered when their features predict their labels. When a learner in the scenario depicted in Fig. 1 expects one label but hears the other, there is a violation of expectation. Faced with an unexpected outcome—a prediction error—the learner will begin to adjust her expectations accordingly. In this case, because the shared shape-feature cues both ''wug'' and ''niz,'' a violation of expectation will occur whenever one of the labels is not heard (as both are predicted by shape). This will cause the learner to scale down her expectations of body shape as a reliable cue, as she shifts the weight of her expectation to the most predictive cues, the colors.

In trial (i), our learner will discover that the features *red* and *body* predict the label ''wug.'' However, because the feature *body* predicts both ''wug'' and ''niz'' indiscriminately, it incorrectly predicts that ''wug'' will occur in trial (ii). As a result, the strength of the association between *body* and ''wug'' decreases, even though ''wug'' is not pres-

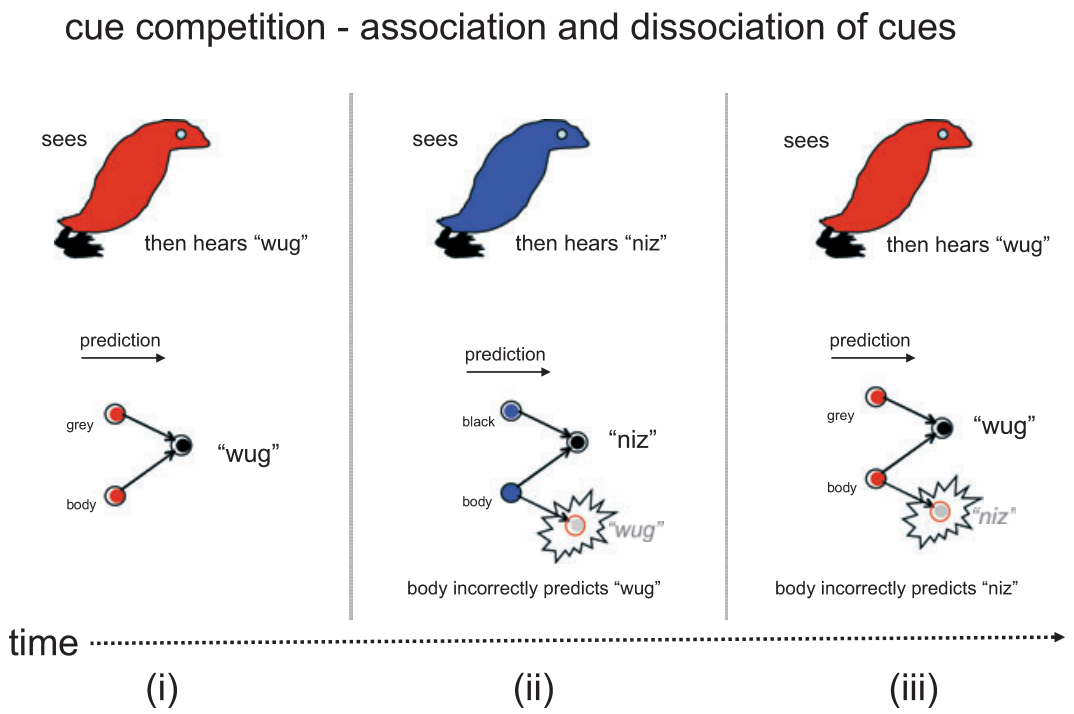

Fig. 1. Cue competition in Feature-to-Label learning. The top panels depict the temporal sequence of events: An object is shown and then a word is heard over three trials. The lower panels depict the relationship between the various cues and labels in word learning.

ent on this trial. The converse occurs in trial (iii), when *body* incorrectly predicts ''niz.'' In that trial, the associative strength between *body* and ''niz'' decreases. In this example, because the feature *body* cues both ''wug'' and ''niz,'' a violation of expectation will occur whenever one of the labels is not heard, as both are predicted. Over time, this will cause the learner to adjust her expectations of *body* downwards to reflect its unreliability as a cue; its cue value will steadily decrease over learning trials, until it eventually approaches zero. As a consequence, in FL learning, *body* will be effectively unlearned as a useful cue, and the colors *red* and *blue* will be learned to be the most predictive cues to ''wug'' or ''niz.''

As the cue value of *body* diminishes, the cue value of *color* will correspondingly increase, resulting in a growing discrepancy between the strength of the expectations produced by *body* and *color*. In the learning trials for ''wug,'' shown in Fig. 2, the color cue *red* gains in associative value as a result of the diminishing value of *body*. Importantly, even though *body* and ''wug'' co-occur with exactly the same frequency as *red* and ''wug,'' learning effectively dissociates *body* and ''wug'' in this situation.

FL learning is thus *competitive*: if a cue loses associative strength, its value can change *relative* to other cues. As one cue's loss can be another's gain, this allows associative value to *shift* from one cue to another. As a consequence, it is predictive power—and

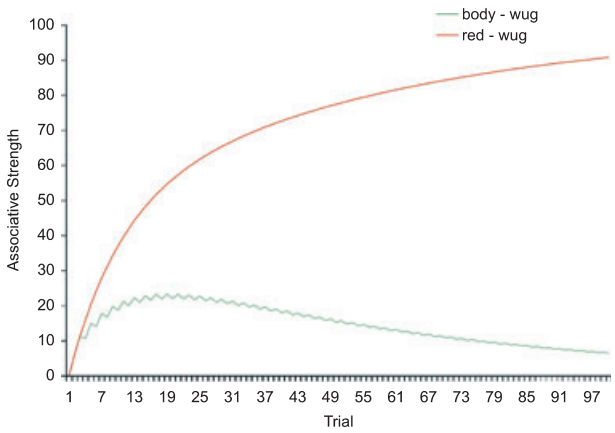

Fig. 2. A simulation of error-driven learning of *body*-''wug'' and *red*-''wug'' in the scenario depicted in Fig. 1. The graph shows the cue values developing in the Rescorla and Wagner (1972) model. The simulation assumed that either a ''wug'' or a ''niz'' was encountered in any given trial, that wugs and nizzes were each equally frequent in the environment, and that color and shape were equally salient features. The errors produced by *body* cause it to lose out in cue competition with *red* so that the association between *red* and ''wug'' is emphasized, while the association between *body* and ''wug'' is devalued. Though *body* and ''wug'' co-occur with exactly the same frequency as *red* and ''wug,'' learning effectively dissociates the two in this scenario.

not frequency or simple probability—that determines cues' values. Because learning emphasizes the set of cues that most reliably predicts each category label, cue competition improves *discrimination*. In this scenario, the learner comes to ignore *body* as a predictive cue. Cue competition reduces the amount of overlap in the cues to ''wug'' and ''niz,'' thereby increasing the difference between the conditions that lead to the expectation of ''wug'' or ''niz'' (see also Rumelhart & Zipser, 1986).

Although in this simple example, a single cue (*red*) is a perfect predictor of the appropriate category (''wug''), real-world categorization is a complex, probabilistic process (see Murphy, 2002; Wittgenstein, 1953). There will be a great deal of overlap in the sets of cues that predict different labels, and competition will serve to shape cue values that minimize error rather than eliminate it altogether.

3. Cue competition and the transformation of cue values

Cue competition is the process by which cues compete for relevance in the prediction of a particular outcome. When a particular cue successfully predicts a given outcome over a number of learning trials, the associative value of the cue will increase. Conversely, when a particular cue *un*successfully predicts a given outcome—that is, the predicted outcome does *not* follow the cue, the associative value of the cue will decrease. When a number of cues are present together, their associative values will increase or decrease depending on how reliable they are as predictors of the outcome.

The transformation of cue values over learning trials can be stated mathematically, as can the scope of what can be learned given a set of cues. The limit on the number of predictions that can be encoded in a given set of cues can be defined as follows: For any discrete outcome[1] to be predicted, a unique cue value (or set of values) must lead to each prediction, such that one set of cue values discriminates one outcome from any other outcomes, while another set of values discriminates another outcome, and so on. It is important to note that it is impossible for one cue value to predict two different discrete outcomes—that is, given only cue A, one might say that either outcome B or outcome C will follow, or both, but only *one* of these possible predictions can actually be encoded in a single cue. It follows then that the total number of discrete predictions D that can be encoded in a set of cues S in which each cue can take V values can be expressed as:

$$D = (V^S - 1) \tag{1}$$

The subtraction reflects the fact that in the absence of any cues, no predictions can be made. Thus, two binary valued cues allow up to three possible outcomes to be discriminated, four cues allow up to 15 outcomes to be discriminated, five cues allow 31 outcomes to be discriminated, and so on. There are two important cases in which the number of outcomes that a set of cue values can discriminate will be reduced: either when *redundant* cues are present, or when the sets of cues used to predict things are not themselves discrete (such as, say, the cues to dogs, wolves, and coyotes), which will result in prediction error and cue competition.

These points about discrimination can be re-described in terms of encoding. Logically, given a large enough set of cues, it is possible to uniquely encode every possible combination of a smaller set of outcomes. However, when this set size relation is inverted, so that a small set of cues is used to encode a larger set of outcomes, it becomes mathematically *impossible* to uniquely encode in the smaller set all of the possible outcomes that might occur in the larger set (see also Abramson, 1963; Kolmogorov, 1965; Rodemich, 1970; Shannon, 1948; logically, this is the basis of the problem of ‘‘reverse abstraction’’ we described earlier).

Accordingly, situations in which there are few cues (and few cue values) provide a poor basis for discrimination learning, specification, and encoding. This is a problem that affects LF learning due to the characteristics of verbal labels.

4. Verbal labels lack cue structure

Verbal labels are relatively discrete and possess little cue structure—by ‘‘cue structure’’ we mean the number of salient and discriminable cues they present simultaneously—whereas objects and events in the world are far less discrete and possess much denser cue structure. Consider a situation in which a *pan* is encountered in the environment. A pan has many discriminable features a learner might treat as cues to *pan*, namely its shape, color, size, and so on.[2]

Now consider the label ''pan.'' A native English speaker can parse it into a sequence of phonemes [p^h *an*] but will be largely unable to discriminate further cues within these sounds. Studies have shown that listeners perceptually divide continuous acoustic dimensions into discrete phonetic categories, for which they exhibit good between-category discrimination and poor within-category discrimination (Kuhl, 2000; Werker & Tees, 1984). For example, the voice onset time (VOT; the temporal difference between the aspirant release of a consonant and its sounding in the vocal chords) between the voiceless and voiced bilabial stop consonants /b/ and /p/ differs along a continuous dimension. While, in theory, this can be subdivided into infinitely smaller time units, English speakers are perceptually insensitive to these kinds of subdivisions. English speakers readily discriminate /b/ or /p/ based on VOT, but they are largely incapable of learning reliable within-category discriminations in /b/ and /p/ (Kuhl, 1994). Thus, English speakers do not perceive /b/ as being composed of reliably discriminable subfeatures (i.e., hearing /b/ as comprising /b^1/, /b^2/, /b^3/, etc.; see Kuhl, 1994), nor do they employ these discriminations semantically in everyday speech. Listeners easily discriminate ''*p*an'' from ''*b*an,'' but they do not (and perhaps cannot) discriminate more features in the sounds of ''pan'' itself. (To draw an analogy with color categorization, while most English speakers can make basic discriminations within color hues—e.g., light green, medium green, dark green—they do not usually discriminate—verbally or otherwise—between increasingly similar hues within those categories, and as a result, take the same signal—''GO''—from both emerald and chartreuse colored traffic lights.)

Because the effects of cue competition become attenuated as the temporal relations between cues vary (Amundson & Miller, 2008), and because phonemes are perceived sequentially rather than simultaneously (Marslen-Wilson, 1975, 1987; McClelland & Elman, 1986; Norris, McQueen, & Cutler, 1995, 2003; Norris, McQueen, Cutler, & Butterfield, 1997), phonemes cannot compete directly as cues. Moreover, the other discriminable cues present in speech—such as emphasis, volume, and pitch contour—do not covary systematically with phonemes (unlike, for instance, the features of dogs which *do* covary systematically with one another, and with the word ''dog''). Given that the key regularity that does systematically covary with semantics in many languages appears to be the phoneme,[3] this means that when labels serve as cues to meanings, they do not provide a would-be learner with many useful cues upon which to base learning (even allowing for systematic variations of tone, articulation, volume, etc.). When a label such as ''pan'' serves as a cue, it essentially provides a learner with a single useful cue: the label ''pan'' itself.

5. Cue competition in FL and LF learning

Because labels lack cue structure, the principles we describe predict very different results when labels predict features (LF learning), as compared to when features predict labels (FL learning). FL learning has a many-to-one learning form: Each feature of an object is a potential cue to a label, and thus features can compete with one another for predictive value. By contrast, LF learning has a one-to-many learning form: Only one label is encountered at a

time, and thus, essentially only a single cue is predictive of *all* of the many features that might be encountered in an object or other outcome. As there are no other cues to compete for associative value, there can be no cue competition and no loss of associative value to other cues over the course of learning trials. The value of a single cue will simply increase when a predicted outcome appears following the cue, and decrease when a predicted outcome fails to appear following the cue.

In the wug/niz example illustrated in Figs. 1 and 2, FL learning discriminated the most reliable cue to each label as the result of differences in the covariance between color and shape cues and labels, which advantaged color cues over shape cues. To contrast the differences between FL learning and LF learning, let us now turn to a more complicated world where color is not a reliable cue. In this new scenario, wugs are *wug-shaped*, but can be *blue* or *red*, and nizzes are *niz-shaped*, but can likewise be *blue* or *red* (Figs. 3 and 4). In this case, the labels ''wug'' and ''niz'' are most effectively predicted by shape (*wug-shaped* or *niz-shaped*) rather than color (as both wugs and nizzes can be *blue* or *red*).

An FL learning scenario is illustrated in Fig. 3, in which wugs and nizzes precede—and thus predict—their labels. At (i), a learner encounters an object with two salient features, shape-1 and red, and then hears the label ''wug.'' The learner acquires information about two equally predictive relations, shape-1⇒''wug'' and red⇒''wug.'' At (ii), the learner encounters two new cues and a new label, and forms two new equally weighted predictive relations, shape-2⇒''niz'' and blue⇒''niz.'' Then at (iii), the learner encounters two previously seen cues, shape-1 and blue.

