Introduction

Languages vary in how they carve up lexical semantic domains; e.g., for situations where English uses the preposition on, Dutch uses aan and op. The crosslinguistic variation in lexical semantic divisions raises two interesting questions regarding the language user. First, are these various divisions all equally easy to learn, and if not, what drives the difference in ease of acquisition? Second, does acquiring a language-specific system affect other parts of cognition, a position known as ‘linguistic relativity’ (Gumperz & Levinson, 1996)? The two questions are related, as the acquisition of language-specific semantic divisions can be expected to go hand in hand with any extra-linguistic effects of acquiring such systems.

In this paper, we use a computational model to study both the acquisition of color terms and behavior on a non-verbal color discrimination task. Our goal is to propose a unified account of these two phenomena by simulating them within a single computational word-learning model. We explore what factors drive the two phenomena, considering both the features with which we represent color in the model and the varying frequencies of the color terms.

As there is an understanding of how color is represented on a perceptual level (e.g., Fairchild, 1998), we can use perceptual features as one representation of color. We also explore features motivated by the Typological Prevalence Hypothesis, which holds that crosslinguistically more common divisions are more cognitively accessible and thus easier to learn (Gentner & Bowerman, 2009). This exploration is motivated by positive results of this approach in the domain of spatial adpositions (Beekhuizen, Fazly, & Stevenson, 2014). Specifically, we use a representation based on the crosslinguistic biases in the divisions of the color space, derived from elicitation data (Kay, Berlin, Maffi, Merrifield, & Cook, 2009).

We find that both types of features yield a good fit to the developmental pattern of color term acquisition as well as to the behavioral pattern, with term frequency having an impact on some results. In some cases, we find that using the features together produces a better fit to human data, indicating that both perceptual properties and crosslinguistic biases may play a complementary role in learning a system of color terms.
Related modeling work To our knowledge, the only other attempt at modeling linguistic relativity is Colunga and Gasser (1998), who train a neural network on artificial languages and semantic domains to study both the effects of ease of acquisition and cognitive consequences of acquiring semantic divisions. Our model has a similar architecture and displays similar effects, but is trained on naturalistic data.

An earlier attempt at modelling color term acquisition is Belpaeme and Bleys (2005), who present a multi-agent model that represents color in an \(L^aL^bL^c\) space (see below), although they do not focus on the developmental trajectories of learners or on behavioral linguistic relativity effects. Our approach can be considered as complementary, focusing on the cognition and behavior of an individual learner rather than on biases in the emergence of community-wide systems.

Beekhuizen and Stevenson (2015) used the Generalized Context Model (GCM; Nosofsky, 1987) to simulate the developmental English color naming data of Bateman (1915). While this approach showed interesting preliminary results, GCM is limited in its ability to acquire language-specific attention weighting. We require a model able to incrementally acquire and represent varying attentional weights over sub-intervals of the values of a dimension, possibly independently of values on other dimensions. The Russian ‘blue’s are a case where such representational potential is needed: attention to a part of the luminance scale is heightened, but only for blue hues. Self-Organizing Maps (SOMs; Kohonen, Schroeder, & Huang, 2001), explored for language acquisition by, e.g., Li and Zhao (2013), constitute a class of models that can capture such effects, while also having the potential to show developmental effects due to their incremental nature.

Our Computational Model

Self-Organizing Map

A Self-Organizing Map \(M\) is a neural network consisting of an \(m \times n\) grid of neuron cells \([c_{11}, c_{12}, \ldots, c_{mn}]\), where every cell consists of a vector of feature values. At every iteration \(i\) of training, an input stimulus \(s\), with values for the same set of features, is compared to all cells \(c \in M\), and is subsequently mapped to the cell to which it is most similar, called the Best Matching Unit (BMU) cell for \(s\), or \(c_s\). The values of \(c_s\) as well as its neighboring cells are then updated with the values of \(s\). This way, \(M\) will come to display a topology that reflects the similarity among the input items.

