Perceptual, Conceptual, and Frequency Effects on Error Patterns in English Color Term Acquisition

Barend Beekhuizen
Leiden University Centre for Linguistics
Leiden University
barendbeekhuizen@gmail.com

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

Abstract

Children's overextension errors in word usage can yield insights into the underlying representation of meaning. We simulate overextension patterns in the domain of color with two word-learning models, and look at the contribution of three possible factors: perceptual properties of the colors, typological prevalence of certain color groupings into categories (as a proxy for cognitive naturalness), and color term frequency. We find that the perceptual features provide the strongest predictors of the error pattern observed during development, and can effectively rule out color term frequency as an explanation. Typological prevalence is shown to correlate strongly with the perceptual dimensions of color, and hence provides no effect over and above the perceptual dimensions.

1 Overextensions in word learning

When learning their language, children often overextend a word by erroneously using it to refer to concepts similar to its actual meaning – e.g., a child learning English might refer to all round things as *ball*. We can learn much about the mechanisms and representations the child uses to arrive at an adult level of understanding by exploring whether the proposed mechanisms lead to observed patterns of such errors over the course of development.

Several factors have been named as potential influences on early overextensions in word meaning acquisition, including underspecification of semantic representations (Clark, 1973), as well as word frequency (mostly invoked as a zero-hypothesis to be rejected; Gülzow and Gagarina (2007), Goodman et al. (2008)).

Another possible factor is conceptual prior biases. Bowerman (1993) suggests that some semantic features (or values of features) may be cognitively more readily available than others, and argues that (crosslinguistic) semantic typology can shed light on the degree of cognitive naturalness of features in a domain. This idea was further articulated by Gentner and Bowerman (2009), who proposed the Typological Prevalence Hypothesis. This proposal states that the more frequently languages make a certain semantic grouping – i.e., collect together a certain set of situational meanings under a single term – the more likely this is a cognitively natural grouping. The reasoning is that if some conceptual categorization comes naturally, languages are more likely to develop linguistic categorization systems that follow these biases. Gentner and Bowerman (2009) further argue that, other things being equal, linguistic terms referring to such cognitively more natural groupings will be acquired more readily by children than terms in a language that do not follow the typical conceptual category boundaries.

The Typological Prevalence Hypothesis explains the error pattern Gentner and Bowerman (2009) observed in the acquisition of Dutch topological spatial markers. Whereas English uses the preposition *on* for all sorts of conceptual relations of support between a figure object and a ground object, Dutch distinguishes *op* 'surface support', *aan* 'tenuous support', and *om* 'surrounding (support)'. Gentner and Bowerman (2009) found experimentally that Dutch children overgeneralize *op* to situations where adults would use *aan* or *om*, but not vice versa. Gentner and Bowerman (2009) note that it is crosslinguistically very common to have a term like *op* that reflects a semantic grouping of various surface support relations, whereas terms such as *aan* that denote 'tenuous support' are typologically rare. They suggest that this pattern reflects a difference in cognitive naturalness (surface support being the more prototypical case of support than tenuous support), which in turn

makes *aan* harder to acquire than *op*.

Beekhuizen et al. (2014) operationalized the Typological Prevalence Hypothesis within a word-learning model by creating a semantic representation for topological situations that used the words themselves from across a number of languages as the features for representing spatial relations. In such a representation, commonalities and differences in the way languages carve up the space of topological relations is reflected in the way the terms within each language group together various situations. This approach yields a semantic representation that can capture crosslinguistic naturalness of the underlying spatial relations, without the need for explicit identification of appropriate semantic features. Situations that, within many languages, are expressed with the same word are closer in this semantic space than those that are more often labelled by different terms within a language. Beekhuizen et al. (2014) simulated the above experimental results on Dutch children by using this semantic space within a computational model for learning Dutch prepositions, whose developmental trajectory displayed the same trends as children.

Here we extend the method of Beekhuizen et al. (2014) to the acquisition of color terms, another domain in which children are known to make overextension errors. Color terms form an interesting test of the Typological Prevalence Hypothesis, because we know reasonably well what the perceptual dimensions of color are, and can test if there is any effect of typological prevalence on top of this. Specifically, we ask if crosslinguistic consistency provides a good basis for the representation of color in word learning, and if such a semantic representation adds information beyond the perceptual properties of color.[1]

Note that other work, such as Regier et al. (2007) among others, has reasoned from the perceptual features of color as well as general considerations concerning category structure to propose an explanation for the observed tendencies across color lexicons. Instead, we explore whether the typological tendencies among color lexicons reflect semantic information relevant to word learning, and especially whether that information goes beyond that provided by perceptual features. We refer to the typologically-based semantic representation as 'conceptual' features (in contrast to perceptual ones) because they refer to the way color concepts are (preferably) structured in the lexicons of the various languages.[2]

Thus, here we explore three potential influences on the error patterns observed in learning of color terms: perceptual factors, conceptual factors, and word frequency effects. We also take the opportunity to strengthen the evaluation method of Beekhuizen et al. (2014) by here using a quantitative measure of model deviation from the observed pattern of word use in order to arrive at more complete insights into the role of these factors.