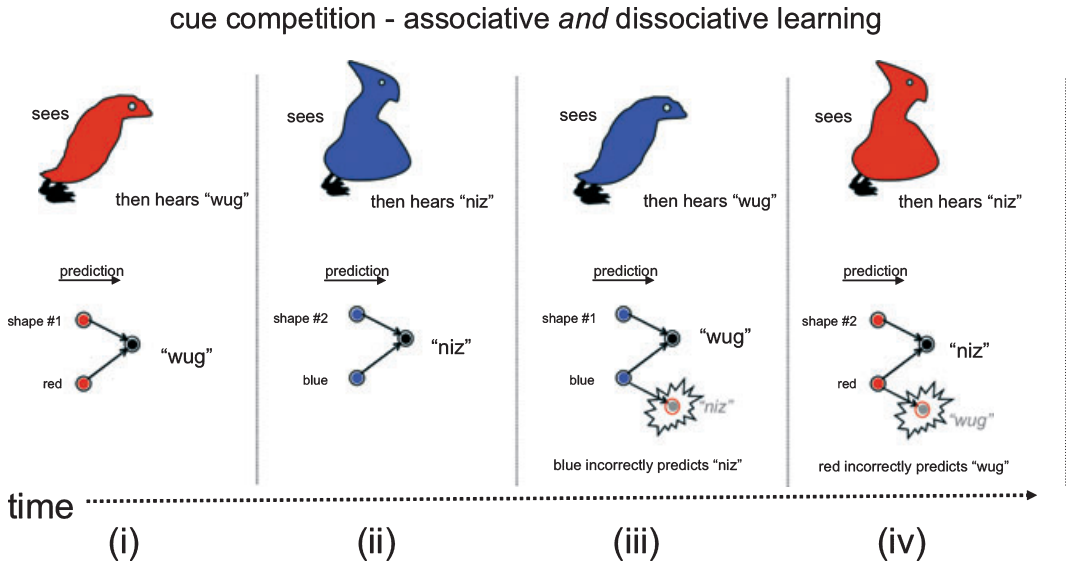

Fig. 3. When features predict their labels in FL learning, the nondiscriminating features will be dissociated from the labels through cue competition.

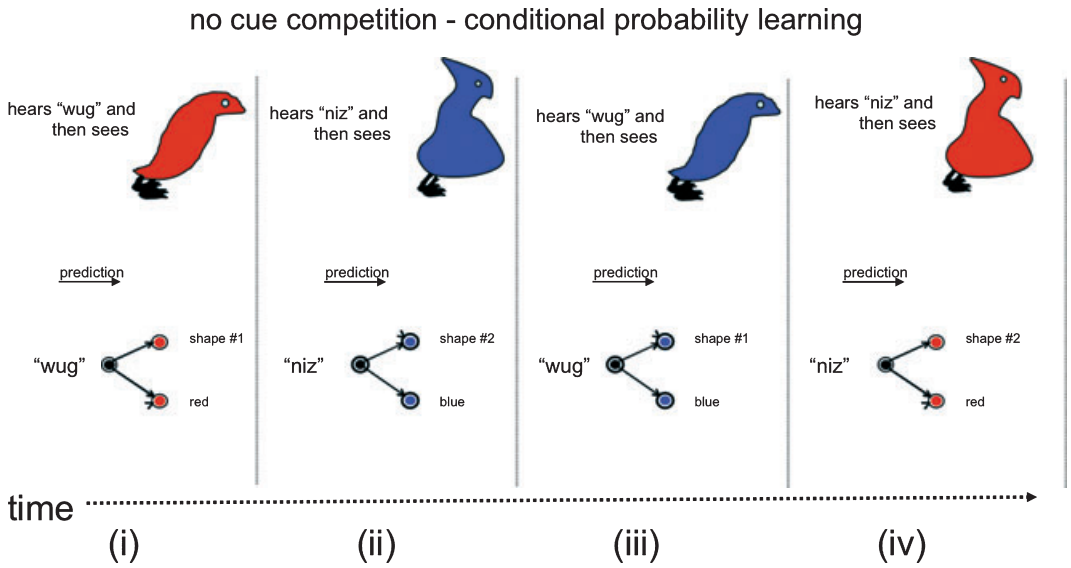

Fig. 4. The absence of cue competition when labels predict features in LF learning will result in the conditional probability of a feature given a label being learned. In this situation, the outcome of learning will simply be a representation of the probability of the features given the labels.

Given what our learner already knows—that is, shape-1⇒‘‘wug’’ and blue⇒‘‘niz’’—she will expect both ‘‘wug’’ and ‘‘niz.’’ In this instance, however, only ‘‘wug’’ occurs. As a result: (a) given positive evidence of the occurrence of ‘‘wug,’’ the associative values for the relation shape-1⇒‘‘wug’’ and blue⇒‘‘wug’’ increase; and importantly (b) negative evidence about the nonoccurrence of ‘‘niz’’ causes blue⇒‘‘niz’’ to lose associative value. Crucially, as the associative value of blue⇒‘‘niz’’ decreases, its value *relative* to shape-2⇒‘‘niz’’ changes as well (making shape-2 a *better* predictor of ‘‘niz’’). At (iv), a similar situation occurs. The learner encounters shape-2 and red and expects ‘‘niz’’ and ‘‘wug.’’ As ‘‘niz’’ is heard, the associative values of shape-2⇒‘‘niz’’ and red⇒‘‘niz’’ increase, while red⇒‘‘wug’’ loses associative value.

Now consider LF learning in a similar scenario (Fig. 2). At (i), a learner encounters the label ‘‘wug’’ and then an object with the two salient features, shape-1 and red. She thus learns about two equally valuable predictive relations ‘‘wug’’⇒shape-1 and ‘‘wug’’⇒red. Similarly, at (ii), the learner acquires two further equally valued relations ‘‘niz’’⇒shape-2 and ‘‘niz’’⇒blue. Now, at (iii), the learner hears ‘‘wug’’ and expects red and shape-1. However, shape-1 occurs and blue occurs. This has three consequences: (a) positive evidence increases the associative value of ‘‘wug’’⇒shape-1; (b) ‘‘wug’’⇒blue becomes a new predictive relation; (c) negative evidence decreases the value of ‘‘wug’’⇒red. However, as ‘‘wug’’ is the only cue, this loss of associative value is *not* relative to any other cues (and likewise at [iv] with ‘‘niz’’).

LF learning is *noncompetitive*. The value of a label-cue will increase when a predicted object (or feature) appears and decrease when a predicted object fails to appear.

However, as there are no other labels (cues) to compete for associative value, there can be no loss of potential associative value to *other labels* over the course of learning trials. Because of this, the effect of prediction error on cue value differs from FL learning. In the absence of cue competition, the cue value of a label will simply come to represent the proportion of successful predictions it has made relative to the proportion of unsuccessful predictions. Accordingly, its value will track the frequency with which labels and features co-occur, approximating the conditional probability of a feature given that label (see Fig. 5; also Cheng, 1997; Wasserman, Elek, Chatlosh, & Baker, 1993). LF learning is thus characterized by *conditional probability learning*, the outcome of solely noncompetitive learning.

Consistent with this, there is a considerable body of evidence showing that in sequential learning tasks in which single phonemes predict other phonemes (i.e., LF learning), participants acquire a good understanding of the transitional probabilities between phonemes in the training sequence (see Saffran, 2001; Saffran, Aslin, & Newport, 1996; Saffran, Johnson, Aslin, & Newport, 1999). Interestingly, it seems likely that the emphasis in cognitive science on simple LF-style learning tasks like this, in which participants can and do learn only the transitional probabilities between cues and subsequent events, has inadvertently contributed to the widespread misconception of learning as being limited to simple probability learning (see also Rescorla, 1988).

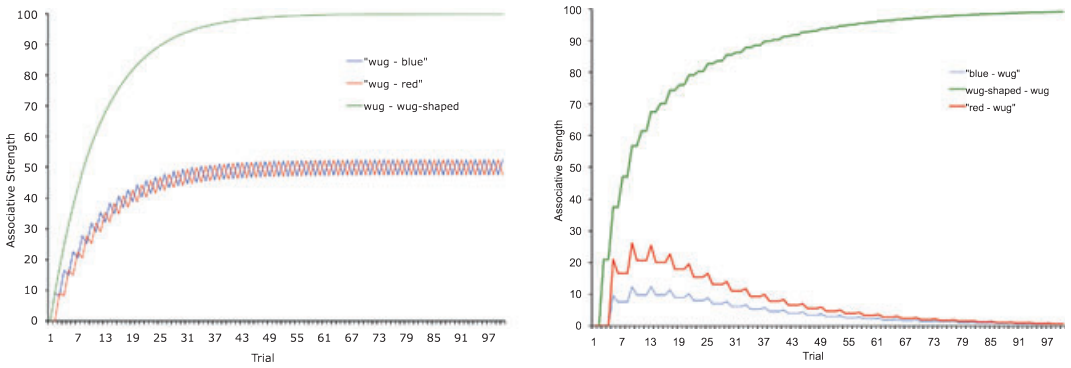

Fig. 5. Simulations of LF and FL learning in the Rescorla and Wagner (1972) model. The left panel simulates learning the cue values of the relationships ''wug''⇒*wug-shaped* and ''wug''⇒*red* in the scenario depicted in Fig. 4, where labels predict features (LF). Because ''wug''⇒*red* is not subject to cue competition, the cue value of ''wug'' simply increases when it successfully predicts *red* and decreases when it predicts *red* in error. Learning asymptotes at the conditional probability of *red* given ''wug.'' The right panel is a simulation of Fig. 3, where features predict labels (FL). In this case, nondiscriminating features are effectively dissociated from the labels in learning. Note that because *red* is encountered earlier in training than *blue* in the FL simulation, it is initially less affected by competition from *wug-shaped*. The peak value of *blue* is less than that of *red* because *wug-shaped* acquired associative value as a cue to *wug* on the red *wug* trial, which preceded the blue *wug* trial. (The simulations assume that a *niz* or a *wug* is encountered in each trial; that both species and their different colored exemplars are equally frequent in the environment; and that color and shape are equally salient.)

Both FL and LF learning capture probabilistic information about predictive relationships in the environment. However, there are fundamental differences between the two. In FL learning, predictive power, not frequency or simple probability, determines cue values; LF learning is probabilistic in far simpler terms. In LF learning what gets learned is the *statistical* structure of the environment. In contrast, in FL learning what gets learned is the *predictive* structure of the environment. This analysis of learning predicts that very different probabilistic understandings of the world will be acquired depending on the order in which features and labels are encountered in learning. We call this the FLO hypothesis.

While we have illustrated the effects of FLO with examples of a child learning to label objects, it should be noted that these principles may apply to the learning of *all* environmental regularities; that is, events, affordances, landmarks, etc. The underlying logic of the FLO hypothesis is not limited to word learning.

6. Learning and response discrimination

Hypothetically, the differences between LF and FL learning might not matter. If, for example, all the objects that shared a label also shared discriminating features, and if the exemplars of each labeled category were encountered equally frequently, an LF-learner might do a *reasonable* job of learning to associate objects and labels. However, in the real world, where the features of objects with different labels often overlap considerably (dogs and foxes look very similar), and where object frequencies vary enormously (a child will see far more dogs than foxes), an LF-learner will struggle. This is because rather than discriminating between expected outcomes, LF learning tends to produce representations in which a number of competing outcomes are all highly probable.

To illustrate the problem of outcome (or response) interference, we return to the wug/niz example. Imagine that in this world of wugs and nizzes, there were 50 times as many blue wugs as blue nizzes in the animal population, and 50 times as many red nizzes as red wugs. In our original example of LF learning, in which there were equal numbers of wugs and nizzes, the color *red* cued ''wug'' 50% of the time and ''niz'' 50% of the time. In this new world, however, the color *red* will cue ''wug'' about 98% of the time and ''niz'' only about 2% of the time, simply based on frequency of occurrence (Fig. 6). Imagine a child trained LF on the animals sees a red wug and is trying to say what it is called. In our original example, it would have seemed easy—near-100% probability that wug-shaped → wug and only 50% probability that red → niz. She will say ''wug.'' But in this new example, there is again a near-100% probability that wug-shaped → wug, but now there is also a 98% probability that red → niz. So what will she say? There is going to be a large degree of uncertainty regarding the correct answer. Because tracking the frequencies of successful predictions does *not* highlight the features of a set of objects that discriminates that set from other objects assigned different labels, the child will experience considerable *response interference* when labeling wugs and nizes. Consequently, while both FL and LF learning can, in theory, discriminate high-frequency items, LF learning will be far less effective given lower frequency items.

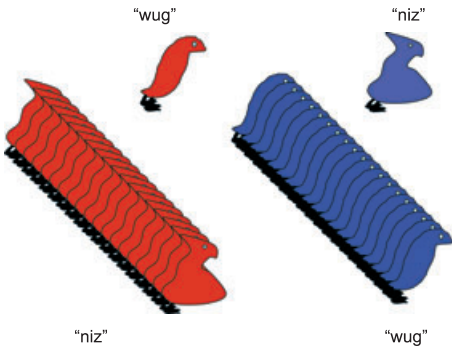

Fig. 6. LF learning makes response discrimination difficult when the frequency of objects in the world varies. Here there is 100% probability of ''wugs'' being *wug-shaped* and ''nizzes'' being *niz-shaped*. However, there is also a 98% probability of ''wugs'' being *blue* and ''nizzes'' being *red.* As a result, when the conditional probabilities are learned LF, red ''wugs'' will be poorly discriminated from red ''nizzes'' and blue ''nizzes'' will be poorly discriminated from blue ''wugs.''

7. Simulation experiment

To test these analyses, we conducted a series of computational simulations using a prominent learning model (Rescorla & Wagner, 1972). The Rescorla–Wagner model has been applied to a variety of learning effects in animals and humans (e.g., Gluck & Bower, 1988a,b; Rudy, 1974), and it is the most widely applied learning model in existence (Siegel & Allan, 1996). The Rescorla–Wagner model is error driven (or expectation-based), meaning that it models the learning of expectations and their refinement in further learning. Specifically, the model learns the relationship between events in the environment and the cues that enable those events to be predicted.

Although it cannot account for all the phenomena observed in associative learning (Miller, Barnet, & Grahame, 1995), the Rescorla–Wagner model offers a straightforward formalization of the basic principles of error-driven learning and yet is sufficiently detailed to allow testing of the analysis we present here in an accessible way. It should be noted that the FLO hypothesis stems from an analysis of how error-driven learning—and in particular, how cue competition—interacts with set-size differences in the number of discriminable features provided by symbols, as compared to objects and events, etc. Importantly, cue competition and error-driven learning are not specific to the Rescorla–Wagner model. Rather, they are well-supported phenomena in learning and are realized in the learning rules of a wide range of models (e.g., Barlow, 2001; Gallistel, 2003; Gallistel & Gibbon, 2000; McLaren & Mackintosh, 2000; Pearce & Hall, 1980; Rosenblatt, 1959; Rumelhart et al., 1986) in which comparable simulations of our analysis could be implemented.

In the Rescorla–Wagner model, the scope of what can be learned given a set of cues can be stated mathematically in terms of the associative relationships between the cues and the outcomes they predict. The model specifies how the associative strength (V) between the set of cues i and the outcome j changes as a result of discrete training trials, where n indexes

the current trial. Note that in animal models, cues would equate to the conditioned stimulus (CS) and outcomes to the unconditioned stimulus (US).

To relate the model to our analysis, it is worth noting that:

1. In FL learning, cues *i* are features and outcomes *j* are labels.
2. In LF learning, cues *i* are labels and outcomes *j* are features.

