Breaking the World into Symbols

Adam D. November (adam.november@gmail.com)
Stanford University - Palo Alto, CA 94305 USA

Nicolas Davidenko (ndaviden@stanford.edu)
Stanford University - Palo Alto, CA 94305 USA

Michael Ramscar (ramscar@gmail.com)
Stanford University - Palo Alto, CA 94305 USA

Abstract

How do people come to assign symbolic labels to continuous dimensions? Previous work has shown that prediction-error-driven models are sensitive to the order of labels and exemplars during training; similar patterns of learning are found found in adult learners trained to associate labels with discrete visual stimuli. Here we provide further evidence in support of the hypothesis that an error-driven mechanism underlies word learning, using continuous stimuli to explore the interactions of temporal structure, stimulus frequency, and distinctiveness in shaping associative learning. We conclude that learning to use features of exemplars to predict labels results in over-representation of diagnostic information, as shown by improved associative performance on stimuli near category boundaries. This is consistent with an error-driven model of label acquisition, and highlights the importance of the associative and prediction-based (rather than exclusively syntactic) aspects of symbolic cognition.

Keywords: Symbolic Cognition; Categorization; Language; Learning; Representation; Concepts; Computational modeling; Prediction

Introduction

The world is full of perceptually similar stimuli that prompt behaviorally diverse responses. Picking ripe fruit, avoiding poisonous creatures, and even interpreting subtle facial expressions all rely upon careful discriminations of subtle cues. What role does language play in learning such discriminations? How can the process of learning a symbol change what is seen?

Classic referential theories of language have little to say on the content of individual words, instead focusing on the syntactical rules for combining abstract symbols. Such theories largely ignore the process of learning words; for example, Fodor (1981) has argued that the concept of carburetor must be innate, restricting the problem of word learning to the problem of finding the proper mapping between new words and the appropriate innately elaborated concepts. Many have discussed this and other limitations of such referential theories of language (Tomasello, 2003; Wittgenstein, 1953). Recent work has rejected the referential view of language, and has instead argued the importance of prediction for developing symbolic representations (Ramscar, Yarlett, Dye, Denny, & Thorpe, 2010). Specifically, this analysis applies error-driven learning to the problem of determining which cues in the world to associate with a word, and shows how varying the structure of information in time elicits different patterns of association.

Within a predictive framework, words can either be used as predictors or as predictions. Used as predictors, the words ‘dog’ or ‘wolf’ often predict very similar outcomes: canine creatures which share many features, and thus error driven-systems will learn quite similar representations for each over time. In contrast, systems using the appearance of a *dog* or a *wolf* to predict the relevant label will have highly distinct outcomes (either the label ‘dog’ or the label ‘wolf’). Over time, these highly distinct outcomes can serve as a useful prediction targets, allowing an error-driven learner to isolate the diagnostic features (e.g. ‘ear floppiness’) that lead to correct prediction.

Recently, Ramscar et al. (2010) demonstrated effects of information structure on symbolic learning by training subjects to learn to name novel stimuli while manipulating temporal order. We argue that this sensitivity to information structure is the result of a fundamental aspect of symbols: informationally impoverished, symbols must serve as abstractions of the things they represent. When learning the symbol’s many associations with the world, it can be advantageous to learn these symbols as consequents predicted from cues in the world, rather than as cues used to predict the world. Such an effect was found for for participants learning to label various fribble categories (Tarr, 2000). In the case of fribbles, the relevant cues consist of a number of discrete features. When trained to predict from **F**eatures to **L**abels (**FL**), subjects were better able to learn the relevant categories, than when trained from **L**abels to **F**eatures (**LF**). Modeling suggested that FL-trained subjects improved by learning to use the diagnostic features to inhibit potentially conflicting labels. In contrast, LF learning developed a more ‘veridical’ but less discriminative model that represents the relative feature frequencies for each label. This Feature-Label-Order (FLO) effect suggested that learning natural symbols is sensitive to information structure. Here, we attempt to explore if such FLO effects can be found while learning to discriminate the boundaries of continuous dimensions.