2 Data on the acquisition of color terms

Across languages, children overextend certain color terms at the cost of others, and there has been a long tradition of research into this domain (Bateman, 1915; Istomina, 1960; Harkness, 1973; Bartlett, 1978; Davies et al., 1994; Davies et al., 1998; Roberson et al., 2004). The case used for our current study is Bateman (1915), who studied 591 English-speaking children in the age range 6-11. Eight color chips of the 'best' examples[3] of the colors BLACK, BLUE, BROWN, GREEN, ORANGE, PURPLE, RED, YELLOW were presented to the subjects, who were then asked to name the color.[4] We use Bateman's elicitation data in the initial application of our approach to this domain because, despite being a century old, it remains the most comprehensive published error data on color terms.

Bateman found that BLACK, WHITE, RED and

[1] For the latter question, the hypothesis is that a color c may be at the same perceptual distance to c' as it is to c'', but for some other reason, languages categorize c and c' with the same term more often than c and c''. There could be various reasons for this difference, such as a preference for certain category structures, or communicative pressures concerning disambiguation. We do not investigate here what those factors might be, but rather explore whether the typologically-derived semantic space provides information in addition to the perceptual features.

[2] A reviewer noted that 'conceptual' may be an inaccurate term, since factors beyond strictly the conceptual biases of language users might influence color lexicons and their crosslinguistic similarities and differences. In adopting the Typological Prevalence Hypothesis as a working hypothesis, we consider that crosslinguistic patterns reflect cognitively natural conceptual groupings, while acknowledging that other factors need to be investigated as well.

[3] "Each color was of the purest tone and strongest saturation obtainable", p. 476.

[4] We adopt the convention of denoting the stimuli with small capitals and the words with italics. The responses contain all eleven English basic color terms (Berlin and Kay, 1969): *black, white, red, yellow, green, blue, orange, purple, pink, brown* and *grey*.

BLUE were learned (nearly) error-free ($\leq 2\%$ erroneous responses at age 6), but YELLOW (7% at age 6), GREEN (6% at age 6), ORANGE (6% at age 6), and especially PURPLE (11% at age 6) displayed errors. For YELLOW, the term *orange* is the most frequent error. For ORANGE various errors are found (*yellow, red, blue, purple, brown, pink*). GREEN displays mostly errors in which *blue* is used. For PURPLE, *blue* is the most frequent erroneous term. Whereas the errors for YELLOW and GREEN have disappeared at age 7, the errors for ORANGE and PURPLE are somewhat more persistent, and are found until age 11 and 9 respectively.

In summary, this data yields the following five phenomena that must be explained:

- BLACK, WHITE, BLUE, and RED display hardly any errors;
- GREEN and YELLOW display some errors at age 6 but none afterwards;
- ORANGE displays (somewhat haphazard) persistent errors;
- PURPLE displays persistent errors, mostly *blue*;
- However, *purple* is not overextended to BLUE.

While previous accounts of the error patterns have mainly focused on perceptual closeness of the various colors (Bartlett, 1978; Pitchford and Mullen, 2003), this cannot be the full explanation: If the overextension of *blue* to PURPLE stimuli was solely due to color similarity, we would expect (contrary to observation) that *purple* would also be incorrectly overextended to BLUE stimuli.

Here, we explore three potential factors that might lead to the observed pattern of color errors: perceptual features of color, conceptual/typological prevalence factors, and/or frequency of the color terms.

3 Operationalizing the Three Factors

We simulate the acquisition of color terms by training a word-learning model on a generated input stream, in which each input item pairs a semantic representation $s \in S$ of a color, with a color term $t \in T$ used to refer to it. S is drawn from the 330 chips of the Munsell color chart, and T contains the eleven basic color terms that comprised the responses in Bateman (1915). We explore the impact of perceptual and/or conceptual (typological) factors by varying the representation of s, using one or both of the feature sets described in Sections 3.1 and 3.2.[5] The role of frequency of t is examined by varying the way the input items are generated, as in Section 3.3.

3.1 Perceptual features

As the perceptual dimensions, we use the CIELab color space. The CIELab space describes all colors visible to the human eye, and consists of three dimensions, lightness (L^*), a red-green scale (a^*) and a yellow-blue scale (b^*). Importantly, the Euclidean distance between any pair of coordinates in CIELab is thought to directly reflect the perceptual similarity between colors. Since color perception is thought to be adultlike before age two (Pitchford and Mullen, 2003), we can assume these perceptual features to be stable over development.

3.2 Conceptual features

The conceptual dimensions reflect the crosslinguistic biases in categorizing the color space. To capture these, we use the World Color Survey data of Kay et al. (2009), which contains elicitations for each of the 330 chips of the Munsell color chart, for 110 typologically diverse languages, with on average 24 participants per language. From this data, we extract an n-dimensional conceptual space by using the first n dimensions of a Principal Component Analysis (PCA, Hotelling (1933)) over the elicited color terms for a number of color stimuli following the method of Beekhuizen et al. (2014), as described below.