Eq. (2) is a discrepancy function that describes the amount of learning that will occur on a given trial; that is, the change in associative strength between a set of cues *i* and some outcome *j*.

$$\Delta V_{ij}^{n} = \alpha_i \beta_j (\lambda_j - V_{\mathrm{TOTAL}}) \tag{2}$$

An update rule is then used to calculate the change in associative strength between the set of cues *i* and the outcome *j* that results. This change is calculated as a function of their associative strength on the current trial:

$$V_{ij}^{n+1} = V_{ij}^{n} + \Delta V_{ij}^{n} \tag{3}$$

In these equations:

1. ΔV_{ij} is the change in associative strength between a set of cues *i* and an outcome *j* on a given trial *n*.
2. α is a parameter that allows individual cues to be marked as more or less salient. In our simulations, α was set to be constant; that is, all features were equally salient.
3. $\boldsymbol{\beta}$ is the parameter that determines the rate of learning for outcome *j*.
4. λ_j denotes the maximum amount of associative value (total cue value) that an outcome *j* can support. In our experiments, λ_j was set to ''1'' (when the outcome *j* was present in a trial) or ''0'' (when the outcome *j* was not present in a trial).
5. V_{TOTAL} is the sum of all current cue values on a given trial.

If there is a discrepancy between λ_j (the value of an outcome) and V_{TOTAL} (the current cue values), the value of that discrepancy will be multiplied against α and $\boldsymbol{\beta}$, and then this total will be added or subtracted from the associative strength of any cues present on that trial. In learning, the associative strength between a set of cues and an outcome increases in a negatively accelerated fashion over time, as the discrepancy between what is predicted and what is observed is gradually reduced. Given an appropriate learning rate, learning in the Rescorla–Wagner model asymptotes at a level that minimizes the sum-of-squares prediction error for the outcome over observed cue configurations.

Category learning was simulated in the Rescorla–Wagner model using abstract representations of the category structures in Fig. 7. The training set comprised three category labels and nine exemplar features (three of which were nondiscriminating features that were shared between exemplars belonging to different categories, and six of which were discriminating features that were not shared with members of another category; see Table 1).

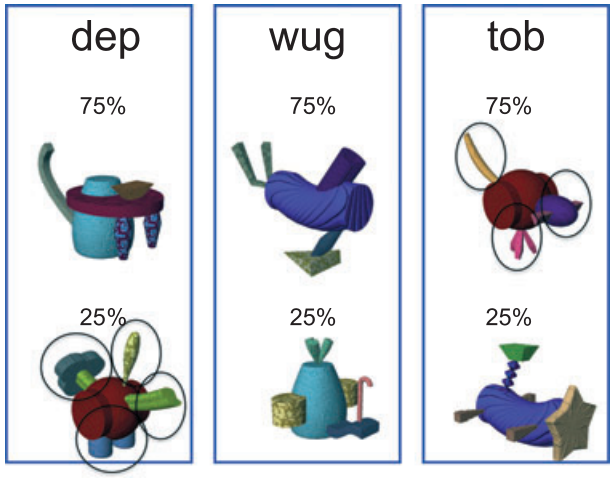

Fig. 7. The category structures employed in the simulations and in Experiment 1. (The stimuli are *fribbles* created by Michael Tarr's lab at Brown University.) Note that body type does not discriminate between categories. The sets of discriminating features that need to be learned in order to successfully distinguish the subcategories are circled on the low-frequency ''dep'' and high-frequency ''tob'' exemplars.

Table 1
The abstract representations of the category structures used to train the Rescorla–Wagner models

	Nondiscriminating Features			Discriminating Features					
	1	2	3	1	2	3	4	5	6
Category 1									
75%	1	0	0	1	0	0	0	0	0
25%	0	1	0	0	1	0	0	0	0
Category 2									
75%	0	1	0	0	0	1	0	0	0
25%	0	0	1	0	0	0	1	0	0
Category 3									
75%	0	0	1	0	0	0	0	1	0
25%	1	0	0	0	0	0	0	0	1

The frequency of the subcategories was manipulated so that each labeled category drew 75% of its exemplars from one subcategory and 25% of its exemplars from another subcategory. The two subcategories that made up each labeled category did not share any features, such that learning to correctly classify one of the subcategories paired with each label would provide no assistance with learning the other subcategory paired with that label. Finally, each low-frequency subcategory shared its nondiscriminating feature with the high-frequency exemplars of a different labeled category (see Table 1).

As a result, learning to correctly classify low-frequency exemplars necessitated learning to value the discriminating feature more than the nondiscriminating feature, despite its lower overall input frequency. Thus, this manipulation was designed to emphasize the problems with discrimination and response interference we hypothesized would result from LF learning, by creating a bias toward the misclassification of the low-frequency exemplars.

Two computational simulations were conducted. Training was configured as illustrated in the lower panels of Figs. 3 and 4, creating two networks of feature and label relationships. The first network learned associative weights in an FL sequence, with the nine exemplar features serving as cues and the three labels serving as outcomes. The second network learned associative weights in an LF sequence, with the three labels serving as cues and the nine features serving as outcomes. Each category had a high-frequency exemplar, presented on 75% of the training trials for that category, and a low-frequency exemplar, presented on 25% of the trials.

On each training trial a label and appropriate exemplar pattern were selected randomly to train each of the two networks. Training comprised 5,000 trials, which allowed learning to reach asymptote. The model has several free parameters that can affect learning. For simplicity, the simulations assumed equally salient cues and outcomes ($\alpha_i\boldsymbol{\beta}_j = 0.01$ for all i and j) and equal maximum associative strengths (=1.0).

To test the FL network, the exemplar features were activated to determine the subsequent activation of the labels. These activations were produced by assigning input values of 1 to the two features corresponding to each exemplar, and then propagating these values across the weights learned by the network to determine the associative values that had been learned for each label given those features. The Luce Choice Axiom (Luce, 1959) was used to derive choice probabilities for the three labels given these activations, revealing that the FL-trained network categorized and discriminated well; the probability of a correct classification for both the low- and the high-frequency exemplars was $p = 1$.

LF network testing involved activating the labels in order to determine subsequent activation of the features. These activations were produced by assigning input values of 1 to each of the labels, and then propagating these values across the weights learned by the network, to determine the associative values that had been learned for each feature. In order to assess the network's performance, the Euclidean distance between the predicted activations and the actual feature activations of the appropriate exemplar were calculated. For each label there were two sets of feature activations: those corresponding to the high- and low-frequency exemplars. To test learning of both exemplar types, a category and a frequency (either high or low) were selected, and the difference between the feature activations predicted by the network and the correct values for the three category exemplars with the same frequency (high or low) was computed. These differences were then converted to z-scores, and from these, the probabilities of selecting the correct exemplar given the category label were calculated as follows:

$$P(x) = \exp(-z(\mathrm{dist}(x, t)) \tag{4}$$

where $P(x)$ is the likelihood of the network selecting exemplar x, $z(\cdot)$ returns the z-score of its argument relative to its population, $dist(\cdot,\cdot)$ is the Euclidean distance function, and t is the

exemplar pattern generated by the network. The $P(x)$ likelihoods were normalized using the Luce Choice Axiom in order to yield normalized probability estimates. These revealed that, as predicted, in comparison with the FL network, the LF network performed poorly. At asymptote, the LF network failed to adequately discriminate the low-frequency items, predicting their correct feature patterns with only $p = .35$ confidence (i.e., the network was at chance on these trials). As expected, confidence was better for the high-frequency exemplars ($p = .75$).

Consistent with our hypothesis, a notable FLO effect was detectable in learning in the simulations: When features predicted labels, the model learned to discriminate exemplars from one another and categorized well; when the direction of prediction was reversed, performance was markedly poorer, especially with regard to the exemplars that were encountered in lower frequency.

It is important to note the role of cue competition and prediction error in these results. An examination of the weight matrices in the asymptotic networks at the end of the simulations revealed that the networks had developed very different *representations* of the categories. The associative weights in the LF-trained network (depicted schematically in Fig. 8B) reflect the absence of cue competition in training; the weights grew in proportion to the probability that the features were reinforced as the labels were presented, so the model learned the approximate conditional probability of each feature given each label (see Cheng, 1997; Wasserman et al., 1993).

In contrast, the weights learned by the FL-trained network (Fig. 8A) were shaped by cue competition between the exemplar features, and they show a bias toward associating the diagnostic (discriminating) features with the labels. This is because the network learned *inhibitory associations* between the nondiagnostic exemplar features and the category labels (the black squares or ''negative weights'' in Fig. 8A). Because each nondiagnostic feature appears in the exemplars of two categories, the presence of a nondiagnostic feature on a

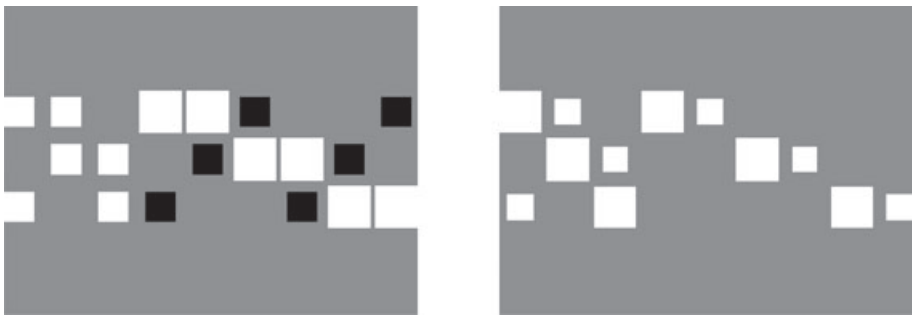

Fig. 8. Diagrams illustrating the representations learned by (A) an FL-trained network and (B) an LF-trained network after learning the categories in Table 1 to asymptote. The area of each square corresponds to the magnitude of a weight in the 3 × 9 matrix; white squares indicate positive associative weights and black squares negative associative weights. As can be seen, the FL-trained network has learned negative weights (black squares), whereas the LF-trained network has not.

given trial predicts both an incorrect label and a correct label. As only one label is actually presented on each trial, positive prediction of an absent label (V_i) produces greater activation of the absent label L than desired (as its activation level ought to be 0 on trials where it is not presented). Because the calculation of the discrepancy between the value of λ for the absent label (0) and the overpredicting cue values ($\lambda - V_i$) returns a negative value, it leads to a reduction in the associative value between any over-predicting cues and the absent label. Thus, ($\lambda - V_i$) for absent label L results in *latent learning* about label L.

It follows from this that the sum total of predictive value produced by both the nondiagnostic and diagnostic features for label L will decrease overall. This will in turn increase the discrepancy in the levels of expectation produced by the diagnostic and the nondiagnostic features for label L, leading to more learning. This learning will be shared between the two cues, such that the predictive value of the diagnostic feature for label L will benefit at the expense of the earlier error produced by the nondiagnostic feature. Over time this process will result in the bulk of the predictive value for each label shifting to the diagnostic feature.

Thus, learning in the networks is not confined to simply recording information about outcomes that are present at a given time but is also shaped by cue competition. The difference in performance between the two networks arises because error and cue competition result in discrimination learning in the FL-trained network, but this does not occur in the LF-trained network, which simply tracks cue probabilities.

In our simulation, the FL network learned to classify well because the configuration of cues in FL learning produced cue competition, allowing the network to learn to ignore the actual probability of labels given nondiagnostic features and invest predictive value in the diagnostic cues instead. It learned representations that traded completeness for discrimination. The LF network, in contrast, built up representations that provided a more veridical picture of the structure of the world (i.e., the actual cue probabilities), yet were of less value when it came to the task. While the LF network learned the actual probabilities in the task, the FL network learned to discriminate between the categories in it.

Does cue competition similarly affect human symbolic learning? Experiment 1 utilized the same category structures as the simulation to examine the FLO hypothesis in human learning.

7.1. Experiment 1

7.1.1. Participants

Thirty-two Stanford undergraduates participated for course credit.

7.1.2. Method and materials

Three experimental ‘‘fribble’’ categories were constructed that structurally matched the categories used in the simulations. Each comprised two subcategories clustered around both a high-saliency nondiscriminating feature (the fribble bodies in Fig. 7) and a set of lower saliency discriminating features (circled in Fig. 9). As in the computer simulations, the two subcategories that made up each labeled category did not share any features, and so learning to correctly classify one of the subcategories paired with each label provided no assistance

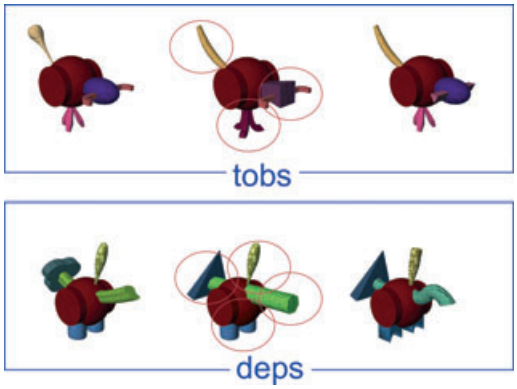

Fig. 9. Examples of high-frequency ''tobs'' (top row) and low-frequency ''deps'' (bottom row). The sets of features that discriminate between the two subcategories are circled on the second exemplar of each category.

with learning the other subcategory paired with that label. The nondiscriminating feature subcategories were again manipulated so that 75% of the exemplars of a category belonged to one subcategory, and 25% to another, and each nondiscriminating feature was shared by high-frequency exemplars of one category and low-frequency exemplars of another category. As in the simulation, learning to correctly classify low-frequency exemplars necessarily required learning to weigh the discriminating feature more than the nondiscriminating feature.

An extra control category was created in which all the exemplars shared just one, highly salient feature (all were blue). Because learning this category involved making a binary pairing *blue*⇔*bim*, there was no ''predictive structure'' to discover. In the absence of competing exemplars, learning was, in this case, predicted to be identical for LF and FL training (both were learned to ceiling in Rescorla–Wagner simulations). This category thus served to check that there were no differences in learning between the two groups other than those hypothesized.

Participants were asked to learn the names of ''species of aliens.'' To enforce LF or FL predictive relationships in training, we minimized participants' opportunities to strategize. As it is clear that the categories that typically make up symbolic systems are not explicitly taught, and children do not consciously strategize in language learning, this also offered the advantage of reproducing a more naturalistic symbolic learning environment for our participants (Deak & Holt, 2008; Jackendoff, 2002; Wittgenstein, 1953).

To achieve this, we trained participants on all four categories simultaneously, with the exemplars interspersed in a semi-randomized order so that the categories were presented in a nonpredictable sequence. Exemplars were presented for just 175 ms to inhibit participants' ability to consciously search for features (Woodman & Luck, 2003). LF-training trials comprised 1000 ms presentation of a label (''this is a wug''), followed by a blank screen for 150 ms, followed by 175 ms exposure to the exemplar. FL-training trials comprised 175 ms exemplar, 150 ms blank screen, and 1000 ms label (''that was a wug''). A 1000 ms blank screen separated all trials (Fig. 10). A training block comprised 20 different exemplars of

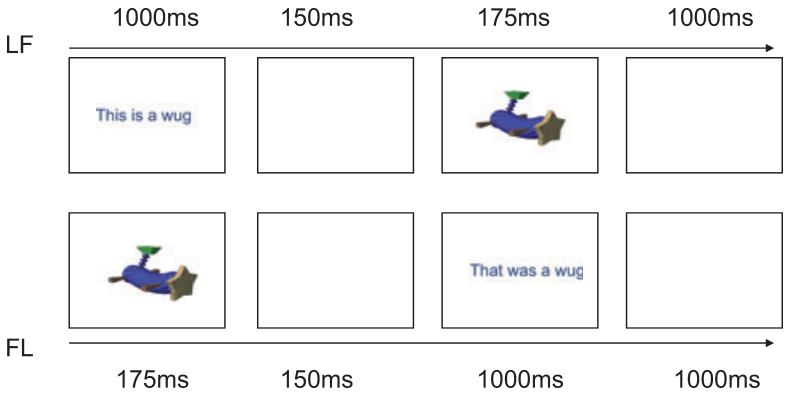

Fig. 10. The temporal (predictive) structure of the training trials in Experiment 1.

each experimental category—15 high-frequency exemplars and 5 low-frequency exemplars—and 15 control category exemplars. Training comprised two identical blocks, with a short rest between the blocks.

