Surprise in the Learning of Color Words

Michael Ramscar (michael@psych.stanford.edu)
Department of Psychology, Building 420
Stanford, CA 94305

Kirsten Thorpe (thorpe@psych.stanford.edu)
Department of Psychology, Building 420
Stanford, CA 94305

Katie Denny (katie32@stanford.edu)
Department of Psychology, Building 420
Stanford, CA 94305

Abstract

In two experiments, we investigate whether, in English, children's difficulty learning color adjectives might stem from their overwhelming tendency to be used in pre-nominal position (e.g., "blue cup"), where word order causes the adjective to arrive ahead of its meaning, rather than post-nominal presentation (e.g., "the cup is blue"), where the meaning cues the word. We consider factors of sequencing from the point of view of models of animal learning. Results indicate that children as young as 2 years begin to reliably master color words when hearing them in training presented post-nominally, but not pre-nominally, and that adults challenged with learning novel color categories are affected by the same ordering effects. We suggest that children's difficulty with color word learning is in part due to the challenge of having to make predictions *from* words *to* the properties they refer to, rather than being able to make predictions *from* the world *to* the words.

Keywords: learning; language; color words; modeling.

Introduction

In the course of the first few years of their lives, children successfully master the use of hundreds of words. Words for objects, actions, emotions, relationships, states of affairs and myriad other aspects of the social, physical and metaphysical worlds they encounter. For some reason, however, in spite of their impressive feats in most other domains, when it comes to learning color words, children perform abjectly. So miserable were his children at learning color names relative to other words that Darwin (1877) initially concluded they were color-blind. Though this is incorrect – 4 month old children can distinguish between basic color categories (Bornstein, Kessen, & Weiskopf, 1976) – Darwin can hardly be faulted for his observation: young children use color words in the same way that blind children do (Landau & Gleitman, 1985): words such as "blue" and "yellow" appear in children's expressive language in answer to answer questions such as "what color is this?" but their mapping to individual colors is haphazard and interchangeable (e.g., Sandhofer & Smith, 1999). If shown a blue cup and asked about its color, typical 2-year-olds seem as likely to answer "red" as "blue." Even as late as age 4, sighted children may still be unable to accurately sort objects by color, even after hundreds of explicit training trials (Rice, 1978).

In English, while word order for color adjectives varies, they are used overwhelmingly in prenominal position (e.g., "blue cup"), such that word order causes the adjective to arrive before its meaning, rather than post-nominal presentation (e.g., "the cup is blue"), where the meaning cues the word. We hypothesize that children's incompetence at color word learning in English is not caused by any unique property of color, or indeed, of the world. Rather, it reflects an interaction between the very mechanisms that enable children to be such masterful word learners in the first place, and the way that these mechanisms interact with these order effects as color words are presented in real-time.

Although it may surprise linguists that the relative order of words matters so much in learning, these effects are consistent with many models of animal learning. In these models, animals do not learn "facts" about the world. Rather, they learn cues that *predict* events; learning involves strengthening representations of the values of these cues (this is true for both "associative" and "information processing" models, e.g., Rescorla & Wagner, 1972; Gallistel & Gibbon, 2000). Thus animal learning can be characterized – irrespective of the particular framework one is examining – as the process by which useful predictive structure in the world is discovered. For example, the Rescorla-Wagner (1972) learning model is a model of elementary associative learning that has been employed in many contexts and used to explain a wide variety of learning effects both in animals and humans (e.g., Gluck and Bower, 1988). In the model, the discovery process can be broadly characterized as follows: an animal observes that a series of cues (CSs) occur, and based on the predictive (or associative) strength that these CSs have acquired relative to various environmental outcomes (USs) which can subsequently

occur, the animal is able to anticipate the likelihood of observing each of these USs. After a delay, one or more of the USs actually occurs, and based on the degree to which the set of USs was mispredicted – informally, based on the degree to which the animal was 'surprised' by the outcome – learning takes place. In particular, the amount of learning, which is manifested in the change of associative strengths between CS and US elements, is proportional to the size of the discrepancy between the actual outcome and the outcome predicted by the animal, such that more learning happens when events are poorly predicted than when they are well predicted. The outcome of this process is far from the simple binary Stimulus-Response process which is often used to characterize Pavlovian conditioning, instead, what is learned by this process are sets of cues in the environment, and the complex sets of relations between them that allow an animal to represent and better understand its surroundings (Rescorla, 1988). Learning thus depends both on what is observed to occur, and also the strength to which the set of actual outcomes was predicted, based on (the strength) of what has already been learned. As a consequence, when multiple cues (CSs) are present, cues compete with one another in order to explain subsequent US outcomes. For example, if one cue already serves as a perfectly reliable predictor of an outcome, an additional cue observed in conjunction with the original cue at a later time will not become strongly associated with the outcome, even if it is also a perfect predictor of the outcome. This is because the predictive strength of the first cue is such that the animal is never 'surprised' by the outcome, so there is no error-signal (discrepancy between observed and expected outcome) to drives changes in the predictive values of the set of cues present (such 'blocking' is well established in animal and human learning).