The elicitations for each language give us a count matrix C containing a set of color stimuli S on the rows, and a set of color terms T in that language on the columns. Every cell is filled with the count of participant responses to stimulus s that use color term t. Matrix C captures the way that the language carves up the space of color: stimuli s and s' are treated similarly in the language to the extent that the labels used to express them are similar, reflected in rows s and s' of C. As we want to know how often stimuli are co-categorized *across* languages, the procedure of Levinson et al. (2003) is adapted: for every language l, an $|S| \times |S|$ distance matrix D^l containing the Euclidean distances between all pairs of situations is extracted. By summing the distance matrices for all languages, we arrive at a distance matrix D^{all}

[5] The values for the 330 chips on the two feature sets are available from the first author upon request.

whose elements d_{ij} are the summed distances between s_i and s_j across all languages. A PCA was applied to D^{all}, from which we use the 4 components with an Eigenvalue ≥ 1 (Kaiser's rule) as our conceptual space to represent color semantics.

3.3 The role of frequency

In the input generation procedure, a pair of a color term $t \in T$ and a stimulus $s \in S$ is sampled from the distribution $P(t, s) = P(s|t)P(t)$. The likelihood $P(s|t)$ is the relative frequency of a specific color chip given a term (as given by the data for English of Berlin and Kay (1969)):

$$P(s|t) = \frac{n(t, s)}{\sum_{s' \in s} n(t, s')} \quad (1)$$

where $P(s|t) = 0$ for s not included in the elicitation data.

To explore the role of term frequency in color errors, we base the prior probability $P(t)$ on the relative frequency of t among the 11 primary color terms in the Manchester corpus of child-directed speech (Theakston et al., 2001). We then compare this to holding frequency constant, i.e. with $P(t)$ a uniform distribution over T.

4 The Experimental Approach

4.1 The learning models

We model word-learning as a categorization problem by considering the 11 color terms as the "categories" to be learned over the various color semantics (the representations of the color chips) each is associated with in the input. Extending Beekhuizen et al. (2014), we try two different categorization models: a Gaussian Naïve Bayes learner (GNB, as in their work), and a Generalized Context Model (GCM, Nosofsky (1987)), for two reasons. First, if the same effects are found with multiple models, the effect is more robust, and not an effect of the model per se. Second, GCM is an exemplar-based categorization model that has been shown to simulate human categorization behavior well.

In the GNB approach, for a given amount of input data of color-semantics/color-term pairs, the model estimates Gaussian distributions over each of the perceptual and/or conceptual feature dimensions. The model is then presented with each of Bateman's 8 color stimuli as the test phase, and it outputs the color term with the Maximal A Posteriori probability as the predicted category for each color.

In the GCM model, the probability of categorizing a color stimulus s_i with category J (response R_J, a color term) is given as the summed similarity η between s_i and all instances of category J (all colors referred to by the color term), divided by the summed similarity between s_i and all exemplars (colors) in the data set.

$$P(R_J|s_i) = \frac{b_J \sum_{j \in C_J} \eta_{ij}}{\sum_K (b_K \sum_{k \in C_K} \eta_{ik})} \quad (2)$$

where b is the category bias, here set to uniform for categories. η_{ij} is given by:

$$\eta_{ij} = e^{-d_{ij}^{\delta}} \quad (3)$$

where δ is the decay function, here set to 1 (exponential). For d we use the Euclidean distance between the coordinate vectors of i and j.

4.2 Experimental set-up

Each model is trained on successively larger amounts of data, in blocks of 10 input pairs. Every 10 input items, the model is presented with the 8 colors of Bateman (1915) and predicts the most likely category label from the set of 11 color terms. As Bateman does not give values in a color space for his stimuli, we assume that the focal colors, as described by Berlin and Kay (1969), were used.[6] For each of the 12 parameter settings ($\texttt{features} = \{\texttt{perc}, \texttt{conc}, \texttt{perc\&conc}\} \times \texttt{frequency} = \{\texttt{relative}, \texttt{uniform}\} \times \texttt{model} = \{\texttt{GCM}, \texttt{GNB}\}$), we run 30 simulations of 1000 input items each, each of which yields 100 test points.