Testing consisted of speeded four alternative forced-choice tasks. Half the participants matched an unseen exemplar to the four category labels, and half matched a label to four previously unseen exemplars drawn from each category. To limit participants' ability to learn by contrasting between similar exemplars during testing, testing trials were composed either of all low-frequency or of all high-frequency exemplars plus control exemplars (this structure corresponded directly to the test trials in the computational simulation). Participants were instructed to respond as quickly as they could; if no answer had been recorded after 3500 ms, a buzzer sounded and no response was recorded. Each high- and low-frequency subcategory (and the control) was tested eight times, yielding 56 test trials.

7.1.3. Results and discussion

The results of Experiment 1 were remarkably consistent with those of the simulation; a 2×2 ANOVA revealed a significant interaction between exemplar frequency and training ($F(1,94) = 20.187$, $p < .001$; Fig. 11). The FL-trained participants classified high- and low-frequency items accurately (FL high $p = .98$; low $p = .78$), while the LF-trained participants only accurately classified high-frequency items ($p = .86$). Consistent with our predictions, LF-trained participants failed to classify the low-frequency exemplars—which comprised 50% of the test trials—above chance levels ($p = .36$, $t(47) = 0.536$, $p > .5$). The control category was learned to ceiling in both conditions. Analyses of confusability (i.e., the rates at which exemplars were *misclassified* to the category with which they shared nondiscriminating features) showed the same interaction between frequency and training ($F(1,94) = 8.335$, $p < .005$), with higher confusion rates after LF training ($M = 22.6\%$) than FL ($M = 6\%$; $t(16) = 5.23$, $p < .0001$). These differences were not due to a speed/accuracy trade-off; participants trained FL were faster as well as more accurate (LF $M = 2332$ ms, FL $M = 2181$ ms; $t(190) = 1.677$, $p < .1$).

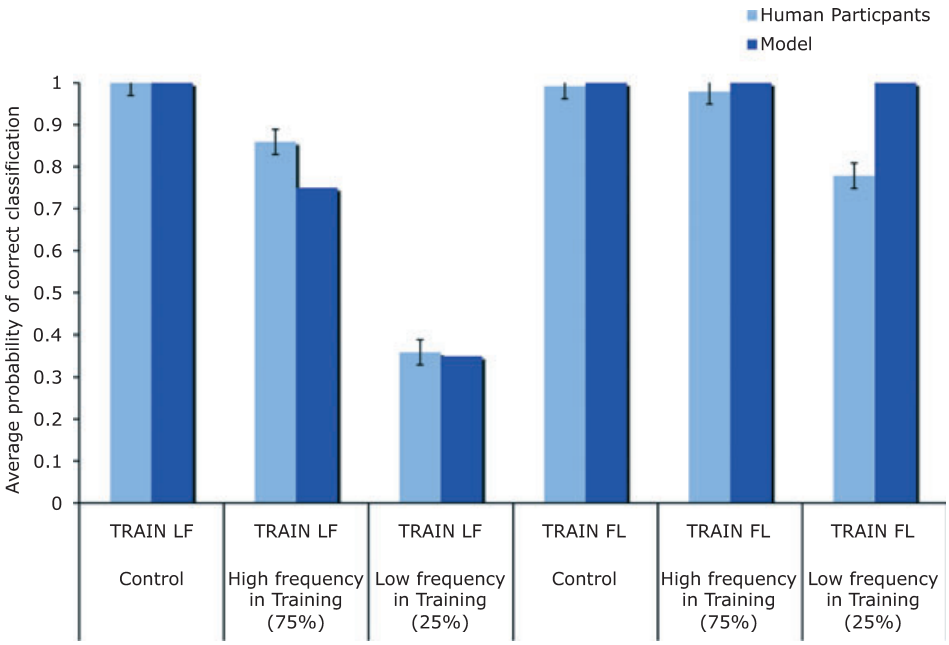

Fig. 11. The predictions of the simulation plotted against the performance of participants in Experiment 1.

Consistent with the analysis and simulations, these results reveal a strong FLO effect. When the predictive order of learning was features to labels (FL), participants learned to classify and discriminate the members of the categories with high levels of accuracy. When the predictive order was reversed, and labels served as cues to features (LF), participants trained on the same items performed poorly, failing to learn to correctly classify the low-frequency exemplars even though they had been exposed to exactly the same information in training as participants in the other condition. Only the order of labels and features in presentation was varied.

St. Clair, Monaghan, and Ramscar (2009) report an identical asymmetry in studies examining the consistent bias across languages for inflections to be added to word endings. A corpus analysis of English confirmed the prediction that suffixes are more informative about the grammatical category of root-words than prefixes, while an artificial language learning task revealed that suffixes (which are predicted by root-words, i.e., FL learning) were learned significantly more accurately than prefixes (which predict root-words, i.e., LF learning). Analogous asymmetries have also been noted in studies in which categories are learned during either inference or classification tasks (Markman & Ross, 2003; Yamauchi, Love, & Markman, 2002; Yamauchi & Markman, 1998) or where participants either describe or explain the properties of category exemplars (Williams & Lombrozo, 2009); it is likely that the principles we describe here underlie these effects (see also Love, Medin, & Gureckis, 2004).

8. Feature-Label-Order and children's color word learning

The effects of FLO offer a fresh perspective from which to consider phenomena associated with children's word learning. Take, for example, children's difficulties with learning words to describe colors. Although 4-month-olds can perceptually distinguish basic color categories (Bornstein, Kessen, & Weiskopf, 1976), young children struggle to learn to map the appropriate label to a given hue. Indeed, younger sighted children's use of color words is much like that of blind children (Landau & Gleitman, 1985); that is, while words like ''blue'' and ''yellow'' are in their vocabularies (color words are frequent in English), and are usually produced in appropriate contexts (e.g., ''yellow banana''), studies of the specific application of these words reveal that young children's use of them is haphazard. Three-year-olds who correctly identify a blue object in one situation may confuse ''blue'' with ''red'' in another (Sandhofer & Smith, 1999), and even at age 4, some children still struggle to discriminate color words appropriately despite hundreds of explicit training trials (Rice, 1980).

Why is learning English color words so difficult? The analysis we presented above contains at least one possible answer. In the wug and niz example in Fig. 1, above, we showed that the errors produced in FL learning actually help a child to learn to ignore the unhelpful association between *body* and wug and niz, and to focus instead on the associations between *red* and wug and *blue* and niz. Yet the relative ease with which a child might learn these mappings through FL learning is illustrative of why in ordinary circumstances, children may actually find color words difficult to master.

In our hypothetical example, we assumed that wugs and nizzes were encountered in isolation and labeled. Children first saw a single wug or niz exemplar in isolation, prior to the presentation of a label. A child would thus either see a niz and hear ''niz,'' or see a wug and hear ''wug.'' These are very helpful circumstances when it comes to learning to discriminate nizzes and wugs. If the presence of an actual niz led to the erroneous prediction of ''wug,'' the fact that the label ''wug'' did not follow would result in latent learning.

Compare this to a child in an ordinary setting hearing the word ''red'' or ''blue'' (or any other color words). In most of the everyday contexts in which children hear these words spoken, they will simultaneously be taking in a wide array of colors present in the surrounding environment. Some parts of a child's visual field will be receiving an input that will correspond to something that an adult might label with one color label, while another part of a child's visual field will be receiving an input that corresponds to something that an adult might label with another color label, and so on and so forth (and this is ignoring contrast, lighting, and other effects that compound the problem). Indeed, for an ordinary child in an ordinary context, it is likely that at any given time, cues that legitimately prompt the expectation of any and all of the high-frequency color labels will be present in the child's visual world. Thus, without some way of reducing the perceptual cues available at a given time, the child will encounter very few ''natural'' situations that will serve as optimal contexts for learning to discriminate between the various hues that might be associated with individual color words (e.g., a context in which the child can see only red and hears ''red''; see Fig. 12).

Discrimination learning of colors could be facilitated if language is used to narrow the child's focus from the environment as a whole to a specific object, thereby reducing the

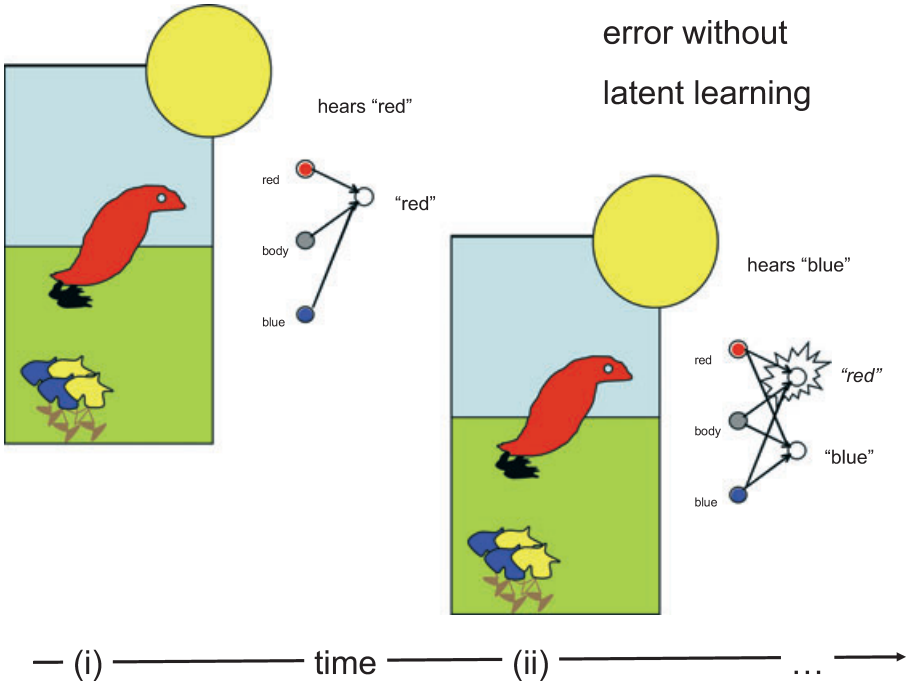

Fig. 12. A depiction of color-word learning in a natural environment. In word-learning, prediction error enables the learner to detect systematic covariance between features and labels. However, if color words are heard in contexts where most colors are available as cues, there will be little systematic covariance to discover. In this example, ''red'' is heard at time (i) and ''blue'' is heard at time (ii). In this context, a child will learn to associate both red and blue (and all of the colors present) with the labels ''red'' and ''blue'' indiscriminately. This may be why children more easily learn labels for objects than textures and colors (see Bloom, 2000).

number of conflicting perceptual cues. If a child's attention were to be drawn to an object at the mentioning of its name, the set of cues the child attends to might be narrowed to those associated with the object. So, for example, if a child hears ''the ball is red,'' attending to the red ball and then hearing ''red'' will increase her association between the property *red* and the word ''red.'' To the degree that the properties of the ball cue other color words, erroneous prediction of those words will result in them being dissociated from *red*, in much the way depicted in Fig. 1. Given that children learn object names before colors (Clark, 2009), placing color words in a postnominal position would create optimal conditions for discrimination learning (for a discussion of other effects of the structure of training on learning, see Sandhofer & Smith, 2001).

However, in English, color words occur *prenominally* roughly 70% of the time in speech (Thorpe & Fernald, 2006). When color words precede nouns in this way, the child's attention cannot be narrowed to focus on the object before the color word is heard. As a result, any and all of the perceptual features available will serve as potential cues to the color word, which in turn will serve to cue the noun. This may have the effect of leaving a color-word-learning child in a predicament similar to that illustrated in Fig. 12.

Cue competition and the FLO effect thus offer a possible explanation for a phenomenon that puzzled Charles Darwin:

> I carefully followed the mental development of my small children, and I was astonished to observe [that] soon after they had reached the age in which they knew the names of all the ordinary things that they appeared to be entirely incapable of giving the right names to the colors of a color etching. They could not name the colors, although I tried repeatedly to teach them the names of the colors. I remember quite clearly to have stated that they are color blind. (Darwin, 1877)

Objects are much more likely than colors to systematically covary with their labels. A dog, for instance, is likely to be present when a child hears ''doggy'' and absent when a child hears ''kitty,'' and vice versa, providing good conditions for latently unlearning the cues jointly associated with each. This is not necessarily the case for colors, as they often co-occur. It seems logical that children learn to label colors later than objects, as children have far more opportunities to learn to discriminate the cues associated with object words than with color words (see Gentner, 1982, 2006, for similar ideas).

To examine whether children's problems with learning color words arise out of this interaction between discrimination learning, the environment, and the way adults tend to talk about color, Experiment 2 engaged young learners in a conventional word-learning task in which we manipulated the sequencing of the color words and object labels. We predicted that learning to discriminate color words would be easier for children if a familiar object was named *predicting* the color label (FL) rather than vice versa (LF). For instance, hearing ''the cup is blue'' should better reinforce a child's mapping of *blue* to the correct hue than hearing ''the blue cup.'' Further, because children lack the ability to consciously strategize in language learning (Ramscar & Gitcho, 2007; Thompson-Schill, Ramscar, & Chrysikou, 2009), we expected that these effects would be evident even when items were presented and then described in natural speech. Thus, unlike in Experiment 1—where speeded presentation was used to enforce LF and FL learning in adults—our manipulation of children's learning in Experiment 2 exploited a natural variation in the ordering of labels in ordinary spoken English.

8.1. Experiment 2

8.1.1. Participants

Participants were 34 typically developing, monolingual English learners from 24 to 30 months old (M = 26 months) recruited from the Stanford area.

8.1.2. Method and materials

Experiment 3 comprised three phases: test-A (pretest), training, and test-B (posttest). Test-A and test-B were identical for each participant. Test materials comprised six sets of novel objects that were each of three colors (red, yellow, and blue). The objects were presented by set (i.e., the same object in each of the three colors), with the location of the correct object counterbalanced across trials. On each trial, children were asked to select

objects in response to questions in which the color word was either prenominal (''which is the red one?'') or postnominal (''which one is red?''). Question order was counterbalanced so that the same color never occurred on consecutive trials, and the form of the initial question (pre- vs. postnominal) alternated between participants.

In training, the children were introduced to a ''magic bucket'' containing five sets of items familiar to 26-month-olds (balls, cups, crayons, glasses, and toy bears) in each of the three colors. Half the children were presented with the items one by one and heard them labeled with color words used prenominally (''This is a red crayon''), while the other half were introduced to the items using postnominally presented color words (''This crayon is red''). Children then participated in test-B. All experimenters were blinded to the hypotheses.

Given the inconsistent nature of children's color word usage and comprehension, this design allowed a measure both of the probability that children would match color word and hue correctly on a given test, and also the probability that they would match color word and hue correctly and consistently on consecutive tests; that is, it measured the *consistent quality* of children's color word knowledge.

