Coloring Outside the Lines: Error Patterns in Children's Acquisition of Color Terms

Julia Watson
Department of Computer Science
University of Toronto
(jwatson@cs.toronto.edu)

Barend Beekhuizen
Department of Language Studies
University of Toronto
(barend.beekhuizen@utoronto.ca)

Suzanne Stevenson
Department of Computer Science
University of Toronto
(suzanne@cs.toronto.edu)

Abstract

A key challenge for children in language acquisition is to learn the mapping of words to mental categories, since this mapping varies greatly from language to language. The errors children make in this process are very informative regarding the development of lexical semantic categories; in particular, how children overextend a word to an inappropriate exemplar provides a window onto the mechanisms that underlie their categorization processes. We perform a large-scale quantitative analysis of the detailed patterns of children's errors in the domain of color, finding evidence that these error patterns are driven by an interaction between domain general principles of categorization, and children's developing knowledge of the semantics of color. Our results suggest that, while domain general processes play a role throughout development, their influence varies across ages according to their use of domain specific (conceptual) knowledge, which gradually increases over time.

Keywords: word learning errors; semantics; color terms

Introduction

A key aspect of language acquisition is learning to map words to mental categories. A child learning English must figure out that *blue* refers to an area of color that Russian-speaking children learn to divide into *sinij* ('dark blue') and *goluboj* ('light blue') (Davies et al., 1998), while Setswana-speaking children learn a broader term *botala* that covers *blue* and *green* (Davies et al., 1994). Because children must determine how their language precisely carves up the semantic space of a domain into the appropriate lexical meanings, acquisition of word–meaning mappings reveals much about children's representation and learning of conceptual categories (e.g., Davies et al., 1998, among many others).

The errors children make in applying words to situations are particularly informative about their developing categorization processes (e.g., Clark, 1973; Pitchford & Mullen, 2003; Gentner & Bowerman, 2009). Young children often overextend words to inappropriate exemplars, such as using *blue* to refer to a color that adults would call *purple* (e.g., Bateman, 1915). If children consistently make such an error but rarely generalize the word *blue* to a RED stimulus, that is revealing about how they make decisions about category membership. In particular, such error patterns may tell us about both the domain general principles children use in their learning (e.g., use of similarity of stimuli to assess co-categorization) and their domain specific (conceptual) knowledge (actually knowing, within a particular domain, what determines similarity of stimuli).

Color term learning is an apt testbed for such research because color terms begin to be used fairly early, but are mastered (at the adult level) relatively late (e.g., Bornstein, 1985). Thus we have the opportunity to explore, over multiple years of development, how children's color categories are (often incorrectly) comprised. Importantly, because even young children appear to have adult-like knowledge of the perceptual space of color (e.g., Pitchford & Mullen, 2003), research can focus on how children's developing conceptual organization of color influences the learning of lexical semantic categories. This leads to the possibility of observing a developmental trajectory of errors, which may reveal how domain general principles of categorization interact with the development of domain specific (conceptual) knowledge.[1]

Our research builds on a number of studies that take differing views on the role of these two factors in the time-course of color term acquisition. One view focuses on the acquisition of domain specific knowledge: it is suggested that younger children lack an understanding of the dimensions of color relevant to color word learning (e.g., Bornstein, 1985; Pitchford & Mullen, 2003); this knowledge then develops quickly (within 3 months) around the third birthday, when most color words are learned (Pitchford & Mullen, 2002). A different view focuses instead on the role of domain general principles of categorization (such as assessing frequency or proximity of exemplars in forming categories); these principles apply throughout the period of color acquisition, even at very young ages, and lead to gradual refinement of color categories (Wagner et al., 2013; Yurovsky et al., 2015). The assumption here is that the color domain knowledge is in place throughout development, but the categorization process is what takes time.

We crucially observe that, while domain general principles of categorization may apply throughout development, successful application of those principles depends on varying degrees of domain knowledge. For instance, across many domains, both frequency and salience of a category's exemplars are properties that influence category learning (e.g., Nosofsky, 1986). But while the frequency of a lexical semantic category can be computed based purely on the child's linguistic input (e.g., how often does she hears the word *blue*), determining a category's salience requires knowledge of the struc-

[1] In the remainder of the paper, we use the terms "domain [specific] knowledge" and "semantics" to refer to the conceptual layer of organization.