Formally, \(c_s = \arg\min_{c \in M} d_{lead}(c, s)\) where \(d_{lead}(c, s)\) is the Euclidean distance between the feature values of \(c\) and \(s\). All cells are updated in proportion to their map distance from \(c_s\):

\[
c_{jk}^{i+1} = c_{jk}^i + h_{jk}^i \times (s - c_{jk}^i)
\]

\[
h_{jk}^i = \alpha \cdot \exp \left(-\frac{d_{map}(c_{jk}, c_s)}{2 \times \sigma_i^2} \right)
\]

That is, \(h_{jk}^i\) yields the excitation of the neuron cell \(c_{jk}\) given a center of activation at the coordinates of \(c_s\), taking into account their distance in the map grid given by \(d_{map}\). Here \(\alpha = [0, 1]\) is a learning rate parameter, and \(\sigma_i\) the neighborhood radius of \(c_s\), given by the exponential function \(\sigma_i = \sigma_0 \times \exp\left(-\frac{t}{\lambda_0}\right)\); \(\sigma_0\) and \(\lambda_0\) are constants defining the intercept and slope of the function yielding the neighborhood radius. To observe developmental effects, slow learning is needed, and therefore we set \(\alpha = .05\), \(\sigma_0 = 1\), \(\lambda_0 = 2000\), and train \(8 \times 8\) maps.

Feature Representations

We formulate acquisition of color vocabulary as a categorization task that associates a color term (category label) with a color stimulus (a set of color property features). An input item consists of a representation of the properties of a Munsell color chip (a property-feature vector) paired with a color term (a term-feature vector). Each cell of the SOM represents a learned association between a set of property-feature values and a distribution over the terms in the term-feature vector.

The term-feature vector has length \(|T|\), where \(T\) is the set of primary color terms in a language. To represent term \(t\) in an input item, the \(t\)th feature is set to a value \(a\) in \([0, 1]\), and all others set to 0; e.g., in a system with 4 terms, input \(t_2 = [0, a, 0, 0]\). The parameter \(a\) (in our experiments set to .2) reflects the relative importance of term features in training. The term-feature vector of each cell of the SOM will come to hold a distribution over terms, which we normalize to arrive at a probability of a term for a cell, \(P(t|c)\) (see below).

The property-feature vector represents the set of stimuli of Munsell color chips, \(S\), in one of two forms. First, we test the idea that the cross-linguistic tendencies in the semantic distinctions are telling of the extra- or pre-linguistic cognitive biases of language learners (cf. the Typological Prevalence Hypothesis). As in Beekhuizen and Stevenson (2015), we operationalize this idea with Principal Component Analysis (PCA) over the World Color Survey data (Kay et al., 2009), which contains color terms for 330 Munsell color chips in 110 languages. The closeness of a pair of chips in the resulting space reflects the frequency with which they are labeled with the same term, and the space thereby represents the cross-linguistic tendencies to group chips under a particular term. (More details can be found in Beekhuizen & Stevenson, 2015.) If the extension of a color term – i.e., the set of chips labeled with that term – is spread widely over the PCA space, it is assumed to be harder to learn than if a set of the same size were spread less widely over the PCA space. We refer to property features based on the PCA components as the conceptual, or conc, features.

We also can represent the various color chips at a purely perceptual level. We use the coordinates of the chips in \(L^aL^bL^c\) space, which is thought to encode the perceptual dissimilarity between colors (Fairchild, 1998). This feature set will be referred to as the perceptual, or perc, features.

Both the property-feature spaces were normalized such that the mean for each feature is .5 and the values are in \([0, 1]\). SOMs are initialized with values of 0 for term features and \(.5 \pm \) a very small random value for property features.
Sampling for training data

Input items are sampled as a pair of a color term $t \in T$ and a stimulus color chip $s \in S$ from the distribution $P(t,s) = P(s|t)P(t)$. We obtain the conditional probability distributions for $P(s|t)$ from adult elicitation data in English (Berlin & Kay, 1969) and in Russian (Davies & Corbett, 1994). For the latter data in Yxy coordinates, we convert those coordinates into $L^*a^*b^*$, and identify the Munsell chip with the closest $L^*a^*b^*$ value.

As one estimation of $P(t)$, we used the relative term frequency over all color terms. For English, these were taken from the child-directed speech portion of the Manchester corpus (Theakston, Lieven, Pine, & Rowland, 2001) of CHILDES (MacWhinney, 2000). For Russian, lacking a corpus of child-directed speech of suitable size, we use the relative term frequencies reported in Vamling (1986). We also assess sampling according to a uniform distribution for $P(t)$.