8.1.3. Results and discussion

Individual analysis of the children's performance in test-A and test-B showed that they had at least partial knowledge of the color words (confirming previous reports). Breaking down their responses by both training and test type, they averaged around two out of three correct choices in every test, significantly more than chance in all conditions (see Table 2). The performance of the FL-trained children improved slightly after training and repeated testing ($M = 57\%$; posttest $M = 63\%$), and their consistency was significantly above chance ($M = 46\%$, $t(101) = 2.502$, $p < .01$). However, the performance of the LF-trained children declined slightly after training and repeated testing (pretest $M = 58\%$; posttest $M = 51\%$, $t(102) = 10.61$, $p > .18$), and their consistency when tested with postnominal questions was *not* above chance ($M = 38\%$, $t(101) = 0.962$, $p > .2$). A 2 (pre- vs. posttest) × 2 (LF vs. FL training) ANOVA showed this interaction between training and testing to be significant ($F(2,201) = 193.78$, $p < .001$).

Interestingly, it appears the FL-trained children's improvement with training was tied to their also being tested with postnominal (''which one is blue?'') questions. Their

Table 2
Children's performance on the three alternate forced choice color matching task, broken down by testing and training type

Training Type	Question Type	Pretest Mean (%)	Posttest Mean (%)	Consistency Mean (%)	Consistency Versus Chance (33%)
FL-trained	Post-nominal	59	71	55	$t(51) = 3.065$, $p < .005$
FL-trained	Pre-nominal	57	55	37	$t(51) = 0.574$, $p > .5$
LF-trained	Post-nominal	59	53	39	$t(51) = 0.852$, $p > .4$
LF-trained	Pre-nominal	55	49	37	$t(51) = 0.574$, $p > .5$

FL, feature-to-label; LF, label-to-feature.

performance on these questions improved significantly after training (pretest $M = 59\%$; posttest $M = 71\%$, $t(51) = 1.948$, $p < .05$), whereas their performance on the prenominal questions did not (pretest $M = 57\%$; posttest $M = 55\%$).

These results support our suggestion that learning to match a color word to a specific color is affected by the sequencing of words and the predictive relationships that sequencing builds between words and objects or words and other words. Consistent with the predictions of our analysis and the FLO effects revealed in the simulations and Experiment 1, when a noun (object) cued a color label (FL), learning was far more successful than when a color label cued a noun (LF). In FL-conditions, the narrowing of the child's focus to the named object facilitated cue competition and learning, while in LF-conditions, the lack of focus and ubiquity of perceptual cues made learning far more difficult.

Other studies of FLO effects have produced similar patterns of results. For example, Ramscar, Dye, Witten, and Klein (2009) explored the way contextual discrimination might assist children in the Dimensional Change Card Sort Task (DCCS). In a training study with three groups of age-matched children under 4 years old, they found that children pretrained on ''color'' and ''shape'' games using FL-structured learning were later able to flexibly switch between the responses required by the DCCS. Children given LF-structured training performed far worse, performing little better than a control group that was given no training, and failed the DCCS as expected.

This analysis is also similar to the one presented by Sandhofer and Smith (1999, 2001), who suggested that color cognition development involves two steps: children first learn color words without making a proper mapping to the underlying color concepts, and then learn to use those color concepts nonlinguistically before mapping them to the appropriate words. Our data are consistent with this overall idea; however, it suggests that learning an underlying concept and learning the label it predicts may be largely aspects of the same process. ''Underlying concepts'' simply are the configurations of cues that allow regularities (events, affordances, labels) to be predicted.

In the analysis we present here, *ubiquity* plays a significant role in making color words hard to learn. The reasons are two-fold. First, the ubiquity of color makes learning a specific association between a hue and a label difficult because of the lack of covariance between individual hues and labels. Second, the ubiquity of color makes individual hues particularly salient and discriminatory, which may bias adult English speakers toward using color words prenominally. For example, given a red and blue crayon, color is an obvious way of discriminating between the two, meaning that ''red'' and ''blue'' are more useful things to convey to a listener than ''crayon.'' Following with this trend, the adjectives that tend to be used prenominally in English pick out properties that are both perceptually ubiquitous (Table 3) and difficult for children to learn (Clark, 2009).

Finally, although cross-linguistic studies have confirmed that the slow mastery of color terms is not unique to English (e.g., Roberson, Davidoff, Davies, & Shapiro, 2004), there has been no systematic investigation of the way that different languages affect this process. While some delays in color word learning may be peculiar to English and languages like it, we would expect the problems posed by the ubiquity of color (and other ubiquitous properties of the world) to affect children regardless of the language they are learning. Similarly,

Table 3
The distribution of the pre- and postnominal forms of 11 common adjectives in an analysis of two corpora from the CHILDES database. While children generally master adjectives with postnominal bias easily, they struggle with prenominally biased adjectives like ''big,'' ''red,'' and ''blue'' (Clark, 2003; Sandhofer & Smith, 1999)

	No. of Tokens	Prenominal		Postnominal	
		M	*SD*	*M*	*SD*
Big	821	.73	.06	.27	.06
Blue	138	.71	.10	.29	.10
Red	207	.67	.16	.33	.16
Pretty	96	.53	.22	.47	.22
Good	740	.52	.10	.48	.10
Nice	423	.51	.14	.49	.14
Dirty	10	.42	.12	.58	.12
Hot	147	.18	.12	.82	.12
Cold	109	.14	.12	.86	.12
Wet	96	.14	.24	.86	.24
Broken	180	.11	.11	.89	.11

while the FLO analysis does not rule out the possibility of languages in which color terms are exclusively prenominal (as opposed to where they are almost invariably postnominal, as in French, or the pattern is mixed, as in English), it suggests that the pattern of covariance of color words and nouns ought to be markedly different in languages where this is the case (see also Arnon & Ramscar, 2008, 2009).

9. Other Feature-Label-Order effects

Ramscar, Dye, and Yarlett (2009) examined whether the discrimination learning benefits seen in FL learning come at a cost to the representation of items at the input-level (or cue-level), as the FLO analysis suggests. FL learning distorts input representations in favor of predictive accuracy. In the wug/niz examples above, cue competition in FL learning effectively resulted in some wug and niz features not being learned. In contrast, while LF learning is less useful for discrimination, it *does* result in a statistically accurate picture of the world. Accordingly, one might expect participants to better recall training items when trained LF, but to better recall the classes they belong to when trained FL. Consistent with this, Ramscar, Dye, and Yarlett (2009) found that in fribble classification tasks in which participants were asked to discriminate correct object and category label pairings from mismatches, participants made more correct judgments and fewer false alarms for FL-trained categories than LF-trained categories. However, the pattern of judgments in the same participants was reversed when they were asked to *distinguish* the actual fribbles they saw in training from other fribbles from the same family; participants made more correct judgments and fewer false alarms for the exemplars of LF-trained categories than for FL-trained categories.

It follows from this that if FL and LF learning produce different representations of what is learned, then the way that people perceive similarities between objects should change as a result of FL and LF learning. Indeed, Ramscar and Dye (2009b) found that *dissimilar* objects that cued the same prediction increased in perceived similarity as a result of FL discrimination learning as compared to objects learned LF. However, *similar* objects that shared overlapping features with objects that were predictive of other outcomes, were perceived to be less similar after FL-training than after LF-training. This is consistent with the idea that FL-training reduces the saliency of features that produce error, leading to a reduction in the number of alignable features in a similarity comparison (Markman, 1996).[4]

10. General discussion

Our analysis of the potential effects of cue competition on word learning led us to predict FLO effects: When features served as cues to labels in learning (FL), we expected that learning would lead to accurate classification and discrimination between categories; when labels served as cues to objects (LF), we expected classification and discrimination to be relatively poor. This asymmetry was strikingly evident in the results of our simulations and human experiments. In these tasks we have shown that FLO has a dramatic effect on a learner's ability to learn and use labels. Set-size differences in the number of cues and outcomes when learning from objects to labels versus labels to objects result in different levels of discrimination learning and asymmetries in the cognitive representations produced by the two forms of learning.

The analysis and results we present suggest that the semantic categories people use to understand and communicate about the world can *only* be learned if labels are predicted from objects, a finding at odds with referential theories of language. This finding is, however, consistent with both the logic of abstraction and with the mathematical constraints of information and coding theory. Abstracting from the world (a large set of information) to labels (a reduced information set) respects the constraints imposed by the process of abstraction, whereas abstracting from labels to the world does not; similarly, using the features of the world to encode labels respects the mathematical constraints of information and coding theory, whereas using labels (infinitesimally small sets of bits) to uniquely encode features of the world (a massive set of bits) violates these constraints (see e.g., Abramson, 1963; Kolmogorov, 1965; Rodemich, 1970; Shannon, 1948). In both these domains, the problem for LF learning is that the way in which the larger set (of features) is to be ''got'' or discriminated from the smaller set (of labels) is necessarily underspecified. In abstraction, the problem arises because, by definition, the information discarded in the process of abstraction is not encoded in the resulting abstraction and is thus unrecoverable; in coding, it arises because there is a bound on the amount of information that can be encoded in any set of bits (Rodemich, 1970; Shannon, 1948).

As elucidated at the outset, philosophical analyses of reference have long encountered similar problems with the notion that a word can point to its meaning, with some of the most influential philosophers of the 20th century arguing that the process of recovering a specific

meaning from a given encounter with a word is *necessarily* underspecified (Quine, 1960; Wittgenstein, 1953). To put it another way, the very idea of reference makes little sense philosophically (see also Fodor, 1998), and it is inconsistent with learning, coding, and information theory (which are the basic principles of computation). On the other hand, treating symbolic learning as a process of learning to predict symbols from features (and symbolic processing as one of predicting symbols given sets of features) is entirely compatible with learning, coding, and information theory, as well as philosophical analyses of the problem of reference (Quine, 1960; Wittgenstein, 1953).

Reference is usually taken as a given in contemporary theories of language and cognition. All ''top-down'' combinatorial, syntactic theories of language—in which meaning is governed by the structures into which lexical items are combined—assume that lexical items contribute their meanings to these structures by reference to underlying concepts or semantic representations. However, as we—and many others—have elucidated, reference is a philosophically and theoretically bankrupt hypothesis. We believe that given the shortcomings of reference, our demonstrations of FLO effects in learning have important implications for our understanding of the nature of categorization and for theories of reference and language. We briefly review some of these below.

11. Similarity, discrimination, and categorization

Our discussion of learning in this paper has focused on learning to label objects, and on classifying objects into labeled categories. In reality, of course, people learn far more about their environments than just the labels languages assign to things. While we learn, for example, that objects of a certain form might be called ''chairs,'' we also learn that we can sit on chairs, that we can stand on them to reach high objects, and that, should we find ourselves in a brawl in a Wild West saloon, chairs will shatter over the back of a black-hatted aggressor in an aesthetically pleasing, fight-ending manner. We also learn to extend these inferences by analogy, such that we might sit on a box when there is no chair to be had, or we might strike a saloon brawler with a bar stool when a chair is out of reach.

To the extent that the kinds of inferences we make about the environment are more or less discrete—that is, that the set of inferences we learn to make is smaller than the set of cues available in the environment, and that these inferences can be readily discriminated from one another—there is reason to believe that FLO-like effects, and the representational principles that accompany them, would be discernable in other learning tasks. Given that the FLO effects we describe in word learning will apply more generally to categorization, it is worth noting some important differences between the approach to category learning that emerges from our analysis and standard approaches to categorization and language acquisition.

First, ''learning to see the similarities between things'' is not an important aspect of category learning from this perspective. The most important aspect of the learning process in the simulations and the experiments we conducted was *discrimination*, and the reason that LF-training was generally poorer than FL-training was that items that were not discriminated were treated equivalently (or similarly). In word learning, from this perspective,

''similarities'' between objects do not need to be discovered; on the contrary, things appear similar because learning has not yet discriminated them.

The idea that similarity is not really relevant to categorization may seem alarming. Similarity is usually seen as a very important mechanism in category learning (see Hahn & Ramscar, 2001; Murphy, 2002; Goldstone & Son, 2005, for reviews). However, the idea that similarity drives categorization suggests that the concept of *Red Burgundy* is learned through noticing *similarities* in the red wines of *Chambolle-Musigny*, *Fixin*, *Santenay*, *Gevrey-Chambertin*, *La Tache,* and *Corton*; or that children learn the concept of *dog* by noticing similarities between *spaniels*, *dachsunds*, *shar-peis*, etc. Yet discrimination seems to be more consistent with the ways in which people actually learn about these things: that the distinction between *Fixin* and *Gevrey-Chambertin* is usually learned after one knows what *Red Burgundy* is, and that learning about their differences is the key part of the process. The same point can be made with regard to spaniels and shar-peis, apples and pomegranates, etc. Instead of discovering the similarities between different examples of dogs, most children initially apply the label indiscriminately, calling many animals ''dogs'' (Clark, 2009) *prior* to discriminating dogs from other animals. Despite the almost universal belief among cognitive scientists concerning the importance of similarity in categorization, we would argue that discrimination offers a far better fit to our experience of how words and their meanings are learned.

Second, learning is not confined to recording information about events that co-occur at a given time. Learning about things that are not present, as a result of overprediction errors, is a second, critical source of information in symbolic learning. Indeed, as discrimination learning is driven by prediction errors, much of the learning that takes place is about things that are not present (the nonoccurrence of expected events, i.e., learning in response to negative evidence).

These points are worth stressing because many psychologists believe the opposite of all these things: that discovering an appropriate similarity metric is a key aspect of category learning (Goldstone & Son, 2005; Hahn & Ramscar, 2001; Murphy, 2002); that children learn without ''negative evidence'' and without systematic information about how words are *not* used (see e.g., Bloom, 2000); and, thus, that learning must come from ''positive evidence'' alone (see Xu & Tenenbaum, 2007, for a review).

12. Symbols and reference

As we noted earlier, the FLO-effect, and the set size problem in symbolic learning we have revealed, can be seen as one more reason (among many compelling reasons) to believe that reference cannot form the basis for word learning and word usage. This does not mean that we do not use words to ''refer'' to things, as in ''look at that chair!'' but rather that formally, reference works predictively: the imperative ''look at that chair!'' works communicatively because there is an object present that successfully cues the word ''chair'' (see also Clark & Wilkes-Gibbs, 1986; Tanenhaus & Brown-Schmidt, 2008, provide a detailed review of the evidence supporting this idea).

Despite the many profound problems that have been pointed out regarding the formal idea of reference (Quine, 1960; Wittgenstein, 1953), it remains the case that symbolic thought is most widely conceived of in referential terms. Indeed, it is ironic that in perhaps the single most influential paper in the history of cognitive science, Chomsky (1959) largely ignores the many cogent criticisms of reference laid out by Skinner (1957); instead, he criticizes Skinner for proposing a theory that neither explains reference nor accommodates Chomsky's fundamentally referential, syntactic view of language. However, as noted earlier, at no point have contemporary referential theories proposed any solutions to the many problems of reference. The claim is simply that reference *must* work (e.g., Chomsky, 2000; Fodor, 1983, 1998).