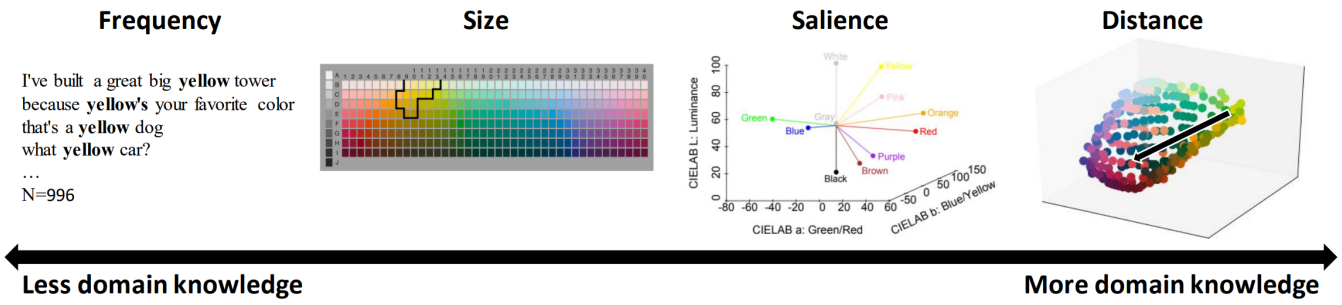

Figure 1: Spectrum of features, ranging from low to high domain specific (conceptual) knowledge. Illustrations focus on *yellow*, with distance presenting the distance between *yellow* and RED. The salience image is reproduced from Yurovsky et al. (2015).

ture of the conceptual domain (e.g., how does BLUE reflect key dimensions of the color space). That is, both frequency and salience are domain general principles of categorization, but they differ in the degree of domain knowledge required to accurately assess each of them.

Our observation suggests that viewing color term learning as the use of domain general principles consistently throughout acquisition, without considering the development of domain specific knowledge, is missing an important part of the story. We propose instead that the observed acquisition patterns arise from the interaction of principles of categorization with growing knowledge of the domain (cf. Fig. 1, discussed in the following section). We do not mean that younger children are guided by domain general mechanisms, while older children rely on domain specific knowledge. Rather we are suggesting that the application of domain general categorization mechanisms throughout development changes as children's grasp of the domain increases, and they can rely on richer domain specific knowledge.

To support this view, we present (to our knowledge) the first large-scale quantitative analysis of a range of factors contributing to the detailed developmental patterns of errors in color term acquisition. For this we draw on the empirical production data from Wagner et al. (2013), filling an important gap in previous research using this comprehensive dataset. While Wagner et al. (2013) examined whether errors were between proximal categories or not, we explore a range of factors that influence color categorization. Several of these were studied by Yurovsky et al. (2015), but in the context of predicting children's *accurate* use of color terms (such as using *blue* only for BLUE objects). While such an approach identifies factors that contribute to ultimately successful learning, it cannot shed light on why it is that children form the particular *incorrect* categories they do along the way – why it is that, before they use *blue* accurately, they are more likely to use it to label PURPLE objects than RED ones. Our analysis of errors is aimed at understanding what drives children's formation of these categories as they are evolving.

To preview our results, we find further evidence that domain general principles of categorization play a role in color term learning across age groups, but we show that their influence varies over time depending on the degree of domain specific knowledge required to apply them. Specifically, factors like frequency that require little knowledge of color semantics have more influence on younger children's errors, while factors such as color similarity, which require greater domain knowledge, largely dominate the errors of older children. We thus provide a more nuanced view of color term acquisition, one that recognizes the important role of the interaction of a range of domain general principles with the gradually developing knowledge of the domain.

Modeling Sources of Errors

As noted above, error patterns in word learning can be informative about what drives conceptual categorization in children. In particular, the specific overextensions of a term to incorrect stimuli – such as using *blue* to refer to a PURPLE object but not a RED one – reveal the factors that children draw on in making decisions about category membership. Thus, using color terminology as a testbed, we can investigate which factors predict such categorization errors, with the specific goal of examining how the influence of those factors might change throughout children's development. Here we examine four different factors that draw on differing levels of knowledge of the semantic domain of color, across four different age groups. Our hypothesis is that: (1) because these factors are all instantiations of domain general principles that apply in categorization they will play a role across development; and (2) younger children will weigh more those factors that rely less on domain specific knowledge, while older children will weigh more those that require richer knowledge of the color domain.