As a result of these interactions between cues, the outcome of learning under the Rescorla-Wagner model is far from the simple binary stimulus-response process that is often used to characterize Pavlovian conditioning (and which did indeed accurately characterize early linear models of learning). Instead, the Rescorla-Wagner model proposes that when an animal encounters important regularities in its environment, what is learned are both the cues, and the complex relations between them, that allow the animal to represent, predict and understand that environment (Rescorla, 1988).

Three elements are key to learning in the Rescorla-Wagner model: (1) the availability of identifiable regularities to drive learning (the prediction of regularly identifiable, behaviorally important aspects of the environment such as events, objects, etc., is the **goal** of the learning process under this analysis); (2) inadequate anticipation of these regularities, which drives learning (violation of expectation is the key **mechanism** that initiates learning; evidence suggests that the neural realization of this error signal may be a dopamine response originating in the midbrain; Hollerman and Schultz, 1998); and (3) the availability of cues, which allow the prediction of the outcomes (the discovery of the correct configuration of cues to successfully predict an outcome is the **product** of learning).

How might these elements relate to word or symbol learning? If learning is a process by which the predictors of behaviorally important outcomes in the environment are identified, the words of a language might be considered as both behaviorally important cues *and* outcomes. To take a benign example, a word such as "chair" is a regularity that applies to the somewhat varied set of entities on the world we happen to call chairs. In other words, the word "chair," when it occurs, might be predicted by, and serve to predict, a set of features in the world that are to some degree systematic. Moreover, it appears that children can learn to identify the sounds of words from speech as useful signals in a way that would allow them to serve as regularities in the world that could enter into predictive relationships with other cues or outcomes (see Saffran, Aslin and Newport, 1996). Hence in encountering the word "chair," a child might learn both that the word "chair" serves as a cue to the presence of certain objects in the environment, and the cues that predict occurrences of the word "chair". From this perspective the goals of word learning are to identify (i) the set of environmental outcomes (objects or features) that the word "chair" predicts, and (ii) the set of features in the world that can serve as cues to the word "chair." (This conception of word-learning differs from "referential" views of word-learning in that while these theories stress the idea that words act as cues to things in the environment – such that hearing "chair" allows a listener to pick out some referent in the world – they do not generally consider the environment as a cue to words; see Bloom, 2000, for a review).

What implications does the view of words as cues that we have outlined have for our understanding of word learning? As we noted above, on the view that a set of objects in the world predicts a word, a surprise, or error-driven learning model such as the Rescorla-Wagner model would learn – through a process of cue competition – the set of cues that most reliably predicts that word (this might loosely be characterized as learning the "concept" associated with that word; see also Gluck and Bower, 1988). On the other hand, if a word *predicts* an object (which can consist, potentially, of a constellation of features, or dimensions), the situation can be very different. From this perspective, words appear to have a unique (or at least extremely uncommon) status relative to other cues in the world: while objects can have many features that can be simultaneously co-present (and hence, when the features of an object serve as cues to a word, cue competition will occur between these features with the result that the cue structure that serves to best predict the word will be learned under Rescorla-Wagner learning), words are single identifiable events in the world that do not tend to be bound to other specific co-occurring events.

As such, cue competition might not apply when a word serves as a cue to other events in the environment (there is little cue structure associated with a word as a cue; just the word itself). This difference may in turn lead to important differences between in the learning that takes place when using a word as a cue to a set of features, versus using a set of features as a cue to a word.

In the Rescorla-Wagner model differences in the associations learned from cues arises as a function of cue competition. Hence the logic of the model suggests that when a single cue (such as a word) predicts multiple outcomes (i.e. numerous features), each individual outcome could serve as a regularity relative to the learning of the predictive value of that cue. Therefore, with regular pairing and in the absence of other competing cues, the cue would simply increase its associative strength to the various features (or values on a dimension) until each association was learned to asymptote. Thus, ultimately, all the features could come to be predicted by the cue with similar associative strengths, such that the cue (word) would not come to discriminate between outcomes in any meaningful sense. The situation could be very different when there are multiple cues to a single environmental outcome, however. In this case the cues would compete with one another, meaning that the associations learned are likely to be more variable. In this case, discrimination between the various features is likely to be greater (see Ramscar & Yarlett, under review, for a formal treatment of this).