The problems reference poses, and the difficulty of accounting for language within the frame of a combinatorial syntax, have led Fodor and Chomsky to make the absurdist claim that all concepts—including ''carburetor'' and ''bureaucrat''—must be innate:

> there is good reason to suppose that the [nativist] argument is at least in substantial measure correct even for such words as carburetor and bureaucrat, which, in fact, pose the familiar problem of poverty of stimulus if we attend carefully to the enormous gap between what we know and the evidence on the basis of which we know it. The same is often true of technical terms of science and mathematics, and it surely appears to be the case for the terms of ordinary discourse. However surprising the conclusion may be that nature has provided us with an innate stock of concepts, and that the child's task is to discover their labels, the empirical facts appear to leave open few other possibilities. (Chomsky, 2000, pp. 65–66)

While it is difficult to falsify this kind of nativist claim, our experiments show that children's acquisition of color words—and adults' acquisition of wug and tob—are easily explicable in terms of learning. Further, the pattern of results in our experiments is hard to square with the idea that the world is packaged into discrete categories, or that people possess innate concepts that allow words to be learned by reference, as these ideas would predict that the directional, set size effects that were strongly in evidence throughout our experiments ought not to occur. Instead, our results suggest that learning theory can inform both our understanding of symbolic representation and our understanding of the relationship between symbols and their meanings. In particular, learning provides a way of conceptualizing and formalizing the many problems that have been raised regarding theories of reference in a predictive framework, and can offer an alternative to the idea that children are innately endowed with concepts like ''allele,'' ''polyploidy,'' ''stoichiometric coefficient,'' and ''floating-point operator.''

13. Language and the nature of symbolic systems

> Holmes: ''How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?''
>
> —*The Sign of Four*, Sir Arthur Conan Doyle

The data and analysis we present here, along with the array of evidence and arguments against reference that predate it, can be seen as providing constraints on the way language itself is conceptualized. Understanding language in terms of learning—and without underspecified appeals to reference—involves a reassessment of what human communication involves, requiring revised theories of language and its role in culture (Quine, 1960; Tomasello, 1999, 2003; Wittgenstein, 1953; see also Fodor, 2000). Learning and abstraction appear to constrain the possible relationships between symbols and their meanings to that of predicting symbols from meanings. It follows therefore that all meaningful symbolic processes—including language comprehension and production—must be predictive. To return to our earlier example of abstraction, the idea that a meaning can be conveyed by a word makes no more sense than the idea that someone might be able to ''get'' detailed information about the results and method sections of a paper she had never seen, simply by reading its abstract. Given an abstract, one can only make guesses about the results and methods sections, making a kind of prediction about the kind of information they might contain. If the reader is an expert, the likelihood that these predictions will be more accurate, or even substantially correct, will increase. However, *given no more than an abstract*, *the reader can do no more than make predictions*, because the process of abstraction involves discarding information that cannot be later recovered from an abstract representation. Symbols are abstractions, and it follows similarly that symbolic meanings can only be inferred. Language must be a predictive process.

While this view is at odds with contemporary ''big picture'' views of language, it is highly compatible with an enormous array of empirical findings relating to language processing and the nature of language. Numerous results in a variety of research paradigms have revealed that when people are listening to or reading a sentence they build up a rich set of linguistic *expectations*, predicting upcoming words based on the structure and semantics of the prior discourse (see e.g., Altmann & Mirković, 2009; DeLong, Urbach, & Kutas, 2005; Garrod & Pickering, 2009; Kutas & Federmeier, 2007; Norris & McQueen, 2008; Pickering & Garrod, 2007; Tanenhaus & Brown-Schmidt, 2008). While this may seem obvious in idiomatic phrases, such as ''cross my heart and hope to ___'' or ''hit the nail on the ___'' a considerable body of evidence suggests this is true of language more generally. For example, DeLong et al. (2005) measured event-related potentials (ERPs) while participants read sentences like ''the day was breezy so the boy went outside to fly a kite'' and ''the day was breezy so the boy went outside to fly an airplane.'' Not only did they find that the less predictable *airplane* produces a larger n400 than *kite* (n400 is an ERP component typically elicited by an unexpected linguistic stimulus), but the same pattern held for the articles *a* and *an* as well. This suggests that participants were using context to predict not just *kite* but also the article that preceded it, causing them to find *an* surprising as a result.

While studies to date have often focused on anticipation of a specific word, object, or event based on prior context (see e.g., Altmann & Steedman, 1988; DeLong et al., 2005; Otten & Van Berkum, 2008; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), it is clear that in natural speech, listeners are anticipating (probabilistically) a

range of different possible words that might follow in a given speech stream (Norris & McQueen, 2008). In much of the literature these predictions are seen as *assisting* the meaningful production and comprehension of language, but from the perspective of symbolic learning described here—and without a predetermined idea of ''reference''—these predictions can be seen as actually *comprising* the meaningful use of language (see also Ramscar et al., in press).

There is a widespread consensus that ''intention reading''—social prediction—is an important component of word learning (see e.g., Bloom, 2000; Tomasello, 2003, 2008); our proposal is that ''intention reading'' can be extended to language processing more generally. We argue that comprehension arises both out of what the listener *knows*—what she will predict—and what the listener *is learning*—what she will come to predict. In order to predict a speaker, even partially, a listener must activate cues similar to those used by the speaker in generating an utterance. As the speaker uses FL-learned cues to generate speech, the listener will activate similar FL-learned cues, allowing her to anticipate upcoming speech. When the listener does not fully predict the speaker's words—as will often be the case—further FL learning will take place. Comprehension, like learning, is thus a process of having and modifying expectations.

The most likely way a listener can predict a speaker is by sharing a cultural ''common ground'' (Clark, 1996) of cues to words, and the more accurately a listener can predict a speaker in a given context, the more likely it is that the listener is using a similar set of contextual cues. Accordingly, meaning will be shared to the degree that predictions are shared, and communication can thus be seen as a process of aligning a speaker and a listener's predictions (see also Davidson, 1986; Garrod & Pickering, 2009; Pickering & Garrod, 2007). Successful communication is a process of reducing uncertainty in a listener's predictions as an utterance or dialog unfolds. We should note that ''predicting'' does not mean ''predicting with certainty.'' From this perspective, understanding is inherently probabilistic: When the degree of prediction is too low, little communication is occurring, and the listener is likely confused; when prediction is too high, communication is occurring, however, the listener is likely deeply bored.

If symbolic communication involves predicting symbols from meanings (and context)—and we have outlined many reasons for assuming that it does—then meaning is something that a speaker elicits in a listener simply by engaging the listener in a game of prediction. In this game, symbols are not used to *convey* meaning, but rather are used to reduce a listener's uncertainty about a speaker's intended message (Shannon, 1948). In order for a listener to predict a speaker, the listener has to activate the same semantic cues to symbolic form as the speaker, such that the listener comes to understand an utterance by *thinking* about that utterance in a way that converges on that of the speaker. This proposal has much in common with the idea that language is a form of joint action (see e.g., Altmann & Mirković, 2009; Clark, 1993; Garrod & Pickering, 2009; Gennari & MacDonald, 2009; Pickering & Garrod, 2007; Tanenhaus & Brown-Schmidt, 2008); it differs in that it is explicitly nonreferential.

14. Prediction, abstraction, and language

Treating language as a process of symbol prediction is consistent with the idea that language serves—very literally—to *abstractly* represent things for the purpose of communication. Because learning itself can be seen as a process of abstraction, learning theory can begin to explain how languages come to serve as abstract representational systems, offering insights into the nature of abstract thought and constraints on the way that abstract thought is to be conceptualized and understood. Learning and cue competition describe the process by which aspects of the world are ''summarized'' in sound symbols. The view of language that emerges is fundamentally directional and predictive. However, while we noted above that there is considerable evidence showing that predictive processes are pervasive in language processing, the idea that all language processes are predictive has been subject to much criticism.

Principal among these criticisms is the issue of data sparsity: It is claimed that the amount of data required to reliably estimate the parameters of any useful probabilistic model far outstrips the amount of language exposure any person could reasonably receive, not just in childhood, but in an entire lifetime. The most acute form of this objection involves unseen events: In any sample of language to which a learner is exposed, many legitimate linguistic constructions will not occur, because language is such a complex, productive system. Simple probabilistic models based on the principle of maximum likelihood will assign these events a probability of 0, incorrectly implying that they are (probabilistically speaking) impossible.

Miller and Chomsky (1963) were among the first authors to consider this problem in detail, and their analysis is taken by many as providing a demonstration of the impossibility of a probabilistic account of language. They pointed out that for an *n*-gram model to represent the intricacies of even a moderate proportion of English sentences, an apparently unlearnable number of statistical parameters would need to be estimated:

> Just how large must *n* and *V* be in order to give a satisfactory model? Consider a perfectly ordinary sentence: The people who called and wanted to rent your house when you go away next year are from California. In this sentence there is a grammatical dependency extending from the second word (the plural subject people) to the seventeenth word (the plural verb are). In order to reflect this particular dependency, therefore, n must be at least 15 words. We have not attempted to explore how far *V* can be pushed and still appear to stay within the bounds of common usage, but the limit is surely greater than 15 words; and the vocabulary must have at least 1,000 words. Taking these conservative values of *n* and *V*, therefore, we have V^n = 1,045 parameters to cope with, far more than we could estimate even with the fastest digital computers. (p. 430; We have altered Miller & Chomsky's notation from *K* and *d* to *V* and *n* for clarity.)

Thus, as *n* increases, the number of potential parameters to be estimated grows as V^n, where *V* is the number of tokens in the language. We can contextualize this (conservative) estimate of the number of parameters that any language model might need to estimate in various ways. For example, it has been estimated that a child with professional parents will only be

exposed to around 11 million spoken words a year (or 11×10^6; Hart & Risley, 1995), and that the average lifetime consists only of around 2.2×10^9 s. By either measure, the figure of 10^{45} advanced by Miller and Chomsky is astronomically larger!

Because of Zipf's law (which holds that the probability that a given word will follow a given word is inversely proportional to its rank in the frequency table; Zipf, 1935, 1949), the number of probabilities to be estimated over time thus grows exponentially—guaranteeing that an *n*-gram model, when its probabilities are estimated by maximum likelihood, will assign a probability of 0 to many perfectly legitimate *n*-gram transitions, simply because the appropriate transitions were not encountered in the language sample a learner was exposed to. As Miller and Chomsky (1963) put it:

> We know that the sequences produced by *n*-limited Markov sources cannot converge on the set of grammatical utterances as *n* increases because there are many grammatical sentences that are never uttered and so could not be represented in any estimation of transitional probabilities. (p. 429)

While Miller and Chomsky's argument has long been taken as conclusive, it makes assumptions about the nature of learning and symbolic representation that, if examined, make these conclusions far from compelling: First, the argument assumes that parameters must be estimated without any generalization from prior experience (although generalization appears to be evident in every other aspect of human and animal learning; Shepard, 1987); and second, it assumes, following from this, that the representations over which probabilistic estimations are computed are confined to ''words'' (while the idea that words are the basic units of linguistic processing has strong intuitive appeal, the number of words in the critical dependency in Miller and Chomsky's own example twice exceeds Miller's famous 7 ± 2 estimate of the *limit* of information processing capacity; Miller, 1956).

There are many reasons to believe that both of Miller and Chomsky's assumptions are wrong. People clearly learn the kind of parameters that Miller and Chomsky assume are unlearnable, and what's more, they acquire a knowledge of linguistic probabilities across a range of linguistic abstractions, including phonemes, words, and larger linguistic structures. There is a wealth of evidence showing, for example, that people are highly sensitive to the probability of phoneme-to-phoneme transitions within the words of a language (Marslen-Wilson, 1975, 1987; McClelland & Elman, 1986; Norris et al., 1995, 1997, 2003), and that language users acquire a probabilistic understanding of language at many degrees of abstraction (see e.g., Altmann & Steedman, 1988; Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Borensztajn, Zuidema, & Bod, 2009; Bresnan, 2007; Bybee, 2002, 2006; DeLong et al., 2005; Hagoort & Van Berkum, 2007; Johnson, 1997; Nakatani & Gibson, 2008, in press; Otten & Van Berkum, 2007, 2008; Pierrehumbert, 2001, 2003; Tanenhaus et al., 1995; Tily et al., in press; see also Baayen & Moscoso del Prado Martin, 2005; Bod, Hay, & Jannedy, 2003; Bybee & Hopper, 2001; Hale, 2003; Jurafsky & Martin, 2009; Levy, 2008; Manning & Schütze, 1999; Plag & Baayen, 2009; Ramscar et al., in press). Consistent with this, there is evidence that children acquire a probabilistic understanding of language at multiple levels of abstraction (Bannard, Lieven, & Tomasello, 2009; Bannard & Matthews,

2008; Goldstein & Schwade, 2008; Trueswell & Gleitman, 2007; Trueswell, Papafragou, & Choi, in press). (Indeed, the evidence suggests that the ''representational units'' of language are not words, as Miller and Chomsky suppose, but rather that linguistic representations are in themselves probabilistic and are better understood in terms of the predictability of linguistic regularities in context rather than as fixed units; this is consistent with many approaches to language in which the traditional idea of the word as a symbol is extended to embrace representations at multiple levels of abstraction; Bannard et al., 2009; Bod, 2009; Goldberg, 2006; Goldberg & Jackendoff, 2004, 2005; Langacker, 1987; Milin, Kuperman, Kostic, & Baayen, 2009; Pollard & Sag, 1994; Tomasello, 2003) Further, in the field of natural language processing (NLP), concrete methods have been devised for estimating linguistic parameters, allowing probabilistic language models to be learned from psychologically realistic language samples (Chen & Goodman, 1996, 1998; Goodman, 2001; Yarlett, 2008; see also Aylett & Turk, 2004; Charniak & Genzel, 2002; Cover & King, 1978; Van Son & Pols, 2003).

There appear to be no principled objections to the idea that language is grounded in mutual prediction. The evidence suggests that people are capable of estimating the probabilities of linguistic events, that they generalize this knowledge to unseen events (Berko, 1958), and that they use this knowledge to make predictions at every conceivable level of language processing. By revealing the pervasive role of expectation in linguistic processing, these studies—and others like them—offer support for the predictive framework suggested by our analysis. Indeed, many of these results may be better understood in terms of the analysis presented here.

15. Learning, convention, and communication (or why a baby learns language, but not a dog)

Presenting a predictive account of symbolic representation and communication grounded in learning theory raises a number of questions: How do people come to learn the sets of conventionalized cues that enable mutual prediction and communication (Wittgenstein, 1953)? If communicative conventions—and language—are the product of learning, why is symbolic communication apparently solely the preserve of humans? Why, given that they share the same environment, do babies and not pet dogs learn language? Framing symbolic knowledge in terms of learning and prediction offers a concrete perspective in which to consider these questions.[5]

One potentially fruitful source of answers to these questions comes from considering the differences between the way that the brains of humans and other animals develop, and then considering the impact these differences in development have on learning. Like many other primates, humans are born with an immature brain. Birth is followed by synaptogenesis (the proliferation of synapses) followed by an extended pruning period (synaptic elimination). Brain development in humans, however, is markedly different from that of other primates. In monkeys, the postnatal development of the brain occurs at the same rate in all cortical areas (Rakic, Bourgeois, Eckenhoff, Zecevic, & Goldman-Rakic, 1986). In contrast, human

cortical development is uneven: Synaptogenesis in the visual and auditory cortex peaks a few months after birth, while the same developments occur later in the prefrontal cortex (Huttenlocher & Dabholkar, 1997; see Ramscar & Gitcho, 2007; Thompson-Schill, Ramscar, & Evangelia, 2009, for reviews).

