Grounded Acquisition of Color Terms by a Computational Model

Deanna DeCarlo*, William Palmer*, Lasse Heinrich van den Berg*, Xiaomeng Zhu*, Maryam Elbenni, Sachien Fernando, Daniel Koldobskiy, Annie Ye, R. Thomas McCoy, Robert Frank
Yale University
{deanna.decarlo, w.palmer, lasseheinrich.vandenberg, miranda.zhu}@yale.edu

Background Compared to the fast and early acquisition of object names observed in children [1], the acquisition of color terms presents an interesting puzzle: they are acquired later and with considerable difficulty (e.g., [2]). A range of proposals have been made about conceptual [3], attentional [2], and linguistic constraints [4] that play a role in this learning process. In this paper, we probe the necessity of such constraints by considering the acquisition of color terms by a learner free of such domain-specific biases. Specifically, we use Vong et al.'s multimodal CVCL language model [5], a neural network that is trained to align utterances with visual input from a developmentally-plausible corpus and is successful at identifying a wide range of object names, despite variability across categories [6]. We ask whether the same training regime that allows for noun learning will also apply to color terms. We find variability in the model's mastery of different color terms, and this motivates us to explore the factors that modulate success and their alignment with human learning.

Evaluating Color Term Acquisition The SAYCam dataset [6] consists of videos from a camera mounted on a child's head, and therefore represents a rough approximation of a child's contextually-enriched linguistic experience. To use this dataset to evaluate color term learning, we focus on utterances in SayCAM that contain some (basic) color term and their corresponding images (extracted from the video time-aligned with the utterance). Given an utterance containing color term T_i, we further restrict attention to only those corresponding images that contain an object of corresponding color C_i as determined by a majority of human judges. Using the resulting set of utterance-image pairs, we construct an evaluation task: an image I that has been annotated as containing a certain target color C_i is paired with 9 basic color terms, and the task is to label I as T_i. For the CVCL model, this is done by identifying the color term whose word embedding has the highest cosine similarity with the I's vector representation, which follows from [5]. Results are shown in Figure 1. When evaluated on utterance-image pairs from the original training set of CVCL, CVCL performs above chance for all colors, with purple achieving the highest accuracy of 0.615 and blue achieving the lowest accuracy of 0.205.

Linguistic Experience and Learning Success Our results show considerable variability in model accuracy across color terms. Why is this? One potential modulating factor is frequency: color terms that occur more frequently during training might be better learned. A second factor concerns variation in world-word fit, as represented by the conditional probability of a color C_i appearing in the visual context given that T_i is uttered. Higher conditional probabilities could, in principle, facilitate learning. We estimate frequency and conditional probability from our labeled corpus data, and the results are given in Table 1. To test the import of these factors, we use logistic regression to predict accuracy on a given color-image pair, using these frequencies and conditional probability as predictors. The results, shown in Table 2, indicate that conditional probability has a significant positive effect on accuracy, as expected, but color term frequency has a significant negative effect. A small positive coefficient for the interaction term indicates that highly frequent terms get some compensation if they have high conditional probability.

Conclusion The results presented above indicate that the correspondence between world and word, as measured by the conditional probability of a color given its corresponding color term, is an important factor in the acquisition of color terms in CVCL. Naturally, one expects performance to improve by finetuning on examples with color terms where this correspondence with the image occurs. Preliminary results show that this type of fine-tuning is especially helpful. Psycholinguistic work on the acquisition of color terms has implicitly assumed the importance of world-word fit [7], something which the results here indicate is crucial for computational models of the word learning process as well.

*Equal contributions.

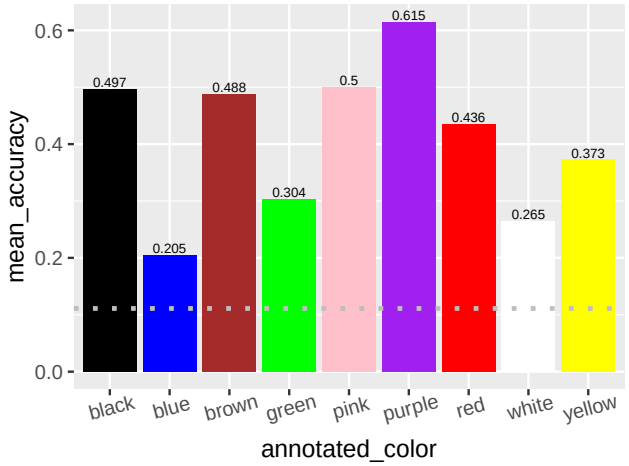

Figure 1: Accuracy on color terms. The dotted line shows chance performance.

Term	Freq	P(Color|Term)
black	702	32.9
blue	1895	33.2
brown	935	47.8
green	1677	49.5
pink	80	58.8
purple	1073	39.6
red	2111	50.7
white	667	34.2
yellow	1440	43.7

Table 1: Frequency of color terms and conditional probabilities of the corresponding colors in the training set.

Term	Coefficient	P <
Intercept	-0.48	0.001
Cond. Prob.	0.25	0.001
Frequency	-0.25	0.001
Cond. Prob. : Freq	0.06	0.05

Table 2: Results of the logistic regression with scaled ($\mu = 0, \sigma = 1$) frequency and conditional probability as predictors.

References

[1] Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, & G. A. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 264–293). MIT Press.

[2] Soja, N. N. (1994). Young children's concept of color and its relation to the acquisition of color words. *Child Development*, *65*(3), 918–937.

[3] Kowalski, K., & Zimiles, H. (2006). The relation between children's conceptual functioning with color and color term acquisition. *Journal of Experimental Child Psychology*, *94*(4), 301–321.

[4] O'Hanlon, C. G., & Roberson, D. (2006). Learning in context: Linguistic and attentional constraints on children's color term learning. *Journal of Experimental Child Psychology*, *94*(4), 275–300.

[5] Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science*, *383*(6682), 504–511.

[6] Sullivan, J., Mei, M., Perfors, A., Wojcik, E., & Frank, M. C. (2021). Saycam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open Mind*, *5*, 20–29. https://doi.org/10.1162/opmi_a_00039

[7] Sandhofer, C. M., & Smith, L. B. (1999). Learning color words involves learning a system of mappings. *Developmental Psychology*, *35*(3), 668–679.