In English, the language most carefully investigated in the domain of children's color word learning, color adjectives occur prenominally ("blue cup") around 70% of the time in the speech adults direct at children (Thorpe & Fernald, 2006). When a color word is used prenominally as language unfolds in time, it acts as a cue to the feature to be learned (the color mapping) rather than being cued by that feature. Together with our analysis of learning theory, this suggests a reason for children's difficulties with color word learning in English: most of what they hear from adults will be unhelpful in learning what color words refer to. To explore this, in Experiment 1 we conducted a training study with children at an age where they are still struggling to master color adjectives (Rice, 1978; Sandhofer & Smith, 1999) and in Experiment 2, we sought to replicate our findings with children in adults learning novel color categories.

Experiment 1

Participants

Participants were 41 children aged between 23 and 29 (M=26) months recruited from the Stanford area. All were typically developing, monolingual, English learners.

Method and materials

Experiment 1 comprised 3 phases: a pre-test, training, and a post-test that was identical to the pre-test. Test materials comprised six objects that were novel to the children. There were three examples of each object in each of three colors (red, yellow and blue). The objects were presented on trays, with the target object location counterbalanced across trials. In both tests, children were asked to select the novel objects in response to requests in which the color word was either a prenominal ("which is the red one?") or a post-nominal ("which one is red?"). Question order was counterbalanced (children were never asked for the same color consecutively), and the form of the initial question was balanced between participants.

In training the children were introduced to a "magic bucket" containing 5 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 then presented with the items one-by-one and heard them labelled with color words used prenominally ("This is a red crayon") while the other half were introduced to the same items described with a postnominal color word ("This crayon is red"). Children then repeated the selection task on the novel items (the post-test was identical to the pre-test).

As we noted above, the problem associated with children's difficulty with color words is not these words are not unfamiliar, or that they can't use color words in appropriate parts of speech. Rather, it is that they fail to consistently map color words to correct colors (to this extent the color word usage of sighted children of this age appears similar to that of blind children, Landau & Gleitman, 1985). Thus to assess the quality of children's knowledge of the color words, and the effect of each type of training, correct choices on items that were consistent across the pre- and post-tests were used to measure children's color knowledge (a child was not considered to "know", or be a competent user of a color word if their pre- and post-test choice for a given color word was incorrect or inconsistent). Since the erratic nature of children's color word usage and comprehension is one of its more striking features, consistency of correct behavior constituted a relatively robust criterion for attributing color knowledge in 2-year-olds.

Results and discussion

Individual analysis of the pre- and post-test data (which confirmed parental vocabulary reports) showed the children had at least some knowledge of the three color words: they averaged 2 out of 3 correct choices in response to both pre- and postnominal question types, significantly more than chance both the pre- (M=63%, $t(81) = 7.77$, $p<0.0001$) and post- tests (M=57%, $t(81) = 5.57$, $p<0.0001$). To establish how robust this knowledge was, and to see whether robustness was affected by the different methods of training and testing we employed, a 2 (training type) x 2 (testing type) ANOVA was conducted of pre- and post-test consistency. This revealed a significant interaction between these factors ($F(1,39)=7.48$, $p<0.01$; see Figure 1), indicating that performance was at its most consistent when children were

both trained and tested on postnominal adjectives, and worst when trained on prenominal adjectives and tested on post-nominal adjectives (that the children performed poorest on postnominals when they were not trained on them is consistent with the low frequency of postnominal color adjectives in the input in English). Post hoc tests revealed significant effects of both training and question type: only the children trained with postnominally presented color words made more consistent correct sorts than one would expect by chance (M = 51%, $t(41) = 2.9$, $p<0.01$) compared to an average of 38% correct in the pre-nominal training group, $t(39) = 0.76$, $p>0.4$). Moreover, when children's responses to the question types were assessed independently, only children who had been trained with postnominal color word presentation and then tested with postnominal question types were significantly more accurate than chance (M = 59%, $t(20) = 3.2$, $p<0.005$; the p values in all other cells were greater than .3, see Figure 1).