4.3 Evaluating the model predictions

Assessing the accuracy of the model in simulating the observed error data requires us to align the predictions P at the 100 test moments of the model with Bateman's observed data O in the 5 age bins (6-, 7-, 8-, 9-, and 10-to-11-year-olds). We represent O as a 5×8 matrix in which each element o_{ij} is the distribution of responses over the children at age bin i to color stimulus j (where j is one of the 8 stimulus colors). The matrix P contains the

[6] If multiple tokens were named as focal in the data of Berlin and Kay (1969), we set coordinates of a test item to the mean of each coordinate for all focal instances of that category.

models responses under a given parameter setting; it is a 100×8 matrix in which each element p_{kj} is the distribution of responses over the 30 simulations at test point i to color stimulus j. For example, p_{kj} for j the RED stimulus might look like:

$$p_{kj} = [\ \textit{red} : 0.8, \textit{orange} : 0.1, \textit{purple} : 0.1, \ldots\ ]$$

indicating that of the 30 simulations at test point k, 24 predicted *red*, 3 *orange*, and 3 *purple*, to the stimulus j=RED (and 0 responses for all other color terms). To recap, each row of O and P is a vector of 8 elements, each of which is a distribution over the 11 color terms that comprises the responses of the children/model at that age/test point, respectively, to the 8 color stimuli.

To determine the degree to which the predictions of the model given in P mimic the error data in O, we need to map each row i of O (the responses for that age bin) to some row k in P, such that each o_{i+1} maps to a higher k than o_i. (This constraint ensures that older age bins map to later test points of the model.) To find this mapping between observed and predicted data, we find the series of 5 (possibly discontiguous) rows in P that minimize the average distance between those 5 rows and the 5 rows of O.

To compare rows o_i and p_k, we find (and average) the distance d between each paired distribution (e.g., RED in o_i and RED in p_k):[7]

$$\Delta(o_i, p_k) = \sum_{s \in S_{\text{test}}} d(o_i^s, p_k^s) \times \frac{1}{|S_{\text{test}}|} \quad (4)$$

where S_{test} is the set of 8 test colors. Using $\Delta(o_i, p_k)$, we compare all o_i and p_k (subject to the ordering constraint on k) and find the series of 5 p_{k_i}'s with the lowest distance to the o_i's they are mapped to.

Now we can calculate the overall **error** of the model's predictions P with respect to the observed data O as:

$$\mathbf{error}(O, P) = \left(\frac{\sum_{i \in [1\ldots5], k_i} \Delta(o_i, p_{k_i})}{5} \right) \quad (5)$$

where the indices k_i are given by the mapping that minimizes the **error**, as explained above.

[7] The experiments reported below use Euclidean distance for d, but the pattern of results is the same under cosine or Canberra distance.

5 Results and discussion

5.1 Global fit and effect of parameters

In order to study the effect of the various parameters ($\texttt{features} = \{\texttt{perc}, \texttt{conc}, \texttt{perc\&conc}\} \times \texttt{freq} = \{\texttt{relative}, \texttt{uniform}\} \times \texttt{model} = \{\texttt{GCM}, \texttt{GNB}\}$), we enter the **error** for the output for each setting into a two-way ANOVA. As we can see in Table 1, there are two main effects: the `features` and the `model`. A post-hoc test (Tukey HSD) shows that for the `features` variable, the difference between `perc` and `conc` ($p < 0.001$) as well as between `perc&conc` and `conc` ($p < 0.001$) are statistically significant, but not the difference between `perc&conc` and `perc` ($n.s.$). For the `model` parameter, we observe a slightly better fit for GCM than for GNB. For the `freq` parameter, there is no difference between `relative` and `uniform`.

The analysis shows that the perceptual features perform better than the conceptual features, and adding the conceptual features to the perceptual ones gives no improvement. It seems that perceptual features play an important role in explaining the overextensions and lack thereof in the development of color terminology, but that the conceptual features explain little on top of this.

The lack of an effect of the conceptual features is unexpected, given that Beekhuizen et al. (2014) found that using their typological conceptual space explained the errors in the acquisition of Dutch spatial relation terms. One could argue that the domain of color is conceptually simpler than space (pertaining to properties of entities rather than relations between them, cf. Gentner (1982)), which is supported by the finding of Majid et al. (2015) that, at least among Germanic languages, space lexicons vary more crosslinguistically than color lexicons. However, the fact that children acquire color terms relatively late (compared to spatial terms) goes against this analysis, but then again, the late acquisition of color may also be due to other factors (e.g., the difficulty of disentangling color from other properties, cf. Soja (1994)). Understanding the lack of an effect of the conceptual features here ultimately requires us to analyze the crosslinguistic data further, which we plan to do in future work.

We also found no significant effect of the frequency manipulation, suggesting that the observed errors are not influenced by the varying frequencies of color terms. This is surprising because

parameter	F	p	parameter setting	mean **error**
`features`	$F(2) = 2790.070$	$p = 0.000$	`perc&conc`	$\mu = 0.015$
			`perc`	$\mu = 0.020$
			`conc`	$\mu = 0.354$
`frequency`	$F(1) = 0.026$	$p = 0.887$	`relative`	$\mu = 0.130$
			`uniform`	$\mu = 0.130$
`model`	$F(1) = 11.208$	$p = 0.044$	`GCM`	$\mu = 0.120$
			`GNB`	$\mu = 0.139$

Table 1: Results of the ANOVA; see Section 5.1 for post-hoc analyses.

Beekhuizen et al. (2014) found that an interaction of frequency and typological factors contributed to the errors they modeled. Moreover, frequency has been shown to correlate with acquisition of color terms (Yurovsky et al., 2015), albeit for younger children than the ones in the Bateman data.