Is learning a mapping between continuous exemplars and discrete labels also affected by the information structure available at learning? Demonstrating a FLO effect in the realm of continuous stimuli would reinforce the idea that prediction is fundamental to symbolic learning and category learning. However, it is possible that a continuous dimension would not provide the consistent ‘hooks’ for discriminative learning that are present when learning binary features (such as fribble appendages); without discrete features from which to generate predictions, the FL advantage could be nullified. However, based upon cue-competition, we predict that even noisy continuous features can serve as useful cues, as incremental learning can distill the consistent cues across trials. In addition, using continuous cues may allow us to learn more about the FL advantage in category learning: is the advantage a broad advantage, improving performance across the board, or will it only boost performance near the category boundary? By looking at how information structure in the world interacts with distinctiveness and frequency, we can learn more about the role of prediction in the learning of natural symbols.

Materials: Novel Shapes

Generation We created several families of novel objects to be used in novel mappings. Our novel objects were generated from a continuous parameter space by drawing smooth splines between points on the 2D plane. This technique, previously used by Davidenko (2007) to probe the mechanisms of face representation, allows one to parameterize a physical space, interpolating continuously between points. Figure 1 demonstrates the process used to generate the novel shapes. First, 16 key-points were randomly selected (A), similarly to the previous face silhouette methodology. The points were then connected to form a closed shape (B) and beta-splines smoothed the edges (C). Shapes were then filled to create a two-tone image (D). Shape 'families' were created by 'morphing' pairs of shapes together via linear combinations of the key-point positions. Families were selected by visual inspection, eliminating any unusually distinctive features such as loops. Pilot studies confirmed that within-family pairs were more similar than between-family pairs.

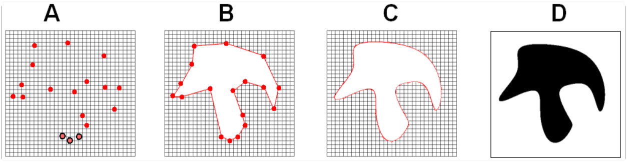

Figure 1: Individual stimulus generation.

Calibration In order to ensure that subsequent tests were equally difficult, we attempted to standardize discriminability of each of the families. We tested 20 participants (mean age of 19.8 years), showing them a series of shapes at 6 levels of difference from 6 families. Each trial consisted of a pair of temporally separated and masked objects; subjects had to respond "identical" or "different". One-third of trials were 'same' trials, where the identical shape was shown twice. The remainder were 'different' trials, adding up to a total of 360 trials per subject. Based upon these data, we constructed estimates of d-prime as a function of parameter distance, and estimated the level of distance at which discrimination performance (d-prime) would be approximately 1.5 via the fitted regression parameters.

Finally, based upon that calibration, we created a new set of stimuli whose half-max distance was estimated to have a same-different discriminability of about 1.5 d-prime, as laid out at the bottom of Figure 3. This procedure provided us with a reasonable assurance that discriminability was comparable across families, and that any remaining uneven discriminability could be controlled via counter-balancing.

Experiment 1: FLO Effects with Continuous Stimuli

Will the temporal order of features and labels influence learning of a continuous dimension as predicted by our error-driven account of learning? Adapting the design from Ramscar et al. (2010), we undertook an experiment to demonstrate the influence of information structure on symbolic learning of a continuous dimension.

Subjects were situated in a quiet room in front of a keyboard and monitor, and were instructed to pay attention to the series of events which occurred on the screen, and that they would be tested on what they had learned. During the training section, temporally staggered pairs of items were presented, with longer blanks between pairs. The primary manipulation was the order of items during study, either the **F**eatures (objects) preceding **L**abels (**FL**), or vice-versa (**LF**). After training, subjects were tested by presenting one shape at a time, and asking subjects to choose the associated label. (We did not test the converse (selecting a shape when prompted with a label) because previous experiments (Ramscar et al., 2010) did not find effects of congruency between training and testing direction.) In Experiment 1, study order was manipulated across subjects, such that an individual subject was trained either entirely in FL or LF order.