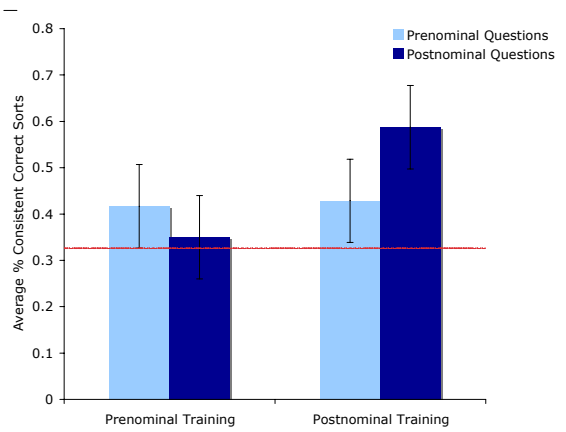

Figure 1: The average number of consistent correct sorts by the children in the training study (error bars are SEM).

Finally, comparing the pre- and post-test scores across each condition revealed a significant decline in performance when children were both pre-tested (M = 65% correct) and post-tested (M = 48%) with questions that placed the color words prenominally ($t(20) = 2.236$, $p<0.05$, two tailed). No significant overall changes were observed in the other cells.

Consistent with the analysis presented above, it appears that when children are exposed to color adjectives in post-nominal position, they learn them rapidly (after just 5 training trials per color); when they are presented with them pre-nominally, as English overwhelming tends to do, children they show no signs of learning.

This result is a striking testament to the predictions we derived from error-driven learning models above. To investigate whether adults might demonstrate similar learning patterns, in a laboratory study, we taught undergraduates novel color categories using a training regime that required them to learn four categories from very short stimulus exposures, either when the category name preceded the exemplar (to simulate pre-nominal learning) or when the exemplar preceded the category name (to simulate post-nominal learning).

Experiment 2

Participants

Participants were 16 Stanford undergraduates.

Method and materials

In order to challenge adult participants, who are already well aware of what hues standard color adjectives refer to, adults were asked to learn three categories that transcended ordinary color category boundaries mixing two different regions of color space within one category. One additional filler category was used that only occupied a single area of color space.

To further challenge adult participants, and to ensure that they could not use their pre-existing knowledge of color words in testing, training was arranged so that 75% of the exemplars they saw for each of the three mixed test categories were from one of the category's regions (high frequency), and only 25% were from the other (low frequency). Training was arranged so that the low frequency exemplars were more similar (closer in color space) either to the high frequency exemplars of another category, or the filler category, while still being clearly discriminable from the other categories (Figure 2).

Adult participants were trained and then tested on their ability to correctly classify the low frequency exemplars by name, and therefore reject competition from the high frequency exemplars from other categories. Training in the post-exemplar label condition consisted of seeing a 175ms flash of a color exemplar followed by a 1 second exposure to a written sentence naming the color, e.g., "That was wug." In the pre-exemplar label condition, the order of presentation of sentence and colored screen were reversed, and participants read e.g., "This is wug."

To control for any bias in the method of testing, and in order to recreate some of the pre- / post-test comparisons we performed in Experiment 1, half of the participants in each training condition were first tested by matching a single category name to four category exemplars (one from each category) and then given further training, after which they were tested on matching a single exemplar to the four category names; for the other half of the participants the order of the tests was reversed. Assessing performance on trials after initial and secondary training allowed us to assess whether there were performance gains or losses in each condition as training increased, as in Experiment 1. Each training block comprised 20 exemplars from each category, each repeated twice, and each test block comprised 26 test trials.

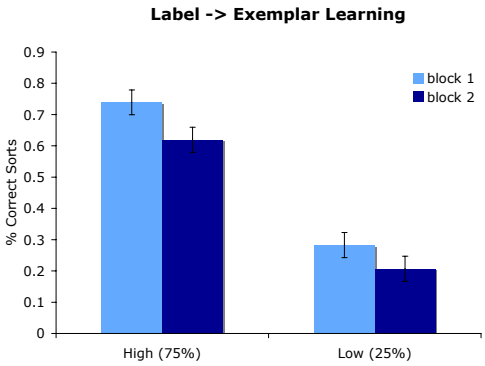

Figure 2: The three mixed-color categories and the filler category from Experiment 2. In the mixed categories, high exposure exemplars form the top row, and low exposure exemplars the bottom row

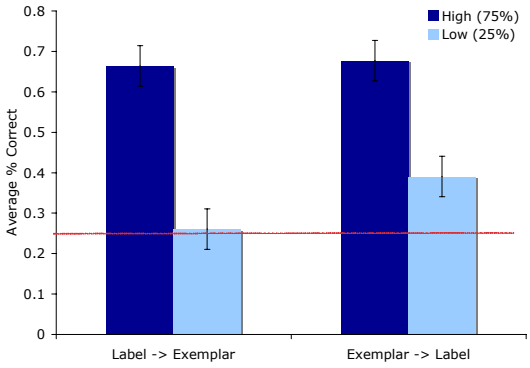

Figure 3: The average number of correct sorts by participants in Experiment 2 (error bars are SEM).