To test this hypothesis, we run a series of regression analyses that consider properties of both the STIMULUS – a color patch children are asked to name in an experiment – and their *response* – the color term they use. We consider that children using a mismatching *response* for a STIMULUS (e.g., saying *blue* for PURPLE) are incorrectly assigning the STIMULUS to the *response* category (assuming PURPLE is part of the conceptual category associated with *blue*). We examine properties of both the STIMULUS and the *response* that could contribute to this miscategorization.

We consider four general principles of categorization that vary in the extent to which they draw on domain-specific knowledge. Three of these are adapted from Yurovsky et al. (2015): the use of frequency, size, and salience of a category. As detailed by Yurovsky et al. (2015), higher values on all of these have been shown to facilitate early and successful acquisition of color categories. We observe that these same principles can also explain *errors* – that is, children are more likely to (incorrectly) overextend highly frequent categories and larger categories (Saji et al., 2011), as well as more salient categories (Anglin, 1977)[2]. In addition, we consider a fourth principle of categorization: the use of similarity in assessing category fit – that is, children are more likely to overextend terms to nearby colors (e.g., Pitchford & Mullen, 2003; Wagner et al., 2013).

The instantiations of these principles in the domain of color draw on varying amounts of domain knowledge; from least to most they are frequency, size, salience, and similarity. Fig. 1 illustrates how these factors might be thought to increase in the complexity of the conceptual knowledge that is called for. Frequency requires little to no domain knowledge because it can be estimated by keeping track of the number of occurrences of the term alone. The size of a category requires some knowledge of the domain; Fig. 1 exemplifies this as the size of a patch in a simplified conceptualization of color. In contrast, assessing salience and similarity require richer knowledge of relevant dimensions in the domain. Salient colors are those far from the "neutral" area of the color space; the most salient categories are warm, chromatic colors like *yellow*, *orange*, and *red*, and the least include achromatic colors such as *white* and *gray*. Assessing salience requires at least a rough grasp of these critical dimensions of color. Finally, judging similarity relies on the most domain knowledge, as it requires assessing distance within an elaborated semantic space.

In what follows, we refer to frequency and size as lower level features, and salience and similarity as higher level features. We expect younger children to be guided more by the former, and older children more by the latter.

Materials and Methods

Children's Naming Data. The data we analyze is production data from Wagner et al. (2013), who collected color term naming data from 141 children between the ages of 1;10 and 5;1. The stimuli were samples (a colored fish or square) corresponding to the 11 English basic color terms (*red, white, yellow, black, green, blue, orange, pink, purple, grey, brown*). Each child was asked the color of each stimulus in each of two tasks (fish and squares), typically yielding up to two *response* words per child per STIMULUS. For comparability to earlier results on this dataset (Wagner et al., 2013; Yurovsky et al., 2015), we use similar filtering and age grouping criteria: We omit children with a family history of abnormal color vision, as well as those who did not cooperate on more than half of trials. Since we aim to predict errors, we also exclude children who made no errors. As in Yurovsky et al. (2015), we focus on children between ages 2 and 4, binning them into half-year groups: 2-year-olds ($n = 22$), 2.5-year-olds ($n = 29$), 3-year-olds ($n = 24$), and 3.5-year-olds ($n = 21$), for a total sample of 96 children. (An alternative would be to group children by size of color vocabulary, as in Wagner et al. (2013). Additional analysis is needed to understand which grouping better captures the underlying causal variable, but for now we note that age and color vocabulary size are highly correlated.)

Regression Analyses. We run a set of novel multiple regression analyses that use the four factors introduced above as variables of interest, along with age bin. We use only the data with a mismatch between a STIMULUS and a *response* (e.g., labeling an ORANGE fish or square with the term *red*), since the factors that contribute to errors may differ from those that best predict correct responses.

In most analyses, we consider properties of both the stimulus and the response as inputs, since either could play a role. For different analyses we use different kinds of regression approaches. In some cases we use logistic regression, and in other cases we use Poisson regression, as follows.