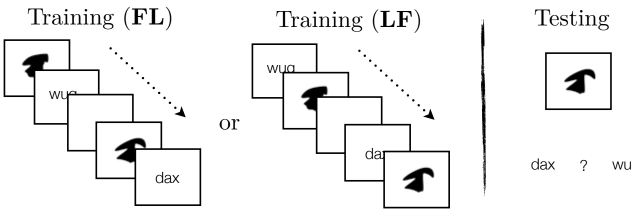

Figure 2: General experimental structure. During training, continuous-featured objects either preceded labels (FL), or were subsequent to labels (LF). At test, subjects made a multiple-alternative forced choice.

Methods

Two morph families were selected for this experiment, with 10 exemplars generated at regular intervals from each respective parameter space. The two families were then split into two categories by first splitting each family at the midpoint, with one half being assigned to the 'rare' frequency condition and the other to the 'common' frequency condition.

Then, the ‘common’ and ‘rare’ exemplars of the two different families were crossed, resulting in two categories comprising half the exemplars of each family (see Figure 3). Across subjects, we randomized the assignment of frequency conditions to stimuli, thus randomizing pairings of individual stimuli within categories. Category labels were also randomized. The disjunctive category structure was used to pit cue frequency against cue diagnosticity, while equating overall label frequency (as in Exp. 1 of Ramscar et al., 2010).

Sixty-eight undergraduates (mean age of 19.9 years) at Stanford University participated for course credit or payment. We instructed subjects to attend to a serial stream of temporally staggered items. During training, subjects were randomly assigned to one of two training conditions: either features were followed by labels (FL) or labels were followed by features (LF). Each study trial progressed as follows: blank for 800 ms; fixation for 200 ms; first stimulus for 300 ms (features or labels depending on condition); fixation for 300 ms; the appropriate paired stimulus (label or features) for 300 ms; and a final fixation for another 200 ms (Figure 2). Subjects were exposed to rare exemplars 4 times each (25% of trials), and common exemplars 12 times each (75%), for a total of 160 trials. Subjects had no other task than to attend to the serial stream of items. There was one break halfway through training.

Subjects were then tested on their learning of the pairings of exemplars and labels. Subjects were presented with a single exemplar, and the two category names, corresponding to the left and right arrow keys on a keyboard. Subjects were told to respond as accurately as possible, under no time pressure. Each exemplar was tested 5 times, for a total of 100 trials.

Results

Subjects learned the categories relatively well, achieving 76.1% accuracy overall. In Figure 4 we see accuracy of activation plotted against morph index, grouped by study order and frequency (collapsing the data across item identity, which is counterbalanced). The curves suggest that subjects are better at classifying the most distinctive stimuli and that FL training results in better overall performance. In addition, it appears that there may be interactions between study order, discriminability, and frequency. In order to test the significance of these effects, we have binned the data by distinctiveness: the stimuli 3 or more units from the category boundary are now ‘distinctive’, while those 2 or less are ‘confusable’. This simplification still captures the general patterns of the data, but recodes it as a binary factorial design: training order (LF or FL), distinctiveness (confusable or distinct), and frequency (rare or common). We then analyzed this binned data using multilevel logistic regression using the lme4 package (Bates & Maechler, 2009), in R (www.r-project.org), with distinctiveness, study order, and frequency as fixed effects, and subjects as random effects. Fixed effects are reported in terms of Log-Odds Ratio (*LOR*), i.e. a *LOR* of 1.0 means that trials in the given treatment condition were $e^{1.0} = 2.7$ times more likely to be correct, and subsequently tested for significance via the Wald z statistic, the ratio of the effect size to its standard error (Jaeger, 2008).