These two conditions are called corpus and uniform.

Experimental Methods

For all experiments, we run 30 simulations for each of the six combinations of features=\{perc, conc, perc+conc\} and sampling=\{corpus, uniform\}. At every test moment (every 100 input items), we present the model with an unlabelled color chip $s$ (i.e., a property-feature vector with no term features) and extract the most probable term that the model associates with those property features. We obtain the Best Matching Unit for $s$ as $c_s = \arg\min_{c \in M} d_{\text{leaf}}(c,s)$, where only the property features of $c$ are compared to those of $s$. The model response for $s$, term $t_s$, is extracted from the probability distribution over the terms $T$ for $c_s$:

$$t_s = \arg\max_{t \in T} P(t|c_s)$$

(3)

$$P(t|c_s) = \frac{\text{value}(t,c_s)}{\sum'_{t' \in T} \text{value}(t',c_s)}$$

(4)

where value$(t,c_s)$ is the value for feature $t$ in cell $c_s$.

Evaluating linguistic convergence

To evaluate whether the model obtains an adult level of understanding of the color terms, we test it with color stimuli corresponding to the complete set of color chips $S_{\text{adult}}$ for which we have adult responses (|$S_{\text{adult}}$| = 49, |$T$| = 12 for Russian; |$S_{\text{adult}}$| = 211, |$T$| = 11 for English). Model convergence with adult linguistic behavior is then given by:

$$\text{score}_C = \frac{|S_{\text{correct}}|}{|S_{\text{adult}}|}$$

(5)

1This formulation of $P(s|t)$ is informative about the mapping of terms to colors: a chip $s_1$ labeled half the time as blue is less likely to be sampled for the term blue than a chip $s_2$ labeled 100% of the time as blue. However, $P(s|t)$ says nothing about how frequently the colors are discussed with that label: if $s_1$ is more frequent in the world, usages of blue may refer to it more than to $s_2$. At this point we know of no way to estimate a sampling of colors people refer to.

The different sources of frequency data may differentially affect outcomes in the two languages, an issue for future research.

Evaluating linguistic development

In the child color naming data, several types of patterns are observed: For some color stimuli, children produce hardly any or no errors, whereas for others, overextensions are observed, sometimes even more frequently than the correct term. Our goal is to assess the fit between the model’s distribution over terms, $P(t|c_s)$ (Eqn. 4), for each stimulus $s$ at various points in learning, and the relative dominance of terms exhibited by children at various points in development.

To that end, we compare the ranking of terms based on $P(t|c_s)$ to a ranking derived from child elicitation data (ranked by the number of children producing an error for that color in Bateman, 1915 and Davies et al., 1998). For every color stimulus presented to children from $n$ age groups, we find the $n$ consecutive, equal-sized bins of test moments for which the predicted ranking for that stimulus matches optimally the observed ranking of each age group for that color. Each bin contains at least 5 test moments, to avoid finding unrealistically narrow ‘age groups’ in the model data. The model ranking of terms is given by $P(t|c_s)$ averaged over all test moments in that bin, across 30 simulations. The low values in this pooled probability distribution ($P(t|c_s) < .05$) are rounded down to 0 to avoid diluting the ranking metric with insignificant predictions; similarly, we consider only errors occurring a minimum of 3 times in the child data. The two – model and observed – rankings are then compared using Kendall’s $\tau_b$, which we use as our evaluation measure.

Evaluating color discrimination

We take the final state of the SOM to correspond to adult organization of the color terms. Reflecting the hypothesis that linguistic knowledge affects the extra-linguistic task of color discrimination, we take the closeness between the BMUs of two stimuli in our learned SOM to correspond to the degree of difficulty people show in discriminating them. We convert the 20 stimuli of W07 into our representation of color properties, yielding the vector $S_{\text{disc}} = [s_1, \ldots, s_{20}]$. Following W07, we consider two stimuli $s_i$ and $s_j$ to be ‘near’ if $j - i + 2$, and ‘far’ if $j = i + 4$. To find the perceived distance between the target and distracter, $s_i$ and $s_j$, we take their SOM distance $d_{\text{map}}(c_i, c_j)$ (as defined above). The greater the distance, the easier to discriminate the target from the distracter.