In sections Q1 and Q2, when we include properties of the stimulus as predictors, we restrict ourselves to logistic regression, predicting 1 for an error (1 or more children made this error) or 0 for no error. We do this because predicting the rate of error for a given stimulus–response pair leads to interdependencies among dependent variables. Considering an extreme example, if all children responded *red* to the ORANGE stimulus, the response rate for any other stimulus–response pair involving the ORANGE stimulus must be 0. Because of this, the data point for the *red* response and ORANGE stimulus "leaks" information about the other data points. We could address this concern by predicting a distribution over responses using multinomial regression, but this is not compatible with the inclusion of properties of the responses (the dependent variables) as independent variables. Additionally, there are many combinations of stimulus and response where no errors were made, and logistic regression is more appropriate for this zero-skewed distribution. Concerns about binarizing this variable are addressed in section Q3, where we taken an alternative approach and use a Poisson regression with a simplified view of the data.

Estimates of the Variable Values. We estimate frequency as the token frequency of a color term in child-directed speech in the Manchester Corpus in CHILDES.[3] Our size measure is based on data from Lindsey & Brown (2014), who collected color naming data for 330 Munsell chips from 51 adult speak-

[2]These motivations are stated in terms of the response category, but there may also be an influence of the properties of the stimulus (the category being incorrectly subsumed), as some work has considered (Pitchford & Mullen, 2002). Here we look at properties of both the response and the stimulus to see which more strongly influences children's errors.

[3]Unlike Yurovsky et al. (2015), we use the token frequency summed over all age bins, rather than cumulative frequency. We do so because cumulative frequency is highly correlated with age, which would prevent us from assessing interactions with age in a way that is comparable to how we do so for other features.

ers of American English. We take size to be the number of chips for which a color term is the modal response.[4] Both higher level features are estimated as distances in CIELAB space (Fairchild, 1998), which was designed to capture color differences. We estimate salience of a color by the distance of its focal point (taken from Berlin & Kay, 1969)[5] to light gray (specifically, the center point of the space, (50,0,0)), as in Yurovsky et al. (2015). Distance between a STIMULUS and a *response* is given by the distance between their focal colors.[6]

Features are assessed for both STIMULUS colors, and *response* terms. For a *response*, we calculate its size, salience, and distance to the STIMULUS using the color category that term (correctly) refers to; conversely, for a STIMULUS, we take its frequency to be that of its correct color term. All features were scaled to have mean zero and unit variance, to allow for maximal comparability between regression coefficients. (Note that the collinearity of our variables is within a reasonable range: in all regression analyses to follow, all VIFs range $1.0 - 2.36$.)

Q1: What factors drive overextensions?

Here we address the following questions: (a) Are the factors identified above used differentially by older and younger children? As children's domain knowledge increases, we expect them to rely less on frequency and size, and more on salience and distance (cf. Fig. 1). (b) Are properties of the response or the stimulus more predictive of errors? Most work on overextension has looked at properties of the subsuming category (here, the *response*), rather than properties of the subsumed exemplar (here, the STIMULUS). We assess whether properties of the response are indeed more important.

We treat each stimulus–response pair, per age group, as a data point, and run a logistic regression predicting 1 for an error (1 or more children made this error) or 0 for no error, based on properties of the stimulus and/or response.[7] Our dataset includes 4 age bins $*$ 11 stimuli $*$ 10 possible incorrect responses per stimulus = 440 data points in total.

To assess whether properties of the stimulus or response are more predictive of overextensions (or whether both are required), we train three models: one with only stimulus features, one with only response features, and one with both. In all models we furthermore include the distance feature, the age bin, and interactions of age with each of the variables.