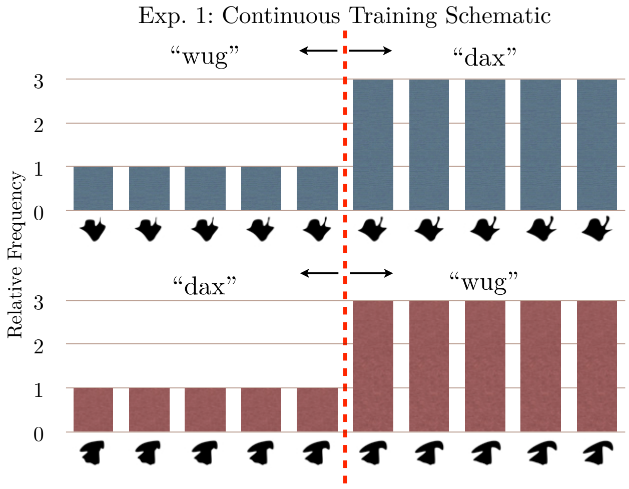

Figure 3: Two stimulus dimensions of 10 exemplars each are split into two disjunctive categories, “wug” and “dax”. In this example, the left half of each dimension is rare, and the other half is common, with exemplars occurring 3 times as often as rare exemplars. Label frequency is equalized, preventing any overall label bias. Alignment of categories and frequencies were counterbalanced across subjects.

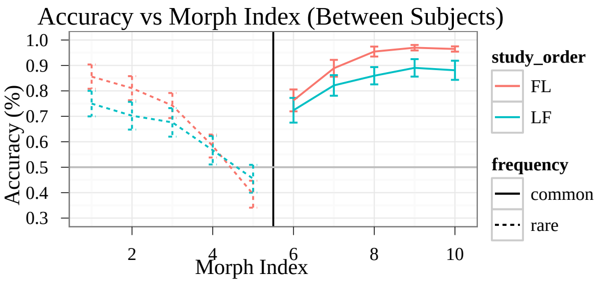

Figure 4: Exp. 1, Demonstration of FLO effects over continuous stimuli. Data has been collapsed across items, leaving accuracy as a function of the stimulus dimension, study order, and frequency. Notice the overall advantage for the FL subjects, and the substantial dip in performance on the rare side of the category boundary.

As suggested in the un-binned plot (Figure 4), a significant interaction between study order and distinctiveness indicates that FL-trained subjects were much better away from the category boundary ($LOR = 1.01$, Wald $z = 4.4$, $p < 10^{-6}$). In addition, a significant interaction between training order and frequency suggests that LF-trained subjects were especially impaired on rare items ($LOR = -0.43$, Wald $z = -2.5$, $p < .02$), across distinctiveness levels. This finding closely matches the original FLO effect experiment (Ramscar et al., 2010), echoing the difficulty of learning rare LF-trained items. In addition, there are highly significant main effects of frequency (a high-frequency performance bias) and distinctiveness (better performance for distinctive items). The main effect of training order is masked by the various interaction effects.

These results are generally consistent with prediction-based constraints on symbolic learning. FL trained subjects seem to be learning the associations better in general, but may be slightly overweighing diagnostic information near the category boundary. So far, this data has provided evidence that learning to map continuous stimuli to discrete labels is subject to the same informational constraints that are present when learning more discrete stimuli, and a replication of the original FLO effect with novel, continuous stimuli. Subjects trained FL had a general advantage, learning the classification more effectively, along with what appears to be a slight diagnosticity bias. It also corroborates our understanding of the mechanism, suggesting that cue-competition from error-based learning can influence the acquisition of associations between labels and exemplars.

However, the between-subjects design of this experiment is open to criticism; it may be that the LF subjects quickly become less motivated due to the relative difficulty of their task. When predicting a diverse set of items from a discrete labels, it is impossible to perfectly predict the exact item that arrives on any trial. This may be frustrating, and cause subjects to pay less attention. Indeed, the training-order effect is largest at the most discriminable stimuli, suggesting the worst subjects may simply not be paying sufficient attention. In addition, the presence of only two labels makes interpreting errors difficult; we can not check our assumption that the two stimulus dimensions have been learned independently. Errors could be due to confusion either with items from the same dimension but with the opposite label, or from confusion with elements from the other dimension. To avoid these and similar shortcomings and criticisms, Experiments 2 and 3 will replicate the continuous FLO effects in a within-subjects design, mitigating the influence of global attentional effects and increasing the interpretability of our results.

