This is the accepted version of the following article: Barend Beekhuizen and Suzanne Stevenson (2018). 'More than the eye can see: A computational model of color term acquisition and color discrimination'. *Cognitive Science* 42: 2699–2734, which has been published in final form at https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.12665.

More than the eye can see: A computational model of color term acquisition and color discrimination

Barend Beekhuizen and Suzanne Stevenson

Department of Computer Science, University of Toronto

Abstract

We explore the following two cognitive questions regarding crosslinguistic variation in lexical semantic systems: Why are some linguistic categories – i.e., the associations between a term and a portion of the semantic space – harder to learn than others? How does learning a language-specific set of lexical categories affect processing in that semantic domain? Using a computational word-learner, and the domain of color as a testbed, we investigate these questions by modeling both child acquisition of color terms and adult behavior on a non-verbal color discrimination task. A further goal is to test an approach to lexical semantic representation based on the principle that the more languages label any two situations with the same word, the more conceptually similar those two situations are. We compare such a crosslinguistically-based semantic space to one based on perceptual similarity. Our computational model suggests a mechanistic explanation for the interplay between term frequency and the semantic closeness of learned categories in developmental error patterns for color terms. Our model also indicates how linguistic relativity effects could arise from an acquisition mechanism that yields language-specific topologies for the same semantic domain. Moreover, we find that the crosslinguistically-inspired semantic space supports these results at least as well as – and in some aspects better than – the purely perceptual one, thus confirming our approach as a practical and principled method for lexical semantic representation in cognitive modeling.

More than the eye can see: A computational model of color term acquisition and color discrimination

Introduction

Languages vary in how they carve up lexical semantic domains: that is, sets of terms in different languages frequently differ in the mapping of the terms to parts of the semantic space. To give a few examples, whereas English describes all spatial relations of SUPPORT with *on*, Dutch makes use of *aan* (TENUOUS SUPPORT) and *op* (STABLE SUPPORT).[1] For interclausal coherence relations covered by *and* and *but* in English, Tuvaluan relies on a single conjunction *kae* (Mauri, 2008). More complex, many-to-many mappings between language-specific categories are also attested (e.g., Berlin & Kay, 1969; Kay, Berlin, Maffi, Merrifield, & Cook, 2009; Malt, Sloman, & Gennari, 1999). For instance, Fig. 1 shows how four languages carve up a small region of the color space, and illustrates many-to-many mappings between the language-specific categories (Kay et al., 2009). These and numerous other such examples suggest that the vocabulary of a domain establishes a set of language-specific semantic categories over the domain; consequently, much research has investigated how the structure of lexical categorization differs across languages.

The crosslinguistic variation in lexical categorization raises important cognitive questions. One such question concerns how these various lexical semantic categories are acquired: Are they all equally easy to learn, and if not, what drives the difference in ease of acquisition? The semantic category boundaries across languages seem to vary widely, but are also characterized by certain commonalities. Such crosslinguistic commonalities have been taken to reflect underlying conceptual similarity: i.e., if there are many languages that use a single term for two semantic situations, those situations are likely very similar (Anderson, 1982). Taking this insight one step further, the Typological Prevalence Hypothesis proposes that crosslinguistically more-prevalent semantic categories are easier to learn than less-prevalent categories (Bowerman, 1993; Gentner & Bowerman, 2009). In short, crosslinguistic patterns

[1] Throughout the paper, we use italics to indicate terms in a language (e.g., *red*), small caps to refer to semantic categories (e.g., RED), and single quotes to provide glosses in English (e.g., *krasnyj* 'red').

in lexical semantic categorization appear to reflect cognitive biases that influence the acquisition of the terms in a language. On the other hand, for the domain of color, it has been argued that basic perceptual properties suffice to explain the relative ease of learning the various mappings between color terms and perceptual stimuli (Dougherty, 1978; Harkness, 1973; Pitchford & Mullen, 2003). This raises the issue of whether crosslinguistic analysis in this domain captures relevant biases that go beyond basic color perception. In this paper, we focus on modeling a particular aspect of color term acquisition – namely, children's overextensions of color words to stimuli for which adult speakers would use other terms – to explore how these factors play a role in developmental error patterns.

A second important question concerns the nature of the cognitive effects of acquiring a language-specific set of lexical semantic categories: Does acquiring such a system influence cognitive processing in that domain (a position widely attributed to Whorf 1956), or is the access/use of concepts unaffected by the system of words used to refer to them? Over the last 20 years significant evidence has accumulated supporting the notion of linguistic relativity – i.e., attesting that the semantic categories reflected in the vocabulary can influence conceptual categorization and/or access (e.g., Everett 2013; Gumperz and Levinson 1996; McDonough, Choi, and Mandler 2003; though see Pinker 2007, for an alternative view). In the domain of color in particular, the early study of Brown and Lenneberg (1954) showed that a measure of color-naming agreement across informants predicted color recognition behavior better than a purely perceptual measure of color distinctiveness. More recently, a number of studies have shown various influences of linguistic color categories on the processing of color stimuli (Bae, Olkkonen, Allred, & Flombaum, 2015; Roberson, Davidoff, Davies, & Shapiro, 2005; Roberson, Davies, & Davidoff, 2000; Winawer et al., 2007). Here, we model empirical data on linguistic relativity in the lexical semantic domain of color, and again consider the factors that might underlie the observed behavioral effects.