Results are shown in Table 1. We find a McFadden's pseudo-R^2 of 0.15 with only stimulus features, 0.22 with only response features, and 0.25 with both. A Likelihood Ratio Test confirms that the model with both is significantly better than the model with response features only ($p = 0.003$). (We cannot run this test to compare the stimulus-only and response-only models, since the test is only applicable when one model's features are a subset of those in the other model.) Across all three models, age and distance are significant predictors. When response features are included, frequency_response and size_response are significant; when stimulus features are included, size_stimulus is significant.

feature	stimulus	response	both
frequency_stimulus	−0.34 ·		−0.34 ·
size_stimulus	−0.35 *		−0.33 *
salience_stimulus	0.18		0.30
frequency_response		0.47 **	0.46 *
size_response		0.53 ***	0.52 ***
salience_response		−0.10	−0.11
distance	−0.28 *	−0.56 ***	−0.54 **
age	−0.86 ***	−0.98 ***	−1.05 ***
age:frequency_stimulus	−0.30		−0.36 ·
age:size_stimulus	0.07		0.09
age:salience_stimulus	−0.01		0.01
age:frequency_response		−0.63 **	−0.71 ***
age:size_response		−0.28	−0.27
age:salience_response		0.54 **	0.51 **
age:distance	−0.02	−0.16	−0.02

Table 1: Logistic regression experiments with features of the stimulus, response, and both. Significant results are bolded.

Positive coefficients for frequency_response and size_response show that children are more likely to overextend frequent terms, and terms that can be used to describe a diverse set of colors, while a negative coefficient for salience_response means that children are less likely to incorrectly overextend terms for warm, chromatic (more salient) colors. Distance also has a negative coefficient, indicating that children are more likely to make errors with neighboring color terms. The negative coefficient for size_stimulus shows that children make fewer errors on stimuli from large categories.

We also find two significant interactions with age. The predictor age:frequency_response has a negative slope, indicating that younger children are more likely to overextend highly frequent terms, and the predictor age:salience_response has a positive slope, indicating that older children are more likely to overextend highly salient terms. Given that we do not find a main effect for salience of the response here, we return to whether this is a meaningful finding in the next section.

Our results are consistent with earlier findings, while extending our understanding of color term learning. Although frequency, size, and salience are predictive of accuracy (Yurovsky et al., 2015), frequency and size of response terms appear to be more relevant than salience for predicting how terms are incorrectly overextended. While earlier analyses showed a difference in error rates based on a binary distinction of adjacent/non-adjacent categories (Pitchford & Mullen, 2003; Wagner et al., 2013), the negative correlation

[4]This is a more standard measure of size than that of Yurovsky et al. (2015), who used proportion of the chips for which all subjects used the same label (which may be undefined for some colors).

[5]In cases where multiple focal chips were indicated for a term, we took the average point in CIELAB space of the designated chips.

[6]Given evidence that the location of focal colors is universal (Berlin & Kay, 1969; Regier et al., 2005), we assume (at least an approximation of) this knowledge is available to children.

[7]We ran three alternative experiments with error thresholds of 2, 3, and 4. For example, when the error threshold was 2, we predicted 1 for 2 or more errors, and 0 for fewer than 2 errors. We found that this had little impact on the results.

here of errors with distance in color space suggests a continuous measure of distance may be more appropriate.

We also show that errors depend more on properties of the overextended term than of what it incorrectly labels. This confirms that features of the *response* category being extended to this STIMULUS instance are more important (or perhaps simply better understood by the child) than properties of the stimulus being incorrectly labelled as that category.

Finally, the interactions with age partially support our hypothesis that the features are used differentially by younger and older children. We next further explore these patterns.

Q2: Does factor influence vary with age?

Our initial regressions show that some properties of the response have significant interactions with age, specifically that younger children are more likely to overextend frequent terms, and older children may be more likely to overextend salient terms. This is partial support for our hypothesis that features requiring less domain knowledge (frequency and size) are used more by younger children, while older children shift to more reliance on features requiring a better understanding of the domain (salience and distance). Here, we perform additional qualitative and post hoc analyses to explore in finer detail how these variables impact errors over the different ages.

Qualitative Error Patterns

We begin with a qualitative look at the most common errors per stimulus for the youngest and oldest age groups, to provide an intuition for the patterns of errors seen across the different ages; see Fig. 2. Errors of 2-year-olds mainly involve overextending the high frequency terms *blue* and *green*. These correspond to the two largest color categories, and are among the three terms with very high frequency. Interestingly, the highest frequency term, *red*, is not the most common error for any stimulus. The category RED is very small (the third smallest), suggesting that, as predicted, both response frequency and size play an important role in youngest children's overextensions.