This suggests that a possible explanation for the lack of both frequency and typological prevalence effects is that the error data we are modeling are from older children (ages 6–11). Perhaps effects of frequency and/or conceptual factors (on the basis of typological prevalence) are only found in younger children. It may be that, by age 6, the young language user has organized her semantic space in accordance with her native language, thus no longer displaying effects of typological prevalence. In the future we will need to look at earlier error data to explore whether the factors involved vary in their importance during the development of a vocabulary: frequency and conceptual biases may have certain effects early on, but factors pertaining to perceptual dimensions leading to the overextension of category boundaries may be more persistent.

5.2 Findings per color

Here we look at the results of the model per test color, considering the role of the different feature spaces, `perc` and `conc`, and of the different frequency settings, `uniform` and `relative`, used for calculating the prior probability of the color terms. In addition to looking at the overall **error** of the model's predictions (Table 2), we also look at the actual responses in some of the interesting cases. Even though the `frequency` setting made no difference overall in the amount of **error**, we show results for both settings, since it affects the pattern of responses for some individual colors. All these results use the GCM model, since it performed slightly (but statistically significantly) better than the GNB model.

Recall that the first two observed error patterns to be explained (see Section 2) are that there are no overextensions for BLACK, WHITE, RED, and BLUE, and few, non-persistent overextensions for YELLOW and GREEN. Regarding these color stimuli, we find that the model provides a good fit under all settings for `features` and `frequency`. In all cases, the **error** is caused by underestimation of the model of the few overextensions that are there, that is: the model predicts no overextensions for these six stimuli, whereas there are some.

The next two phenomena concern the persistent errors for ORANGE and PURPLE, where other color terms are overextended by even older children to these stimuli. For these two stimuli, the model fit is slightly worse than for the other colors when using the **perc** features, but the setting of **conc** features alone worsens the fit with a dramatic increase in the model **error**.

For ORANGE, the model behaves similarly as with the previous 6: it predicts no overextensions (for `perc&conc` and `perc`) or a complete overextension of *red* (for `conc`). As such, we cannot explain the observed overextension pattern for ORANGE well at this point. However, we can exclude term frequency as an explanation: under both settings for `frequency`, the model has the same fit with the observed pattern.

The results for PURPLE, the other color with persistent overextensions, display a number of noteworthy effects. Here, in addition to the model **error** in Table 2, we also show figures with the proportion of responses to PURPLE over time, for both the child data and for the model under several interesting settings; see Figure 1.

First, the model under all settings does predict overextensions of other color terms to PUR-

	BLACK	BLUE	GREEN	ORANGE	PURPLE	RED	WHITE	YELLOW
perc&conc, uniform	0.000	0.005	0.013	0.029	0.024	0.003	0.000	0.011
perc, uniform	0.000	0.005	0.013	0.029	0.026	0.003	0.000	0.011
conc, uniform	0.000	0.019	0.030	1.000	0.854	0.003	0.000	0.011
perc&conc, relative	0.000	0.005	0.013	0.029	0.036	0.003	0.000	0.011
perc, relative	0.000	0.005	0.013	0.029	0.015	0.003	0.000	0.011
conc, relative	0.000	0.028	0.013	1.000	0.852	0.003	0.000	0.011

Table 2: Mean **error** per stimulus, in the GCM model.

PLE. Focussing on the settings with a good fit (perc&conc and perc), we find that the term *blue* is in all cases persistently overextended to PURPLE. However, the various settings do provide different overextension patterns, as can be seen in Figure 1. The setting with the closest fit (**error** $= 0.014$) is pred, relative (Fig. 1d): here we see a pattern most similar to that found in child data (Fig. 1a). From the fact that the model **error** for this setting is about twice as low as the settings with uniform frequency and with conceptual dimensions we can infer two things. First, we do find a frequency effect: *blue* being more frequent than *black* in child-directed speech explains why there are more overtextensions of *black* given the setting perc, uniform (Fig. 1c) than given perc, relative. Second, the conceptual dimensions hurt the prediction of the overextension pattern. Including the conceptual dimensions correctly predicts *blue* to be the most frequent overextension, but underestimates the total amount of errors (Fig. 1b).

The final phenomenon concerns the asymmetry in overextensions between PURPLE and BLUE. Whereas *blue* is overextended to the PURPLE stimulus, *purple* is not overextended to BLUE. We can rule out the frequency difference between *blue* and *purple* as an explanation, despite that *purple* is much less frequent: Under both frequency settings, *purple* is not overextended to BLUE. Given that the conceptual features do not help the model fit, it is likely that the source of the asymmetry is to be found in the perceptual feature space.