Experiment 2: Within-subjects training

With the between-subjects design of Experiment 1, we saw that training condition showed an overall effect on performance. Although this effect is consistent with our error-driven model, it is possible that the difference could be driven by a global attentional effect, rather than the specific learning mechanism used. Our experiments so far have used disjunctive categories to explore the differences in learning, pitting salient and frequent but non-diagnostic information against less-salient but diagnostic information. The disjunctive categories may help explain the confusion even on relatively distinctive exemplars. In addition, we hope to equalize performance on the most distinctive exemplars across subjects, because any performance difference at the endpoints makes it much more difficult to interpret any changes in bias or sensitivity at the category boundary.

In order to improve both the power of the analysis and the strength of our argument, we moved to a within-subjects design (Figure 5). Subjects were trained to categorize two continuous families: one trained FL, one trained LF. Each family was split down the middle, and each half was associated with a novel label, yielding a total of four labels. To ensure we detect any differences near the category boundary between training types, we increased training to ensure all subjects learned the most distinctive items well. To reach this goal, we gave each subject three training and testing sessions, with the same items repeated (in a random order) during each of these three blocks. We did not analyze the progression of learning across blocks, instead analyzing them as one large block; the patterns of effects looked consistent across blocks.

In this experiment, we temporarily removed the frequency manipulation to obtain a minimalist, within-subjects test of FLO effects. With these changes, the critical questions became: will the FL advantage in learning replicate in a within-subjects design? After eliminating disjunctive categories? While training subjects across multiple blocks? The cue-competition account predicts that the FLO effect should be independent of these factors; by controlling their influence we further constrain the space of possible alternative explanations.

Methods

Eight undergraduates (mean age of 18.9 years) at Stanford University participated for course credit or payment. The within-subjects manipulation of training-type, fewer counterbalancing conditions, and extended training greatly increased our power, substantially reducing the number of subjects needed to demonstrate an effect of training.

The general structure of training was similar to Experiment 1, with the following differences: Each of three training blocks presented both distributions 9 full times each, for a total of 180 trials total (5 exemplars/label * 4 labels * 9 repetitions). For each subject, one family was learned in FL order, while the other was learned LF, counterbalanced across subjects. Individual trials proceeded as in Experiment 1.

At test, subjects were tested on each exemplar 3 times, for a total of 60 trials per testing block. Testing took the form of an unspeeded 4-AFC, with one exemplar presented as a cue, and the four labels presented as alternatives on the screen. Location of the labels on the screen was randomized for each subject, but consistent throughout the experiment.

Results

Experiment 2, with the increased power of a within subjects design, yielded a similar pattern of data (Figure 6), where FL training once again had a positive effect on performance. Beyond the strong effect of distinctiveness, a mixed-effect logistic regression of the data (binned as in Experiment 1) reveals a main effect of study order: FL training results in better associative learning ($LOR = 0.37$, Wald $z = 2.07$, $p < .05$). This minimal manipulation supports the idea that learning to categorize continuous stimuli is constrained by FLO effects. However, the lack of an interaction between distinctiveness and training order is problematic for our preferred interpretation: general attentional suppression of the FL condition could generate this pattern of results. In the final experiment, we attempt to address this possibility.

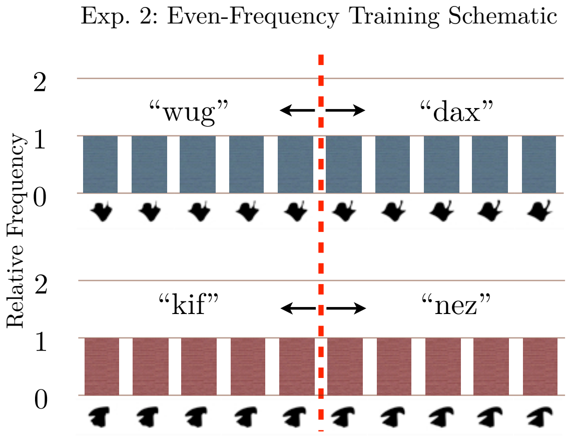

Figure 5: Two stimulus dimensions are split among four labels, with no frequency bias. Training order (FL vs LF) is manipulated within subjects, between dimensions.