We find the category boundary in the model by having it predict the most likely Russian term per stimulus in $S_{\text{disc}}$, and placing the boundary between the last light blue (goluboj)
response and the first dark blue (sinij) response. English does not have these distinct terms, but the observed category boundaries for Russian and English hardly differ according to W07. We thus use as the English boundary the mean location of the Russian category boundary under the given combination of features $\times$ sampling. A target–distracter pair, $s_1$–$s_2$, is considered ‘within’-category if $s_1$ and $s_2$ are on the same side of the boundary, and ‘across’-category otherwise. Analogously to W07, the map distances for the 8 ‘near’ and 8 ‘far’ pairs closest to the category boundary were calculated from the model for all simulations.

We evaluate whether the model’s behavior corresponds to human behavior in W07 by seeing if the same significant effects are found: we compare the $d_{map}(c_x,c_y)$ values (using $t$-tests) between near and far cases, and between within- and across-category cases, and see whether these two interact.

**Results: convergence and development**

The model reaches its closest fit to adult behavior after some 20K (Russian) or 30K (English) input items. Table 1 shows that the model captures adult behavior well; a naive baseline always guessing the most frequent term would reach a score $C$ of .20 (English) or .22 (Russian). We find that uniform sampling achieves slightly closer to adult naming behavior. With

<table>
<thead>
<tr>
<th></th>
<th>Russian</th>
<th></th>
<th>English</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>corpus</td>
<td>uniform</td>
<td>corpus</td>
<td>uniform</td>
</tr>
<tr>
<td>perc</td>
<td>.84 (.04)</td>
<td>.89 (.03)</td>
<td>.91 (.03)</td>
<td>.93 (.02)</td>
</tr>
<tr>
<td>conc</td>
<td>.86 (.03)</td>
<td>.87 (.02)</td>
<td>.92 (.02)</td>
<td>.93 (.02)</td>
</tr>
<tr>
<td>perc+conc</td>
<td>.89 (.03)</td>
<td>.92 (.03)</td>
<td>.95 (.02)</td>
<td>.97 (.01)</td>
</tr>
</tbody>
</table>

Table 1: Results for convergence: mean and standard deviation of $score_C$ (Eqn. 5), over 30 simulations.

For English, the two perc settings give a better fit than the conc settings. Considering the match with children’s development on particular colors helps understand why. For English, Bateman (1915) presented children (age 6–12) with 8 color chips. The model only displays the correct overextensions of blue to purple with (perc+conc, corpus) and (perc, corpus), and fails to simulate the correct pattern for orange in both conc settings. The other color terms were learned with the correct developmental pattern under all settings: For black, white, red, and blue, no or hardly any overextensions were found either in children or in the model, and the few observed overextensions for yellow and green were predicted in any parameter setting.

For Russian, we observe a difference between corpus and uniform sampling. Davies et al. (1998) presented 3- to 5-year-olds with 12 color chips. 3-year-olds label light blue and purple more frequently sinij ‘dark blue’ than the correct terms, but do not label dark blue as goluboj ‘light blue’ or fioletovyj ‘purple’ as often. Furthermore, 3-year-olds more frequently use krasnyj ‘red’ than rozovyj ‘pink’ for pink. The model predicts these effects completely under
Importantly, the model matches the main finding of W07 that learned SOM is a good proxy for discriminability of stimuli. Languages support our assumption that map distance in the and for 5 of 6 settings in English.

For example, the entry for Russian of “near-within discrimination task (1st column), along with an indication of whether the model results match those findings (2nd column). The asymmetry for DARK BLUE and LIGHT BLUE disappears when we train on perc, as DARK BLUE and LIGHT BLUE are (too) easily discriminated in the perceptual space. This is different for the conceptual features: as many languages group DARK BLUE and LIGHT BLUE under one term, the inferred cognitive bias is to group them together. Figures 1e-f show that for conc, the asymmetry is present, but goluboj never gets fully learned. A combination of both feature sets thus seems necessary to understand this effect: perceptual dissimilarity is needed to discriminate them, but cognitive biases bias the learner against forming two categories. The asymmetry may then emerge because of the slightly higher term frequency of sinij (.08) over goluboj (.06).