Looking more closely at the color stimuli and the perceptual feature space, we can identify that the reason for the observed asymmetry is the location of the focal colors within each color category. As Figure 2 shows, the BLUE and PURPLE categories form a sphere in the three perceptual dimensions. The focal exemplars of each category, however, are located at different values for L^*, the luminance dimension. Focal PURPLE is darker than focal BLUE, and hence closer (on the dimensions a^* and b^*) to BLUE exemplars with a lower luminance. Focal BLUE is more luminant, and hence further away from PURPLE exemplars with the same luminance.

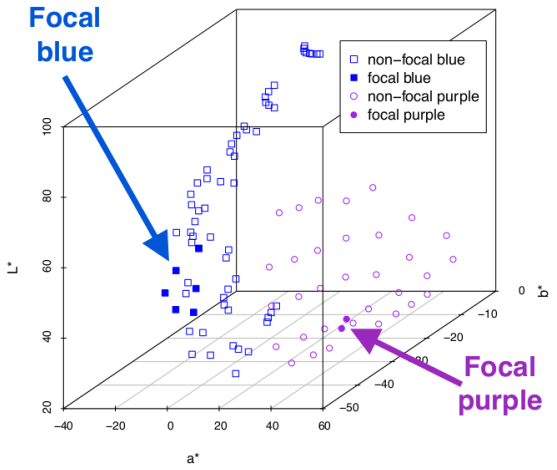
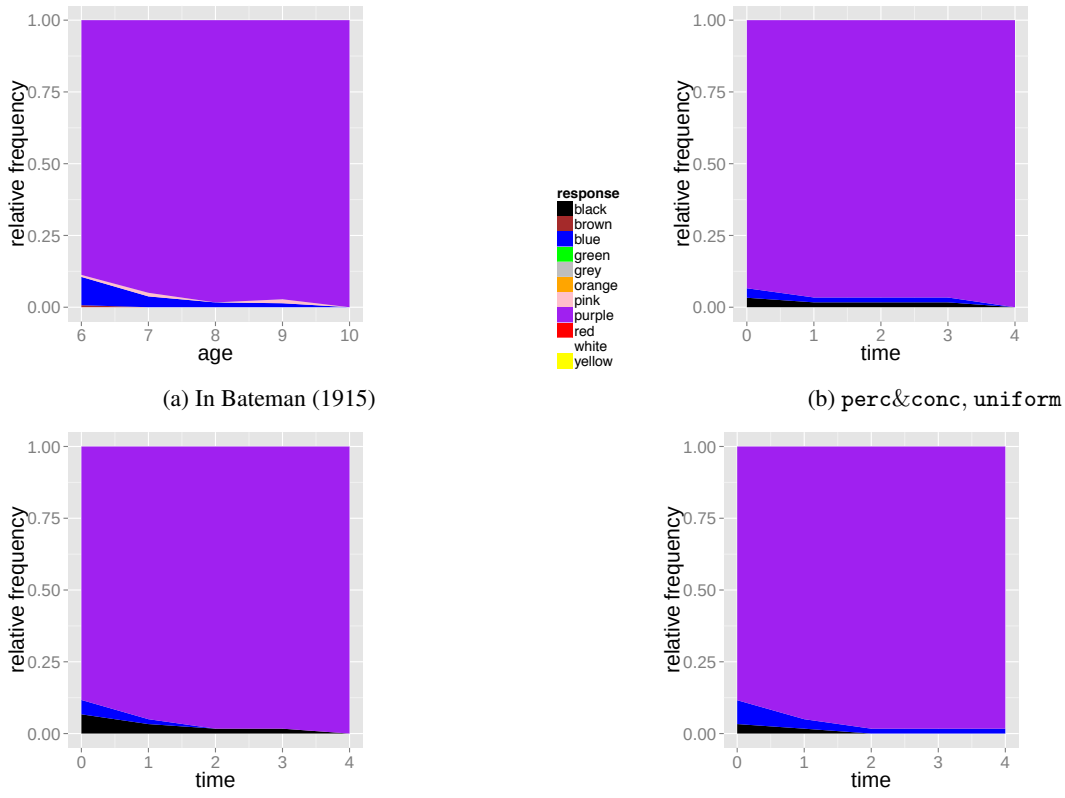


Figure 2: Positions of the various BLUE and PURPLE exemplars in the CIELab space.

On the assumption that Bateman's test items were focal exemplars of the categories, this means that the lack of overextension of *purple* to BLUE can be attributed to the lay-out of the perceptual dimensions, and to the position that the focal exemplars have in that space. Thus, the model's results suggest a new explanation for the asymmetry in overextensions that goes beyond simple perceptual closeness and frequency of color terms.

5.3 The role of the conceptual features

If the conceptual dimensions have little additional predictive power over the perceptual ones, two



(a) In Bateman (1915)

(b) `perc&conc`, `uniform`

(c) `perc`, `uniform`

(d) `perc`, `relative`

Figure 1: Observed and predicted responses to PURPLE over time.