Results and discussion

Participants failed to learn to classify the low frequency exemplars when they were given a label prior to a color sample in training, effectively mimicking children's experience with pre-nominal color words, but succeeded in learning them from exactly the same number of exposures at the same duration when the color samples preceded the label, mimicking children's experience with post-nominal color words. A repeated measures ANOVA of high and low frequency performance with training type as a between subjects factor revealed a main effect of training condition, ($F(1,24)=46.00$, $p<0.001$), with no interaction between training and frequency. Critically, analysis of the low frequency categories revealed that participants in the condition where the category label followed the category exemplars (the "post-nominal" condition) sorted the low exposure exemplars correctly 39% of the time in a 4AFC task where chance was 25% ($t(23)=2.84$, $p<0.01$), whereas when trained with the label preceding the exemplars, participants sorted only 26% of the low exposure exemplars correctly ($t(23)=0.23$, $p>0.8$).

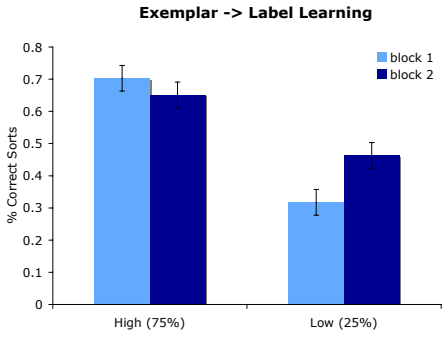

Figure 4: Test performance after Block 1 and Block 2 in the post-exemplar (top) and pre-exemplar (bottom) conditions.

Further, the failure of participants in the label first training condition did not appear to be something that further training would remedy. An analysis of the change in performance between tests one and two revealed that while (as predicted), participants in the exemplar first training condition showed a significant improvement in accuracy for low frequency items between trials one (32%) and two (46%, $t(24)=1.8$, $p<0.05$; there was no significant change for the high frequency items), participants in the label first condition showed a decrease in performance between trials one (31%) and two (21%, $t(24)=2.05$, $p<0.05$, two-tailed). This decline (which was consistent with the decline in performance in the pre-nominal test/train data in Experiment 1) was mirrored in the high frequency items as well (trial one 72%, trial two 60%, $t(24)=2.2$, $p<0.05$, two-tailed; Figure 4).

General Discussion

In two Experiments testing the learning capacities of both young children and adults, we have demonstrated that the difficult (in English at least) task of mapping a color word to a specific color (or space of colors) in the world can be massively affected by the order in which words and their referents are arranged. The importance of sequencing and timing is well established in animal learning: in animal models (or any model that implements cue competition), what gets learned depends on the timing of one event relative to another. These factors may apply to learning words and meanings more than has often been supposed.

Acknowledgements

We thank Nicole Gitcho, Brad Love and Dan Yarlett for discussion. This research was supported by a National Science Foundation CAREER Award (#0547775) to MR

References

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

Bornstein, M. H., W. Kessen, & S. Weiskopf. (1976). Color Vision and Hue Categorization in Young Human Infants, *Journal of Experimental Psychology* 2:115-19.

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

Gallistel, C. R., & Gibbon, J. (2000). Time, rate, and conditioning. *Psychological Review, 107,* 289–344

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

Hollerman, R. and Schultz, W. (1998) Dopamine neurons report an error in the temporal prediction of reward during learning, *Nature Neuroscience* 1(4): 304-309

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

Ramscar, M & Yarlett, D (under review), The temporal and predictive structure of symbolic category learning.

Rescorla, R.A. 1988. Pavlovian conditioning: It's not what you think. *American Psychologist* 43:151-160.

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

Rice, M. L. (1978). The effect of children's prior nonverbal color concepts on the learning of color words. (PhD dissertation, University of Kansas, 1978) *Dissertation Abstracts International, 39* (8-A), 4915.

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

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

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.