One behavioral consequence of slow prefrontal development is that children appear unable to engage in behaviors that conflict with prepotent responses, and they appear unable to filter their behavior and their learning (Thompson-Schill et al., 2009). In adults, prefrontal control mechanisms bias responses and attention according to goals or context (Yeung, Botvinick, & Cohen, 2004); the prefrontal cortex functions as a dynamic filter, selectively maintaining task-relevant information and discarding task-irrelevant information (Shimamura, 2000). The *absence* of this capacity in young children can be illustrated by contrasting their performance with that of adults on biased selection tasks, such as guessing the hand an M&M is in (the hands are biased 25:75). Children up to age 5 tend to overmatch, fixating on the high-probability ''good'' hand. After age 5, however, a probability matching strategy emerges (Derks & Paclisanu, 1967). This is one of the rare instances in which children's inability to think flexibly is actually to their advantage (probability matching actually *reduces* the number of M&Ms won by children over 5 years old and adults).

Another area of learning in which cognitive flexibility may well prove disadvantageous is in *convention learning*. Symbolic knowledge is, in its essence, conventional. Given a symbol, a social animal needs to be able to infer and understand (and often, to do) the *appropriate thing* in the *appropriate context*. For this to happen, ''symbolic values,'' must be both conventionalized and internalized (Wittgenstein, 1953). In learning the appropriate cues to predicting symbols, this is far more likely to happen if learners are unable to filter their attention during the course of learning. Given a similar set of cues and labels to learn, learners will tend to sample the environment in much the same way, and thus it is more likely that they will come to have similar expectations regarding the relationship between cues and symbols. We argue that these shared expectations are in large part what linguistic conventions *are* (see also Wittgenstein, 1953).

In contrast to children, adults struggle to master new linguistic conventions (Johnson & Newport, 1989). This may reflect an inevitable handicap that adults' increased ability to selectively respond or attend to the world imposes on convention learning. Development appears to increase the complexity of the human learning architecture, allowing learners to filter their attention in learning, and this dramatically reduces the potential for conventions to arise in the way we describe above (Ramscar & Gitcho, 2007; Thompson-Schill et al., 2009). The greater the variety there is in what adults focus on in learning, the less conventionality there will be in what they learn. Conversely, the less children are able to direct their attention in learning, the more their learning will approximate the basic error-driven learning process we have described here. The representations children learn will be shaped more straightforwardly by their immediate physical, social, and linguistic environment, and their learning about common regularities in that environment will be more conventional (see also Hudson Kam & Newport, 2005, 2009; Singleton & Newport, 2004).

Given that adults were once children, these considerations may explain why other animals—which appear to be able to selectively attend to their environments and filter their

responses from almost the moment they are born—fail to learn much by way of complex, conventionalized social and linguistic behavior.[6] These considerations may further allow us to begin to describe what precisely the genetic endowment that makes humans *the* symbolic species actually amounts to.

16. Summary: symbols, prediction, and meaning

> One keeps hearing the remark that philosophy really makes no progress, that the same philosophical problems that had occupied the Greeks are still occupying us. … The reason [why this is true] is that our language has remained the same and seduces us into asking the same questions over and over. As long as there is a verb 'to be' which seems to function like 'to eat' and 'to drink,' as long as there are adjectives like 'identical,' 'true,' 'false,' [and] 'possible', as long as one talks about a flow of time and an expanse of space, etc., etc., humans will continue to bump up against the same mysterious difficulties, and stare at something that no explanation seems able to remove. (Wittgenstein, 1993, p. 185)

The analysis, simulations, findings, and phenomena we report here offer insight into the ways in which symbolic knowledge is learned, represented, and processed, and into how the very concept of *concept*, itself, is to be conceptualized. In the traditional analyses of ''concept'' and ''reference'' that dominate thinking in cognitive science, ''concept'' is treated as a noun akin to *dog*, and ''refer'' as a verb akin to *point*. In the same vein, it is often assumed that a symbol *conveys* meaning in a manner akin to the way that a train conveys goods. None of these analyses stand up to theoretical and empirical scrutiny. Our analysis points to an alternative characterization of these ideas: that *concepts* are better understood in terms of the probabilistic knowledge used in making predictions (rather than as discrete units of representation), and that *reference* should be understood in terms of the way that words function as tools in a predictive human practice (rather than as pointers to the world). These proposals are not new ways of approaching traditional problems in cognitive science, but rather represent a fundamental reanalysis of the problems cognitive science seeks to solve.

The traditional approach to symbolic thought in cognitive science begins at the macrolevel, with an analysis of one way in which words (as discrete, referential symbols that pick out discrete concepts) might come to be conjoined in meaningful sequences (Chomsky, 1957). This referential, combinatorial analysis of symbolic thought ultimately produces a picture of thought and language in which much of the structure and content of symbolic thought must be assumed to be innate and inscrutable (Anderson & Lightfoot, 2002; Chomsky, 2000; Fodor, 1998). Despite what might be seen as alarming consequences of this analysis, the referential, combinatorial approach to symbolic thought dominates current theorizing about the mind. (This is perhaps because many researchers who subscribe to the referential, combinatorial analysis of symbolic thought do not subscribe to nativism; however, as Fodor, 1998, 2000, has noted, referential theories that avoid nativism tend to do so by avoiding any explanation of how reference is supposed to work at all.)

In the analysis presented here, we started at a micro-level, considering how the associations between things in the world and the symbols that we use to represent them in abstract are learned. The picture of symbolic thought—and of language—that this analysis ultimately produces is very different to any received view of both. Conceiving of symbolic thought in terms of prediction does not promise an account of how referential, combinatorial symbols, or the syntax by which they are combined, are learned. Instead, it promises an alternative conception of language and symbolic thought. We hope that novelty will not dissuade readers from appreciating the consistency and coherency of this view—or at least, that resistance be tempered by giving the problems posed by reference the same degree of consideration.

We do not claim to have offered a ''complete'' version of the alternative here, though we should note that a ''complete'' version of the referential, combinatorial analysis of language has not been forthcoming either (for further evidence that treating language as prediction can assist in understanding and accounting for the learning of aspects of language that are usually seen as syntactic, see Arnon & Ramscar, 2008, 2009; Ramscar & Dye, 2009a; Ramscar & Yarlett, 2007). What we have tried to offer here is a detailed account of some basic principles of symbolic learning, and to sketch out how these might flesh out into the alternative conception of thought and language we envisage. We have also sought to show how our conception of symbolic thought and language in terms of prediction is consistent with both a large and growing body of evidence arising out of the study of language and speech processing, and with the majority of successful approaches to language engineering.

From a theoretical standpoint, we hope that the results and analysis we describe can provide food for thought for other researchers, and that they might be inspired to consider the very different conceptions of thought and language we describe as an alternative avenue of investigation. Practically, we believe our analysis offers a framework that can allow educational and linguistic training to be structured in ways that make them more efficient and effective, enabling people to better utilize and refine their capacity for symbolic thought.

Notes

1. An outcome can be considered *discrete* to the extent that it is discriminated from other outcomes in prediction.
2. For simplicity, we consider only object features as cues in FL learning in this discussion; in the real world, the cues in FL learning features would include anything present in the environment prior to the occurrence of a label, including other labels (i.e., words; see Arnon & Ramscar, 2009; Ramscar, Matlock, & Dye, in press).
3. In tonal languages, the phoneme in conjunction with its tone might be taken as a similarly discrete unit.
4. It is highly unlikely that the way people form representations of the world is as simple and straightforward as this brief picture might imply. The brain appears to make use of redundant coding in multiple memory systems to compensate for the costs of

learning (see Ramscar & Dye, 2009b). That these differences are noticeable suggests that the amount of redundancy in the coding of representations is limited (see also Barlow, 2001).

5. This contrasts starkly with the nativist alternative, which holds that this matter is (once more) inscrutable: ''As far as is known, even the most rudimentary properties of the initial and attained states [of language acquisition] are not found among other organisms or, indeed, in the biological world, apart from its points of contact with inorganic matter. Nor are there more than very weak relations to anything discovered in the brain sciences. So we face the problems of unification that are common in the history of science and do not know how—or if—they will be resolved'' Chomsky (2000, p. 79).
6. It is interesting to note that the bonobo Kanzi—the most successful non-human-learner of a conventionalized symbol system—was accidentally introduced to that system as a neonate (Savage-Rumbaugh & Lewin, 1994).

Acknowledgments

This material is based upon work supported by the National Science Foundation under grant nos. 0547775 and 0624345 to Michael Ramscar. We are grateful to Roddy Lindsay, Brad Love, Gordon Bower, Bruce McCandliss, Sharon Thompson-Schill, Delphine Dahan, Colin Bannard, Nick Davidenko, and Daphna Shohamy for discussion of these ideas, and Art Markman and three anonymous reviewers for much insightful feedback on earlier drafts of this paper.

References

Abramson, N. (1963). *Information theory and coding*. New York: McGraw-Hill.

Altmann, G. T. M., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, *33*, 1–27.

Altmann, G. T. M., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, *30*, 191–238.

Amundson, J. C., & Miller, R. R. (2008). CS–US temporal relations in blocking. *Learning and Behavior*, *36*(2), 92–103.

Anderson, S. R., & Lightfoot, D. W. (2002). *The language organ: Linguistics as cognitive physiology*. Cambridge, England: Cambridge University Press.

Arnon, I., & Ramscar, M. (2008). *How order-of-acquisition shapes language learning: The case of grammatical gender*. Paper presented to the Boston University Conference on Language Development, Boston, MA.

Arnon, I., & Ramscar, M. (2009). *Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned*. Proceedings of the 31st Annual Conference of the Cognitive Science Society, Amsterdam.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*(1), 31–56.

Baayen, R. H., & Moscoso del Prado Martin, F. (2005). Semantic density and past-tense formation in three Germanic languages. *Language*, *81*, 666–698.

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences USA*, *106*(41), 17284–17289.

Bannard, C., & Matthews, D. E. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, *19*, 241–248.

Barlow, H. (2001). Redundancy reduction revisited. *Network: Computation in Neural Systems*, *12*, 241–253.

Bell, A., Brenier, J., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111.

Berko, J. (1958). The child's learning of English morphology. *Word*, *14*, 150–177.

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

Bod, R. (2009). From exemplar to grammer: A probailistic analogy-based model of language learning. *Cognitive Science*, *33*(5) , 752–793.

Bod, R., Hay, J., & Jannedy, S. (Eds.) (2003). *Probabilistic linguistics*. Cambridge, MA: MIT Press.

Borensztajn, G., Zuidema, W., & Bod, R. (2009). Children's grammars grow more abstract with age—Evidence from an automatic procedure for identifying the productive units of language. *Topics in Cognitive Science*, *1*, 175–188.

Bornstein, M. H. (1985). On the development of color naming in young children: Data and theory. *Brain and Language, 26, 72–93.*

Bornstein, M. H., Kessen, W., & Weiskopf, S. (1976). Color vision and hue categorization in young human infants. *Journal of Experimental Psychology*, *2*, 115–119.

Bresnan, J. (2007). Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In S. Featherston, & W. Sternefeld (Eds.), *Roots: Linguistics in search of its evidential base generative grammar* (pp. 77–96). Berlin: Mouton de Gruyter.

Bybee, J. (2002). Phonological evidence for exemplar storage of multiword sequences. *Studies in Second Language Acquisition*, *24*(2), 215–222.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, *82*(4), 529–551.

Bybee, J., & Hopper, P. (Eds.) (2001). *Frequency and the emergence of linguistic structure*. Typological studies in language, Vol. 45. Amsterdam: John Benjamins.

Charniak, E., & Genzel, D. (2002) Entropy rate constancy in text. In *Proceedings of the Association for Computational Linguistics (ACL02)* (pp. 199–206). Somerset, NJ: Association for Computational Linguistics.

Chen, S. F., & Goodman, J. T. (1996). An empirical study of smoothing techniques for language modeling. In *Proceedings of the Association for Computational Linguistics (ACL96)* (pp. 310–318). Somerset, NJ: Association for Computational Linguistics.

Chen, S. F., & Goodman, J. T. (1998). *An empirical study of smoothing techniques for language modeling*. Technical Report TR-10-98, Computer Science Group. Cambridge, MA: Harvard University.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405.

Chomsky, N. (1957). *Syntactic structures*. The Hague, The Netherlands: Mouton.

Chomsky, N. (1959). Review of verbal behavior, by B.F. Skinner. *Language*, *35*, 26–57.

Chomsky, N. (2000). *New horizons in the study of language and mind*. Cambridge, England: Cambridge University Press.

Clark, E. V. (1993). *The lexicon in acquisition*. Cambridge, England: Cambridge University Press.

Clark, E. V. (2003). *First language acquition*. Cambridge, England: Cambridge University Press.

Clark, E. V. (2009). *First language acquisition* (2nd ed.). Cambridge, England: Cambridge University Press.

Clark, H. H. (1996). *Using language*. Cambridge, England: Cambridge University Press.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*, 1–39.

Cover, T. M., & King, R. C. (1978). A convergent gambling estimate of the entropy of English. *IEEE Transactions on Information Theory 24*(4), 413–421.

Darwin, C. (1877). *Biographiche skizze eines kleinen kindes*. *Kosmos*, 367–376.

Davidson, D. (1986). A nice derangement of epitaphs. In E. LePore (Ed.), *Truth and interpretation: Perspectives on the philosophy of Donald Davidson* (pp. 433–436). Oxford, England: Blackwell.

Deak, G. O., & Holt, A. (2008). Language learning. In H. L. Roediger, III (Ed.), *Learning and memory: A comprehensive reference, volume 2: Cognitive psychology of memory* (pp. 557–578). Oxford, England: Elsevier.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*, 1117–1121.

Derks, P. L., & Paclisanu, M. I. (1967). Simple strategies in binary prediction by children and adults. *Journal of Experimental Psychology*, *73*(2), 278–285.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Fodor, J. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.

Fodor, J. (1998). *Concepts: Where cognitive science went wrong*. New York: Oxford University Press.

Fodor, J. (2000). *The mind doesn't work that way: The scope and limits of computational psychology*. Cambridge, MA: MIT Press.

Fodor, J., & Pylyshyn, F. (1988). Connectionism and cognitive architecture. *Cognition*, *28*(1-2), 3–71.

Gallistel, C. R.. (2001). *Mental representations. Encyclopedia of the behavioral and social sciences*. New York: Elsevier.

Gallistel, C. R. (2003). Conditioning from an information processing perspective. *Behavioural Processes*, *62*, 89–101.

Gallistel, C. R., & Gibbon, J. (2000). Time, rate and conditioning. *Psychological Review*, *107*, 344–389.

Garrod, S., & Pickering, M. J. (2009). Joint action, interactive alignment and dialogue. *Topics in Cognitive Science*, *1*, 292–304.