	L*	a*	b*
PCA1	−0.01	0.80*	−0.01
PCA2	−0.97***	0.40	−0.08
PCA3	0.16	−0.03	−0.88**
PCA4	0.60	−0.86*	0.70

Table 3: Correlation matrix for the four used PCA components and the three perceptual dimensions. Stars indicate level of significance of the correlation (* = $p < .05$, ** = $p < .01$, *** = $p < .001$).

scenarios are possible. The conceptual dimensions may correlate with the perceptual ones, or they may be independent from them. In the former case, it means that the crosslinguistic commonalities in structuring the domain of color mirror the perceptual biases. This would mean that adding the conceptual dimensions can be expected to have no explanatory effect on top of the perceptual dimensions. In the latter scenario, it means that there are other biases causing the commonalities in the crosslinguistic data, but that these biases do not affect language acquisition. This scenario would imply a negative assessment of the Typological Prevalence Hypothesis.

As we can see in Table 3, the former scenario of correlated features seems closer to the truth than the latter. The luminance dimension L^* displays an almost perfect negative correlation with component 2 of the PCA, whereas the red-green scale a^* has a strong positive correlation with component 1 and a strong negative one with component 4. The yellow-blue scale, finally, has a strong negative correlation with component 3. That is: all four features of our `conc` space (i.e., those PCA components with Eigenvalues greater than 1) have correlating perceptual dimensions. This means that they can be seen as symptoms of these dimensions and that the category structure of color terms across languages depends to a large extent on the perceptual dimensions of color.

What this means is that using crosslinguistic data does lay bare an important part of the conceptual structure of the domain. If we did not know of the perceptual properties of color, a Principal Component Analysis on the basis of crosslinguistic data would provide us with an insight in all three dimensions of the perceptual space.

One concern remains, however. Even though

the perceptual feature space by itself constitutes a good predictor of the error terms, the use of only conceptual dimensions does not explain as much of the error pattern.

6 Conclusion

In this paper, we looked at overextensions in the acquisition of the meaning of color terms. For this initial study, we focused on the English data of Bateman (1915) – the most comprehensive published error data on color terms – in which we identified five phenomena that characterize the pattern of children's errors, and that must be explained by a theory of word meaning acquisition. We considered three factors that might play a role in this domain: (1) the identified perceptual dimensions relating to the various exemplars of the color terms; (2) the effect of typological prevalence (i.e., the more frequently a certain grouping of color exemplars is crosslinguistically, the more cognitively natural it is thought to be, and hence the more readily/robustly acquirable, Gentner and Bowerman (2009)); and (3) the frequency of color terms.

We used an extension of the modeling approach taken in Beekhuizen et al. (2014). In that work, the effects of typological prevalence and frequency were studied in the domain of spatial relations. In this paper, we applied the same technique to the crosslinguistic elicitation data of the World Color Survey (Kay et al., 2009) to arrive at a set of features (the 'conceptual' space) reflecting typological frequency of semantic groupings. We considered in addition the possible impact of a perceptual representation of color.

We find several notable effects within our setup. First, the perceptual influence provides the best explanation of the errors: Including the perceptual features gives the model a very good fit with the developmental overextension pattern for all five phenomena observed in the Bateman data, and adding either or both of the conceptual (typological) features and the frequency information does not improve the fit. This last finding is revealing, as it means that the overextensions cannot be ascribed to the frequencies of the color terms.

We argued that the reason the conceptual features do not improve the model fit is that the perceptual and conceptual spaces are strongly correlated. This suggests that the typological prevalence patterns in the crosslinguistic data follow the perceptual dimensions. However, the model fit is actually worse when only using the conceptual features, an issue that we must explore further.

Furthermore, it may be that the conceptual features do help for the acquisition of color words in other languages. The lack of an effect of the conceptual space on top of the perceptual features may also be due to the (older) age of the children in the data. Overextension patterns in younger children may display effects of the conceptual dimensions, as well as frequency. We are currently planning to extend this research to a variety of error data sets, both in English and other languages, to see if similar results are found and to further evaluate the role of the various perceptual, typological, and frequency factors.

Another issue we plan to work on is the fact that the model performs 'too well': It predicts no overextensions for 6 out of the 8 color stimuli, despite children displaying a few errors on 4 of these colors. Using our typologically-derived semantic space within a fuller model of word learning, such as that of Fazly et al. (2010) or Nematzadeh et al. (2012), rather than using a simple categorization model as we do here, might further our insight into potential sources of overextensions.

Given our general methodological approach, reviewers noted other interesting possibilities and suggested that alternative design choices are possible as well for the dimensionality reduction technique, the alignment method between predicted model data and observed experimental data, and the statistical evaluation procedure. We plan to follow up on these suggestions in future research, in addition to the exploration of a wider set of crosslinguistic error patterns, the consideration of earlier developmental stages, and the use of a more realistic word-learning model.

Acknowledgments

We gratefully acknowledge NSERC of Canada for the funding of both authors, as well as the four anonymous reviewers for their comments and suggestions.

References

Elsa Jaffe Bartlett. 1978. The acquisition of the meaning of colour ters: a study of lexcal development. pages 89–108.