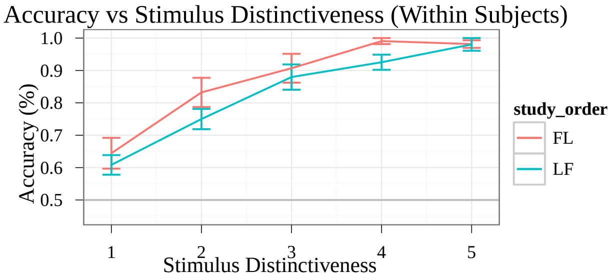

Figure 6: Exp. 2, A within-subjects manipulation reveals better performance on items trained FL rather than LF.

Experiment 3: Frequency Distributions

Having successfully demonstrated a within-subject continuous subject FLO effect, we hope to gain insight into the representations that are created by FL vs LF learning by adding a frequency manipulation (see Figure 5). Experiment 2 demonstrated a within-subjects continuous FLO effect. However, this main effect of training order is consistent with a couple of possible hypotheses. While the FLO analysis explains this effect as a result of cue competition stemming from prediction error in the FL condition, another hypothesis is that LF training generally reduced performance for some unrelated reason; perhaps words appearing first are less salient than shapes. Performance could still be reasonably high in both conditions; trials at the highest discriminability could reach ceiling, but the difference across conditions could be simply due to differences in attention. In order to rule this confounding hypothesis out, we need to demonstrate that the representations of the associations have been shaped by the process of learning. As in the original fribble experiment, differences in the pattern of bias due to frequency differences can shed light on the representations being used.

Methods

Unlike Experiment 1, the frequency manipulation in this experiment did not bias entire halves of each family. Instead, a sawtooth-shaped frequency distribution was adopted (as shown in Figure 7). Thus each stimulus family was split into two halves; on the 'near' side, the most frequent exemplar is the one against the border, while on the 'far' side, the most common exemplar is the most distinctive, farthest from the category boundary. Like in Experiment 1, the overall frequency of each label was thereby equalized, but in addition, the use of overly confusing disjunctive categories was avoided.

Eighteen undergraduates (mean age of 19.4 years) at Stanford University participated for course credit or payment. Training proceeded as in Experiment 2, but with altered frequency profiles. Each training block trained each distribution 6 full times each, for a total of 180 trials total (15 presentations/label * 4 labels * 6 repetitions). The 'near' side for each stimulus dimension was counter-balanced across subjects. For each subject, one family was learned in FL order, while the other was learned LF, counterbalanced across subjects. Testing proceeded identically to Experiment 2.

Results

To analyze the data, we collapsed the data across stimuli within conditions. In Figure 8, we see a similar overall pattern of results as Experiment 1. Once again, general learning performance is high; at the ends, performance is much closer to perfect; LF may even be overtaking FL at the most extreme morph values. Elements near the category boundary are disproportionately affected by training order. To quantify these effects, we again ran a mixed effects logistic regression on the data (again binned by distinctiveness); this revealed a three-way interaction between training order, distinctiveness, and frequency conditions ($LOR = 1.05$, Wald $z = 2.08$, $p < 0.05$). This effect is driven by the disproportionate accuracy of subjects on the high-frequency FL trained stimuli near the category boundary, suggesting that FL trained items exploit diagnostic features near the category boundary, but that these critical features are less informative for the most discriminable exemplars. The presence of this interaction provides the strongest evidence yet that the temporal information structure influences the development of representations, as predicted by our error-driven model of learning.

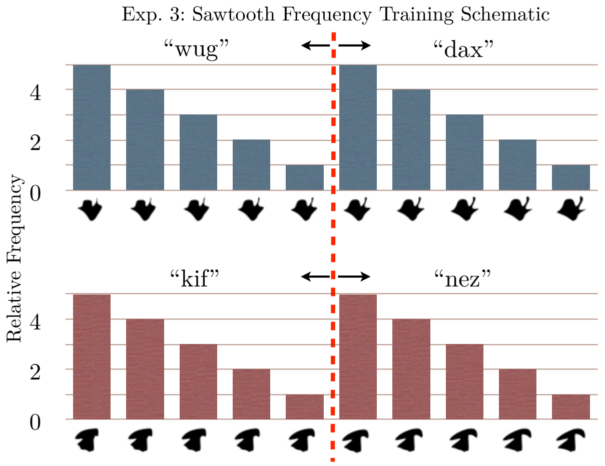

Figure 7: Two stimulus dimensions are split across four labels. In this case, the elements labeled “dax” and “nez” are in the ‘near’ frequency condition, while the elements labels “wug” and “kif” are in the ‘far’ frequency condition. This design will help demonstrate the interaction of training with frequency information across the distinctiveness dimension.