The 3.5-year-olds' most common errors cover a more diverse set of terms. Generally, the patterns seem to support our hypothesis, that older children's incorrect terms may be more driven by semantic similarity of categories (e.g., using *purple* for PINK or *yellow* for ORANGE) and their salience (YELLOW and ORANGE are the most salient colors). However, these observations are suggestive only, due to the low number of errors for any given stimulus in this age group.

A Closer Look at Age-Related Influences

In order to quantitatively assess the magnitude and direction of the effects of our features across the ages, we turn to a set of logistic regressions of the same form as in the analyses for Q1. Here we consider a single factor at a time, to clearly see its relation to the presence of error for a given stimulus–response pair. To reduce the number of features we consider, we focus on response features (frequency_response, size_response, salience_response, distance) since they contributed most in the earlier models. With each feature as the predictor, we run a regression for each of the four age groups, for a total of 16 logistic regressions.[8] We then plot the resulting beta coefficients by age for each predictor; these indicate the magnitude and direction of the relation of the predictor to the presence of error. This allows us to see how the relation for each feature changes with age.

Color Stimuli	2 year olds	3.5 year olds	
RED	*green*	no errors	RED
ORANGE	*blue*	*yellow*	ORANGE
YELLOW	*blue*	*orange*	YELLOW
GREEN	*blue*	*blue*	GREEN
BLUE	*green*	no errors	BLUE
PURPLE	*blue*	*blue*	PURPLE
PINK	*blue*	*purple*	PINK
BROWN	*black*	*orange*	BROWN
WHITE	*blue*	*black*	WHITE
GRAY	*white*	*white*	GRAY
BLACK	*blue*	*brown*	BLACK

Figure 2: Modal errors for the youngest/oldest age groups.

Based on the significant interactions between age and frequency_response and between age and salience_response found earlier, we expect the coefficient for frequency_response to decrease with age, and the coefficient for salience_response to increase with age. In addition, given our hypothesis regarding the lower level and higher level features, we expect that size_response will also be more predictive of errors for younger children, and dist_response will be more predictive of errors for older children. (Note that we expect distance to be a negative predictor for older children, since it is the inverse of similarity). For these latter two features, given the lack of significant interactions in the analyses for Q1, we are interested to see whether age modulates the relation between the variables and the error pattern in a more complex fashion.

Fig. 3 shows how the beta coefficients for each variable vary by age group. We discuss each predictor in the order shown, which reflects the increasing degree of domain knowledge posited in Fig. 1.

For frequency, the coefficients decrease monotonically with age, in line with our hypothesis. Frequency is only a significant predictor for the two youngest age groups. This is consistent with the significant, negative slope for age:frequency_response found in the analyses for Q1. As hy-

[8] Each regression thus has one quarter the data points compared to those in the previous section – i.e., 110 data points. We used Holm-Bonferroni correction per-variable (i.e., for each variable, the p-values considered significant summed over age bins must sum to < 0.05).

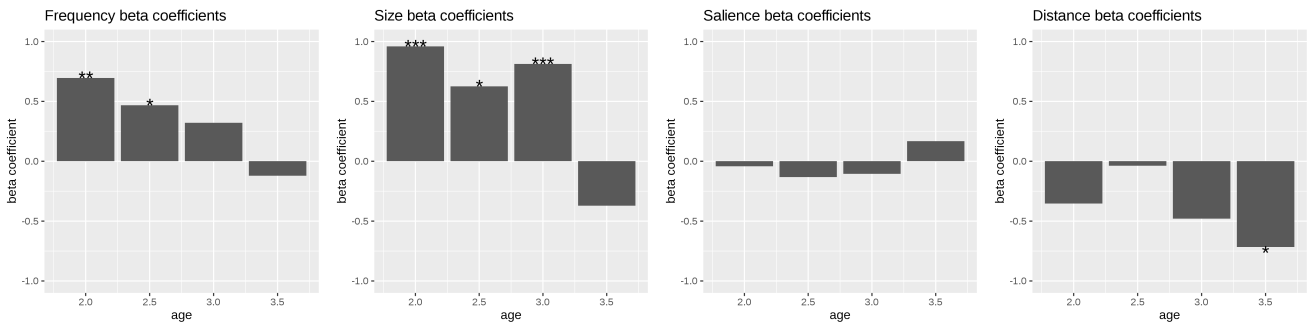

Figure 3: Beta coefficients by age group, for each predictor.

pothesized, size is also only significant for the younger children. Here, the pattern of decrease in the coefficients is more abrupt than with frequency, perhaps explaining its lack of a significant interactions with age in the earlier analyses.