W. G. Bateman. 1915. The Naming of Colors by

Children the Binet Test. *The Pedagogical Seminary*, 22(4):469–486, December.

Barend Beekhuizen, Afsaneh Fazly, and Suzanne Stevenson. 2014. Learning Meaning without Primitives: Typology Predicts Developmental Patterns. In *Proceedings of the 36th annual meeting of the Cognitive Science Society*.

Brent Berlin and Paul Kay. 1969. *Basic color terms: Their universality and evolution*. University of California Press, Berkeley, CA.

Melissa Bowerman. 1993. Typological perspectives on language acquisition: Do crosslinguistic patterns predict development? In Eve V. Clark, editor, *Proceedings of the Twenty-fifth Annual Child Language Research Forum*, pages 7–15, Stanford, CA. CSLI Publications.

Eve V. Clark. 1973. What's in a word? On the child's acquisition of semantics in his first language. In Timothy E. Moore, editor, *Cognitive Development and the Acquisition of Language*, pages 65–110. New York: Academic Press.

Ian R. L. Davies, Greville Corbett, Harry McGurk, and David Jerrett. 1994. A developmental study of he acquisition of colour terms in Setswana. *Journal of Child Language*, 21:693–712.

Ian R. L. Davies, Greville Corbett, Harry McGurk, and Catriona MacDermid. 1998. A developmental study of the acquisition of Russian colous terms. *Journal of Child Language*, 25:395–417.

Afsaneh Fazly, Afra Alishahi, and Suzanne Stevenson. 2010. A probabilistic computational model of cross-situational word learning. *Cognitive Science*, 34(6):1017–1063.

Dedre Gentner and Melissa Bowerman. 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, and S. Ozcaliskan, editors, *Crosslinguistic approaches to the psychology of language. Research in the tradition of Dan Isaac Slobin*, chapter 34, pages 465–480. Psychology Press, New York, NY.

Dedre Gentner. 1982. Why Nouns are Learned Before Verbs : Linguistic Relativity versus Natural Partitioning. In Stan Kuczaj, editor, *Language Development. Volume 2: Language, Thought, and Culture*, volume 2, chapter 11, pages 301–334. Lawrence Erlbaum Associates, Hillsale, New Jersey.

Judith C Goodman, Philip S Dale, and Ping Li. 2008. Does frequency count? Parental input and the acquisition of vocabulary. *Journal of child language*, 35(3):515–31, August.

Insa Gülzow and Natalia Gagarina, editors. 2007. *Frequency Effects in Language Acquisition. Defining the Limits of Frequency as an Explanatory Concept*. De Gruyter Mouton, Berlin.

Sara Harkness. 1973. Universal Aspects of Learning Color Codes: A Study in Two Cultures. *Ethos*, pages 175–200.

H. Hotelling. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441,498–520.

Z.M. Istomina. 1960. Perception and naming of color in early childhood. *Izvestiia Akademii Pedagogicheskikh*, 113:37–45.

Paul Kay, Brent Berlin, Luisa Maffi, William R. Merrifield, and Richard Cook. 2009. *World Color Survey*. CSLI Publications, Stanford, CA.

Stephen C. Levinson, Sergio Meira, The Language Group, and Cognition. 2003. 'Natural Concepts' in the Spatial Topological Domain – Adpositional Meanings in Crosslinguistic Perspective: An Exercise in Semantic Typology. *Language*, 79(3):485–516.

Asifa Majid, Fiona Jordan, and Michael Dunn. 2015. Semantic systems in closely related languages. *Language Sciences*, 49:1–18.

Aida Nematzadeh, Afsaneh Fazly, and Suzanne Stevenson. 2012. A computational model of memory, attention, and word learning. In *Proceedings of the Third Workshop on Cognitive Modeling and Computational Linguistics*.

Robert M Nosofsky. 1987. Attention and Learning Processes in the Identification and Categorization of Integral Stimuli. *Journal of Experimental Psychology*, 13(1):87–108.

Nicola J. Pitchford and Kathy J. Mullen. 2003. The development of conceptual colour categories in preschool children: Influence of perceptual organization. *Visual Cognition*, 10(1):51–57.

Terry Regier, Paul Kay, and Naveen Khetarpal. 2007. Color naming reflects optimal partitions of color space. *PNAS*, 104:1436–1441.

Debi Roberson, Jules Davidoff, Ian R L Davies, and Laura R Shapiro. 2004. The development of color categories in two languages: a longitudinal study. *Journal of experimental psychology. General*, 133(4):554–71, December.

Nancy N. Soja. 1994. Young Children's Concept of Color and Its Relation to the Acquisition of Color Words. *Child Development*, 65:918–937.

Anna L. Theakston, Elena V.M. Lieven, Julian M. Pine, and Caroline M. Rowland. 2001. The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Languages*, pages 127–152.

Daniel Yurovsky, Katie Wagner, David Barner, and Michael C. Frank. 2015. Signatures of Domain-General Categorization Mechanisms in Color Word Learning. In *Proceedings CogSci*.