Discussion

Error-driven learning has been extensively studied across a variety of disciplines such as classical-conditioning (Pavlov & Anrep, 1927) and decision making (Gluck & Bower, 1988); successful computational models have been developed which accurately predict a wide range of experimental findings (Barlow, 2001; Rescorla, 1988). Here, we have extended this powerful model to help understand how symbols are learned.

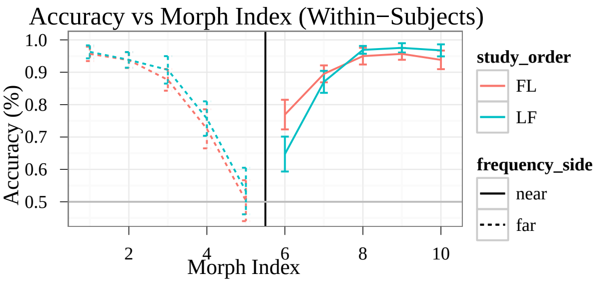

Figure 8: Exp. 3, Accuracy by training order and frequency. A significant three-way interaction on the mixed logistic regression suggests a robust FLO effect. Error bars are simple between-subjects SEM. Analysis of deviance of binned data results in a significant three-way interaction, driven by the improved performance near the boundary of the FL trained items in the ‘near’ frequency condition.

We found significant advantages for training subjects to use exemplars to predict words, even when performance on the most distinctive examples was equalized in within-subject designs. Furthermore, the interaction of training-order, distinctiveness, and frequency in Experiment 3 provides direct support for our cue-competition account, confirming the intriguing prediction that FL learning tends to distort representations, making them less ‘veridical’, but more useful compared to the ‘naive’ models developed by LF learning. More generally, these experiments demonstrate that the temporal structure of information present during learning can shape representation, suggesting that linguistic representations must be learned. This work also has more general implications for categorical learning, suggesting that the simple association is not sufficient for explaining patterns of learning; competition amongst predictive elements will shape the pattern of association, a conclusion compatible with previous work on classification (Sakamoto & Love, 2010).

References

Barlow, H. (2001). The exploitation of regularities in the environment by the brain. *Behav Brain Science*, *24*(4), 602-607.

Bates, D., & Maechler, M. (2009). lme4: Linear mixed-effects models [Computer software manual].

Davidenko, N. (2007). Silhouetted face profiles: A new methodology for face perception research. *Journal of Vision*, *7*(4), 6.

Fodor, J. (1981). *Representations: Philosophical essays on the foundations of cognitive science*. MIT Press.

Gluck, M., & Bower, G. H. (1988). From conditioning to category learning: an adaptive network model. *J. Exp. Psychol Gen.*, *117*(3), 227-47.

Jaeger, T. F. (2008). Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434-446.

Pavlov, I., & Anrep, G. (1927). *Conditioned reflexes*. Oxford University Press.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The Effects of Feature-Label-Order and Their Implications for Symbolic Learning. *Cognitive Science*, *34*(6).

Rescorla, R. (1988). Pavlovian Conditioning. *American Psychologist*, *43*, 151–160.

Sakamoto, Y., & Love, B. (2010). Learning and retention through predictive inference and classification. *Journal of Experimental Psychology: Applied*, *16*(4), 361.

Tarr, M. J. (2000). *Fribbles*. Retrieved 1/5/2010, from `http://www.tarrlab.org/`

Tomasello, M. (2003). *Constructing a language*. Harvard University Press.

Wittgenstein, L. (1953). Philosophical investigations.