Although the coefficients for salience generally increase numerically with age, it is not found to be a significant predictor of error for any age group. We suspect the significant interaction between age and salience of the response found in the previous section may be due to interactions with the other variables. While collinearity of predictors might be suspected to be an issue, the VIF was within a reasonable range, as noted earlier. Further analyses are needed to understand this result.

The distance coefficients show a clear pattern consistent with our hypothesis: their effect mostly increases with age, and they are only significant for older children. As with size for younger children, the decrease is not completely smooth, in line with the lack of significant interaction with age found above. This differs from the Wagner et al.'s (2013) finding that overextensions of a response term to adjacent (proximal) stimuli is above chance rates for all age groups. However, the measure that they use is not the same as our distance measure (adjacency is a binary property and distances is a continuous property), so additional analyses are required to understand how our results relate to theirs.

Based on the analyses for Q1 and these follow-up regressions, we find support for the hypothesis that younger children are more likely to overextend frequent terms or large terms – features that depend less on knowledge of the color domain – while older children are more likely to overextend terms based on similarity – which requires fine-grained knowledge of the color space.

In addition, we find evidence for a gradual shift in which features children attend to. The beta coefficients for frequency and distance are both suggestive of a shift taking place over multiple age bins. We also find that features come in and out of significance at different ages. On the other hand, we find that frequency, size, and salience all change sign at age 3.5, which is the same age where distance becomes significant. This is compatible with the observation of Pitchford & Mullen (2002) that children's knowledge of the domain of color changes substantially around this age. However, our results suggest their strong conclusion that children undergo a shift from lack of knowledge to full knowledge of the domain in a three month period may be too strong.

Q3: Individual differences beyond age effects?

Our analyses above suggest that children from different age groups show different error patterns. Here, we address individual differences in error patterns, focusing on two main questions: (a) Do children exhibit individual differences beyond age effects? Others have shown that there is high individual variation in color naming (e.g., Roberson et al., 2004), and we look at how this relates to age variation. (b) Do we find similar results to our analyses for Q1 when we take individual variation into account? This allows us to verify that the results from our analyses for Q1 are robust, and not driven by a small number of children.

Here, we predict the number of errors each child made per response.[9] We use Poisson regression since it is appropriate for predicting count data, especially in cases where the dependent variable is often zero. Comparing the Q1 results to results using Poisson regression also helps to address concerns about formulating this as a binary prediction problem in the earlier sections. In all models, we include all features of the response (frequency_response, size_response, and salience_response). To address question (a), we compare three models: one with fixed effects for age and the interactions with age that came out as significant in the analyses for Q1;[10] one with random intercepts per child (and no age variables); and one with both child random intercepts and age variables. By comparing these models we explore whether (i) individual differences in propensity to make errors can as fully account for the data as the age variables – i.e., age effects reduce to individual differences or v.v.; or (ii) both age and individual differences contribute to children's error patterns – i.e., there are individual differences above and beyond the age effects shown above.

[9] We predict per response, rather than per stimulus–response pair, because of the earlier mentioned interdependencies between DVs. There are 1056 data points = 96 children $*$ 11 response terms.

[10] This excludes the interaction age:size_response. We made this choice because a regression including all three interactions did not converge.

We find that the model with random intercepts for children performs better than the model with age variables (AIC of 2535 vs. 2590), and the model with both random intercepts and age variables performs best (AIC of 2444). This model is the most complex model justified by the data, and achieves a conditional R^2 of 0.18. This means that children's probability of making errors varies by individual, as well as by age.

To address question (b), assessing the robustness of the results in the analysis for Q1, we examine the coefficients and significance of fixed effects in this model. Each variable found to be significant in the analyses for Q1 also comes out as significant here ($p < 0.05$), with matching signs. We also find that salience of the response is a significant, negative predictor, while in Q1 only its interaction with age was significant. This analysis provides evidence that the relevance of these features is not simply driven by a small number of children, and confirms that the results presented in earlier sections were not due to our choice of a binary prediction scheme.