Gennari, S. P., & MacDonald, M. C. (2009). Linking production and comprehension processes: The case of relative clauses. *Cognition*, *111*, 1–23.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. A. Kuczaj (Ed.), *Language development: Vol. 2. Language, thought, and culture* (pp. 301–334). Hillsdale, NJ: Erlbaum.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*, 155–170.

Gentner, D. (2003). Why we're so smart. In D. Gentner, & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and cognition* (pp. 195–236). Cambridge, MA: MIT Press.

Gentner, D. (2006). Why verbs are hard to learn. In K. Hirsh-Pasek, & R. Golinkoff (Eds.), *Action meets word: How children learn verbs* (pp. 544–564). New York: Oxford University Press.

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

Gluck, M. A., & Bower, G. H. (1988b). Evaluating an adaptive network model of human learning. *Journal of Memory and Language*, *27*, 166–195.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in Language*. Oxford, England: Oxford University Press.

Goldberg, A. E., & Jackendoff, R. (2004). The English resultative as a family of constructions. *Language*, *80*, 532–568.

Goldberg, A. E., & Jackendoff, R. (2005). The end result(ative). *Language*, *81*(2), 474–477.

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science*, *19*, 515–522.

Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the non-independence of features in similarity judgments. *Cognitive Psychology*, *23*, 222–264.

Goldstone, R. L., & Son, J. Y. (2005). Similarity. In K. J. Holyoak, & R. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 13–36). Cambridge, England: Cambridge University Press.

Goodman, N. (1972). *Problems and projects*. Indianapolis: Bobbs-Merril.

Goodman, J. T. (2001). A bit of progress in language modeling. *Computer Speech and Language*, *15*(4), 403–434.

Hagoort, P., & Van Berkum, J. J. A. (2007). Beyond the sentence given. *Philosophical Transactions of the Royal Society. Series B: Biological Sciences*, *362*, 801–811.

Hahn, U., & Ramscar, M. (2001). *Similarity and categorization*. Oxford, England: Oxford University Press.

Hale, J. (2003). The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, *32*(2), 101–123.

Hart, B., & Risley, T.. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Paul H. Brookes.

Haugeland, J. (1985). *Artificial intelligence: The very idea*, Cambridge, MA: MIT Press.

Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, *1*, 304–309.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*, 151–195.

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, *59*(1), 30–66.

Hume, D. (1740). *A treatise of human nature* (1967 ed.). Oxford, England: Oxford University Press.

Huttenlocher, P. R., & Dabholkar, A. S. (1997). Regional differences in synaptogenesis in human cerebral cortex. *Journal of Comparative Neurology*, *387*(2), 167–178.

Jackendoff, R. S. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford, England: Oxford University Press.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson, & J. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–165). San Diego, CA: Academic Press.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*, 60–99.

Jurafsky, D., & Martin, J. H.. (2009). *Speech and language processing: An introduction to natural language processing, speech recognition, and computational linguistics* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B. Campbell, & R. Church (Eds.), *Punishment and aversive behaviour* (pp. 279–296). New York: Appleton-Century-Crofts.

Kolmogorov, A. N. (1965). Three approaches to the definition of the quantity of information. *Problems of Information Transmission*, *1*, 3–11.

Kuhl, P. K. (1994). Learning and representation in speech and language. *Current Opinion in Neurobiology*, *4*, 812–822.

Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences USA*, *97*(22), 11850–11857.

Kurtz, K. J., Gentner, D., & Gunn, V. (1999). Reasoning. In D. E. Rumelhart & B. M. Bly (Eds.), *Cognitive science: Handbook of perception and cognition* (2nd ed., pp. 145–200). San Diego, CA: Academic Press.

Kutas, M., & Federmeier, K. D. (2007). Event-related brain potential (ERP) studies of sentence processing. In G. Gaskell (Ed.), *Oxford handbook of psycholinguistics* (pp. 385–406). Oxford, England: Oxford University Press.

Landau, B., & Gleitman, L. R. (1985). *Language and experience: Evidence from the blind child*. Cambridge, MA: Harvard University Press.

Langacker, R. (1987). *Foundations of cognitive grammar*. Stanford, CA: Stanford University Press.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309–338.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: The MIT Press.

Markman, A. B. (1996). Structural alignment in similarity and difference judgments. *Psychonomic Bulletin and Review*, *3*(2), 227–230.

Markman, A. B. (1999). *Knowledge representation*. Mahwah, NJ: Erlbaum.
Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, *129*, 592–613.
Marslen-Wilson, W. D. (1975). Sentence perception as an interactive parallel process. *Science*, *189*, 226–228.
Marslen-Wilson, W. (1987). Functional parallelism in spoken word recognition. *Cognition*, *25*, 71–102.
McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.
McLaren, I. P. L., & Mackintosh, N. J. (2000). An elemental model of associative learning: I. Latent inhibition and perceptual learning. *Animal Learning & Behavior*, *28*, 211–246.
Milin, P., Kuperman, V., Kostic, A., & Baayen, R. H. (2009). Paradigms bit by bit: An information theoretic approach to the processing of paradigmatic structure in inflection and derivation. In J. P. Blevins & J. Blevins (Eds.), *Analogy in grammar: Form and acquisition* (pp. 214–252). Oxford, England: Oxford University Press.
Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.
Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla–Wagner model. *Psychological Bulletin*, *117*(3), 363–386.
Miller, G. A., & Chomsky, N. (1963). Finitary models of language users. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. II, pp. 419–491). New York: John Wiley.
Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.
Nakatani, K., & Gibson, E. (2008). Distinguishing theories of syntactic expectation cost in sentence comprehension: Evidence from Japanese. *Linguistics*, *46*(1), 63–86.
Nakatani, K., & Gibson, E. (in press). An on-line study of Japanese nesting complexity. *Cognitive Science* .
Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395.
Norris, D., McQueen, J. M., & Cutler, A. (1995). Competition and segmentation in spoken-word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(5), 1209–1228.
Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238.
Norris, D., McQueen, J. M., Cutler, A., & Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology*, *34*(3), 191–243.
O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review*, *108*, 311–345.
Otten, M., & Van Berkum, J. J. A. (2007). What makes a discourse constraining? Comparing the effects of discourse message and scenario fit on the discourse-dependent N400 effect. *Brain Research*, *1153*, 166–177.
Otten, M., & Van Berkum, J. J. A. (2008). Discourse-based word anticipation during language processing: Prediction of priming? *Discourse Processes*, *45*, 464–496.
Pearce, J. M. (1987). A model of stimulus generalization for Pavlovian conditioning. *Psychological Review*, *94*, 61–73.
Pearce, J. M. (1994). Similarity and discrimination: A selective review and a connectionist model. *Psychological Review*, *101*, 587–607.
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*, 532–552.
Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral & Brain Science*, *31*, 109–178.
Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*, 105–110.
Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency effects and the emergence of lexical structure* (pp. 137–157). Amsterdam: John Benjamins.

Pierrehumbert, J. (2003). Probabilistic phonology: Discrimination and robustness. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probability theory in linguistics* (pp. 177–228). Cambridge, MA: The MIT Press.

Plag, I., & Baayen, R. H. (2009). Suffix ordering and morphological processing. *Language*, *85*, 106–149.

Pollard, C., & Sag, I. (1994). *Head-driven phrase structure grammar*. Stanford, CA: Center for the Study of Language and Information.

Quine, W. V. (1960). *Word and object*. Cambridge, MA: MIT Press.

Rakic, P., Bourgeois, J. P., Eckenhoff, M. F., Zecevic, N., & Goldman-Rakic, P. S. (1986). Concurrent overproduction of synapses in diverse regions of the primate cerebral cortex. *Science*, *232*(4747), 232–235.

Ramscar, M., & Dye, M. (2009a). *Expectation and negative evidence in language learning: The curious absence of mouses in adult speech*. Proceedings of the 31st Annual Conference of the Cognitive Science Society, Amsterdam.

Ramscar, M., & Dye, M. (2009b). *No representation without taxation: The costs and benefits of learning to conceptualize the environment*. Proceedings of the Analogy 09 Conference, Sofia, Bulgaria.

Ramscar, M., Dye, M., Witten, J., & Klein, J. (2009). *Two routes to cognitive flexibility: Learning and response conflict resolution in the Dimensional Change Card Sort Task*. Proceedings of the 31st Annual Conference of the Cognitive Science Society, Amsterdam.

Ramscar, M., Dye, M., & Yarlett, D. (2009). *No representation without taxation: The costs and benefits of learning to conceptualize the environment*. Proceedings of the 31st Annual Conference of the Cognitive Science Society, Amsterdam.

Ramscar, M., & Gitcho, N. (2007). Developmental change and the nature of learning in childhood. *Trends in Cognitive Science*, *11*(7), 274–279.

Ramscar, M., Matlock, T., & Dye, M. (in press). Running down the clock: The role of expectation in our understanding of time and motion. *Language and Cognitive Processes*.

Ramscar, M., & Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science*, *31*, 927–960.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist*, *43*(3), 151–160.

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* (pp. 64–99). New York: Appleton-Century-Crofts.

Rice, N. (1980). *Cognition to language*. Baltimore, MD: University Park Press.

Roberson, D., Davidoff, J., Davies, I. R. L., & Shapiro, L. R. (2004). The development of color categories in two languages: A longitudinal study. *Journal of Experimental Psychology: General*, *133*, 554–571.

Rodemich, E. R. (1970). Covering by rook-domains. *Journal of Combinatorial Theory*, *9*, 117–128.

Rosch, E. (1978). Principles of categorization. In E. Rosch, & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillside, NJ: Lawrence Erlbaum Publishers.

Rosenblatt, F.. (1959). Two theorems of statistical separability in the Perceptron. In V. Blake & A. M., Uttley (Eds.), *Mechanisation of thought processes: Proceedings of a symposium held at the National Physical Laboratory* (pp. 419–456). London: Her Majesty's Stationery Office.

Rudy, J. W. (1974). Stimulus selection in animal conditioning and paired-associate learning: Variations in the associative process. *Journal of Verbal Learning & Verbal Behavior*, *13*, 282–296.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L.. (1986). A general framework for parallel distributed processing. In D. E., Rumelhart, & J. L., McClelland (Eds.), *Parallel distributed processing: Explorations in the microarchitecture of cognition* (Vol. I, pp. 44–55). Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning past tenses of English verbs. In D. E. Rumelhart, & J. L. McClelland (Eds.), *Parallel distributed processing: Vol 2: Psychological and biological models* (pp. 216–271). Cambridge, MA: MIT Press.

Rumelhart, D. E., & Zipser, D.. (1986) Feature discovery by competitive learning. In D. E. Rumelhart, & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations* (pp. 51–193). Cambridge, MA: MIT Press.

Saffran, J. R. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language*, *44*, 493–515.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*, 27–52.

Sandhofer, C., & Smith, L. B. (1999). Learning color words involves learning a system of mappings. *Developmental Psychology*, *35*, 668–679.

Sandhofer, C., & Smith, L. B. (2001). Why children learn color and size words so differently: Evidence from adults' learning of artificial terms. *Journal of Experimental Psychology: General*, *130*, 600–620.

Savage-Rumbaugh, S., & Lewin, R. (1994). *Kanzi: The ape at the brink of the human mind*. Toronto: John Wiley and Sons.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Shimamura, A. P. (2000). The role of prefrontal cortex in dynamic filtering. *Psychobiology*, *28*, 207–218.

Siegel, S., & Allan, L. G. (1996). The widespread influence of the Rescorla–Wagner model. *Psychonomic Bulletin and Review*, *3*, 314–321.

Singleton, J. L., & Newport, E. L. (2004). When learners surpass their models: The acquisition of American sign language from inconsistent input. *Cognitive Psychology*, *49*, 370–407.

Skinner, B. F. (1957). *Verbal behavior*. New York: Appleton-Century-Crofts, Inc.

St. Clair, M., Monaghan, P., & Ramscar, M. (2009). Relationships between language structure and language learning: The suffixing preference and grammatical categorization. *Cognitive Science*, *33*(7), 1317–1329.

Tanenhaus, M. K., & Brown-Schmidt, S. (2008). Language processing in the natural world. In B. C. M. Moore, L. K. Tyler, & W. D. Marslen-Wilson (Eds.), *The perception of speech: From sound to meaning*. *Transactions of the Royal Society B: Biological Sciences*, *363*, 1105–1122.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. E. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632–1634.

Thompson-Schill, S. L., Ramscar, M., & Chrysikou, M. (2009). Cognition without control: When a little frontal lobe goes a long way. *Current Directions in Psychological Science 8*(5), 259–263.

Thorpe, K., & Fernald, A. (2006). Knowing what a novel word is not: Two-year-olds ''listen through'' ambiguous adjectives in fluent speech. *Cognition*, *100*, 389–433.

Tily, H., Gahl, S., Arnon, I., Snider, N., Kothari, A., & Bresnan, J. (2009). Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language and Cognition*, *1*(2), 147–165.

Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Tomasello, M. (2008). *Origins of human communication*. Cambridge, MA: MIT Press.

Trueswell, J. C., & Gleitman, L. R. (2007). Learning to parse and its implications for language acquisition. In G. Gaskell (Ed.), *Oxford handbook of psycholinguistics* (pp. 635–657). Oxford, England: Oxford University Press.

Trueswell, J. C., Papafragou, A., & Choi, Y. (in press). Syntactic and referential processes: What develops? In E. Gibson & N. Pearlmutter (Eds.), *The processing and acquisition of reference*. Cambridge, MA: MIT Press.

Van Son, R., & Pols, L. (2003). How efficient is speech? *Proceedings of the Institute of Phonetic Sciences*, *25*, 171–184.

Waelti, P., Dickinson, A., & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, *412*, 43–48.

Wasserman, E. A., Elek, S. M., Chatlosh, D. L., & Baker, A. G. (1993). Rating causal relations: The role of probability in judgments of response-outcome contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 174–188.

Werker, J. F., & Tees, R. C. (1984). Phonemic and phonetic factors in adult cross-language speech perception. *Journal of the Acoustical Society of America*, *75*(6), 1866–1878.

Williams, J. J., & Lombrozo, T.. (2009) *Explaining promotes discovery: Evidence from category learning*. Proceedings of the 31st Annual Conference of the Cognitive Science Society, Amsterdam.

Wittgenstein, L. (1953). *Philosophical investigations*. Oxford, England: Blackwell.

Wittgenstein, L. (1993). In J. C., Klagge, & A., Nordmann (Eds). *Philosophical occasions 1912–1951*. London: Blackwell.

Woodman, G. F., & Luck, S. J. (2003). Electrophysiological measurement of rapid shifts of attention during visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *29*, 121–138.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*, 245–272.

Yamauchi, T., Love, B. C., & Markman, A. B. (2002). Learning non-linearly separable categories by inference and classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 585–593.

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language*, *39*, 124–149.

Yarlett, D. (2008). Language learning through similarity-based generalization. Unpublished PhD Thesis, Stanford University.

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review*, *111*(4), 931–959.

Zipf, G. K. (1935). *The psychobiology of language*. New York: Houghton-Mifflin.

Zipf, G. K. (1949). *Human behavior and the principle of least-effort*. Cambridge, MA: Addison-Wesley.