Results: discrimination

Figure 2a summarizes the findings of W07 in their color discrimination task (1st column), along with an indication of whether the model results match those findings (2nd column). For example, the entry for Russian of “near-within > near-across” means that people found the near-within cases harder to discriminate than the far-within cases (a statistically-significant difference); “far-within ≈ far-across” means the difference between those two cases for people was not statistically significant. For the former, our model also found a statistically significant effect in the same direction, and for the latter, the lack of an effect in the same direction.\(^3\)

The fact that the model matches “near > far” for both languages supports our assumption that map distance in the learned SOM is a good proxy for discriminability of stimuli. Importantly, the model matches the main finding of W07 that distracters in a different category from the target are more easily discriminated than distracters in the same category, for Russian but not for English (the “within > across” and “within ≈ across” rows in Figure 2a).

To illustrate why this happens, Figure 2b presents a typical converged map for English and for Russian. For English, chips \(s_1:s_4\) are mapped to the cell marked with (2), and chips \(s_5:s_{20}\) to the cell marked with (1). Because the category boundary is placed between \(s_{11}\) and \(s_{12}\) of \(S_{\text{disc}}\), all pairs of targets and distracters are mapped to the same cell (1), whether across-category or within, and such pairs are indiscriminable for the learner. For Russian, the different shades of blue cannot be compressed on the SOM as much as in English, because there are two terms that need to be discriminated: English blue is the most likely term in 5 cells, whereas Russian sinij and goluboj combined are the most likely terms in 10 cells. Thus in Russian, we see that the 20 \(S_{\text{disc}}\) stimuli are mapped to a larger part of the SOM (cells (1)-(4) in that map) than the English stimuli, and distances across the categories – from cells (1)-(2) to (3)-(4) – are further than within categories (within (1)-(2) or within (3)-(4)).

Finally, the model generally fails to predict the empirical interaction whereby Russian displays a significant within-across difference for near but not far cases. Under all settings, the model predicts both differences to be significant. We do find a trend in the right direction: for all settings, the within-across difference is greater in the model for the near cases than for the far cases.

Discussion

In this paper, we looked at the developmental pathway of color term acquisition and the effects of acquiring the color term system of a particular language on a non-verbal discrimination task. A Self-Organizing Map (SOM) trained on naturalistic input models three effects: (1) some patterns of overextension errors in linguistic development and (2) subsequent convergence in Russian and English, as well as (3) a higher ability to discriminate light blue from dark blue stimuli in Russian, but not English. Our model thus provides a
mechanistic conception of learning that gives a unified explanation of both linguistic development and linguistic relativity. The idea that between-language variation is represented by the varying amount of information compression on the SOM (due to the different patterns of words with stimuli across languages) gives us an explanatory principle that could be applied to domains beyond color.

We asked whether possible cognitive biases inferred from crosslinguistic categorization tendencies (cf. Gentner & Bowerman, 2009, reflected in our ‘conceptual features’, play a role, or whether perceptual features of color best explain the effects. Both feature sets contribute to the explanation of linguistic development: in some cases (naming PURPLE in English), the error pattern is predicted only when the perceptual features are present. For others, leaving out the conceptual features hurts the fit with the observed data (naming DARK BLUE in Russian), suggesting that these biases do play a role.

We also investigated frequency effects: The model fails to predict common overextension patterns in both languages when not taking term frequency into account. Nonetheless, sampling on the basis of corpus frequencies makes the model converge less well to adult behavior for infrequent terms, suggesting that, over development, learners may need to be increasingly sensitive to term frequency.

One issue we did not explore is different initializations of the SOMs. As children experience color prior to acquiring terms for them, it is possible that the map is already ‘pre-organized’ by exposure to color stimuli without associated color terms. We plan on studying further whether such prelinguistic exposure affects the developmental patterns.

Finally, we looked at the converged states of the SOMs in predicting color discrimination behavior across languages, finding a weak preference for models trained on perceptual features. Since we are able to track the development of the SOM, we can also investigate the effect of language-specific lexical semantic systems on extra-linguistic behavior over developmental time (see, e.g., McDonough, Choi, & Mandler, 2003, for such developmental effects in another domain). In the future, we plan to explore suitable semantic domains for evaluating how well our model simulates linguistic relativity effects over the course of acquisition.

Acknowledgements

We acknowledge the support of NSERC of Canada and thank the four CogSci reviewers for helpful suggestions.

References