Conclusions

We analyze the detailed patterns of children's overextension errors in the domain of color, finding evidence that these error patterns are driven by an interaction between domain general principles of categorization, and children's developing conceptual organization in the domain of color. Younger children's errors are driven more by features that require little domain knowledge, like response frequency and category size, and older children's errors are driven more by features like similarity, which requires more complex, domain-specific computations. Our results here provide a more nuanced picture of word–meaning mapping in this domain as a categorization process: While domain general processes play a role throughout development, as suggested by Yurovsky et al. (2015), their influence varies across ages according to their dependence on domain specific knowledge. Although we find notable differences between the oldest children and the others, as proposed, e.g., by Pitchford & Mullen (2002), our results suggest that this may be the culmination of a more gradual increase in application of such knowledge, rather than the abrupt shift in understanding that they propose.

Acknowledgments

We are grateful to the authors of Wagner et al. (2013) and Lindsey & Brown (2014) for sharing their data, and Blair Armstrong for consultation on the statistical analyses. We thank the reviewers for their detailed and constructive comments on the paper. JW gratefully acknowledges the support of NSERC grants RGPIN-2019-06917 to BB and RGPIN-2017-06506 to SS.

References

Anglin, J. M. (1977). *Word, object, and conceptual development*. Norton.

Bateman, W. (1915). The naming of colors by children: The Binet test. *The Pedagogical Seminary*, *22*(4), 469–486.

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley, CA: UC Press.

Bornstein, M. H. (1985). On the development of color naming in young children: Data and theory. *Brain and language*, *26*(1), 72–93.

Clark, E. V. (1973). What's in a word? on the child's acquisition of semantics in his first language. In T. E. Moore (Ed.), *Cognitive development and acquisition of language* (pp. 65–110). Elsevier.

Davies, I., Corbett, G., McGurk, H., & Jerrett, D. (1994). A developmental study of the acquisition of colour terms in setswana. *Journal of Child Language*, *21*(3), 693–712.

Davies, I., Corbett, G., McGurk, H., & MacDermid, C. (1998). A developmental study of the acquisition of Russian colour terms. *Journal of Child Language*, *25*(2), 395–417.

Fairchild, M. D. (1998). *Color appearance models*. Reading, MA: Addison-Wesley.

Gentner, D., & Bowerman, M. (2009). Why some spatial semantic categories are harder to learn than others. The Typological Prevalence Hypothesis. In J. Guo, E. Lieven, N. Budwig, S. Ervin-Tripp, K. Nakamura, & S. Özcaliskan (Eds.), *Crosslinguistic approaches to the psychology of language. Research in the tradition of Dan Isaac Slobin* (pp. 465–480). New York: Psychology Press.

Lindsey, D. T., & Brown, A. M. (2014). The color lexicon of American English. *Journal of vision*, *14*(2), 17–17.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, *115*(1), 39.

Pitchford, N., & Mullen, K. (2002). Is the acquisition of basic-colour terms in young children constrained? *Perception*, *31*(11), 1349–1370.

Pitchford, N., & Mullen, K. (2003). The development of conceptual colour categories in pre-school children: Influence of perceptual organization. *Visual Cognition*, *10*(1), 51–57.

Regier, T., Kay, P., & Cook, R. S. (2005). Focal colors are universal after all. *PNAS*, *102*(23), 8386–8391.

Roberson, D., Davidoff, J., Davies, I. R., & Shapiro, L. R. (2004). The development of color categories in two languages: a longitudinal study. *Journal of Experimental Psychology: General*, *133*(4), 554.

Saji, N., Imai, M., Saalbach, H., Zhang, Y., Shu, H., & Okada, H. (2011). Word learning does not end at fast-mapping: Evolution of verb meanings through reorganization of an entire semantic domain. *Cognition*, *118*(1), 45–61.

Wagner, K., Dobkins, K., & Barner, D. (2013). Slow mapping: Color word learning as a gradual inductive process. *Cognition*, *127*, 307-317.

Yurovsky, D., Wagner, K., Barner, D., & Frank, M. C. (2015). Signatures of domain-general categorization mechanisms in color word learning. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 2775–2780).