To investigate the above two questions, we use a computational word-learner to model both child acquisition of color terms and adult behavior on a non-verbal color discrimination task. Specifically, we explore the factors underlying both the observed developmental errors in

children, and the subsequent linguistic relativity effects found when adults draw on their learned knowledge. By studying both learning and use of knowledge in one computational model, we aim to give a unifying explanation of the factors that influence the two.

Our primary focus is a comparison of two different semantic representations of the color domain, to see which best captures human behavior in our computational model. The first is a perceptual semantics – i.e., one based on the properties of human color perception, which have been proposed to be responsible for patterns in child acquisition of color terms. If the lay-out of color categories in a color appearance model plays a central role, as many accounts have it (Dougherty, 1978; Harkness, 1973; Pitchford & Mullen, 2003), training a model on pairings of a color term and a color stimulus from this appearance model should suffice to explain observed developmental patterns.

We contrast this perceptual approach with a crosslinguistic semantic representation inspired by the Typological Prevalence Hypothesis. In particular, we build on previous work (Beekhuizen, Fazly, & Stevenson, 2014) suggesting that crosslinguistic elicitation data can form the foundation for a vector-based representational space that reflects the cognitive biases in how languages carve up a semantic domain into lexical categories. In line with the Typological Prevalence Hypothesis, such a representation may capture the *linguistically-relevant* aspects of color categorization, rather than purely perceptual ones. We then hypothesize that using the crosslinguistically-inspired color space in our model may yield a better match to human behavior in both color term acquisition and color discrimination. In this way we explore whether perceptual factors alone can explain these effects, or whether other kinds of cognitive biases come into play. Even if the crosslinguistic elicitation data does not show an improvement over the perceptual space, there is another factor motivating our alternative approach to a semantic space: namely, that its basis in crosslinguistic elicitation data means such representations can be derived for domains for which perceptual features are difficult or impossible to obtain (cf. Beekhuizen, Watson, & Stevenson, 2017).

For completeness, we also study the relative contribution of color term frequency in simulating both the developmental and relativity tasks. Rather than statically correlating term frequency

with ease of acquisition, as has been done in statistical studies (e.g. Yurovsky, Wagner, Barner, & Frank, 2015), we manipulate term frequency in our incremental learning model. By modeling both term frequency effects and the different approaches to semantics within a unified model of acquisition and use of knowledge, we aim to reveal how these factors interact in both developmental errors and linguistic relativity behaviors in the domain of color.

Color terms: acquisition and linguistic relativity

Here, we discuss literature on error patterns in children's color term acquisition and on linguistic relativity effects in the domain of color, as well as related modeling approaches.

Much research has considered the patterns of acquisition of *basic* color terms – i.e., monomorphemic terms, such as *red* or *purple*, with the following characteristics: they can be applied to a wide range of objects, are psychologically salient to speakers of the language, and have an extension that is not the subset of an extension of another color term (Berlin & Kay, 1969). It is well known that the development of children's use of color terms is generally slow; e.g., English children typically are only able to accurately use four of the basic color terms (*blue*, *red*, *green*, and *yellow*) by 4 years old (see Bartlett, 1978; Bornstein, 1985; Pitchford & Mullen, 2003; Soja, 1994, and references cited therein). The difficulty of learning color vocabulary leads to omissions and errors that can be informative about the mechanisms and representations involved in learning the terms of this domain.

Most of the above developmental studies focus on the order in which color terms are learned, where the 'acquisition' of a color term generally means that the term is used for the appropriate color category and not for any others. These approaches have shed light on the potential role of perceptual factors by linking them to the ease or difficulty of learning certain color terms. Far fewer studies have been concerned with the exact nature of the error patterns in color term learning, and mainly identify that errors are found between color categories that are adjacent in hue or saturation (Bartlett, 1978; Pitchford & Mullen, 2003; Shatz, Behrend, Gelman, & Ebeling, 1996). Very little work has looked at finer-grained patterns – such as the overextension or underextension of color terms – and it is these errors that have the potential

to reveal more nuanced influences on color term acquisition. For example, perceptual closeness may explain why two color terms are confusable, but asymmetric overextension of one to the other – e.g., using *blue* for PURPLE but not *purple* for BLUE, as observed for instance by Istomina (1960) and I. Davies, Corbett, McGurk, and MacDermid (1998) – indicates other factors play a role in color term learning.

Moreover, while we can identify, for example, that certain perceptual properties correlate with the color naming behavior of children, we do not fully understand *how* these biases are operational in the acquisition process. Computational modeling provides a way to explore the interaction of various factors and the mechanisms that bring them to bear on learning. For example, Yurovsky et al. (2015) present a mathematical model that identifies several factors that underlie under- and overextension of color categories; importantly, color term frequency, as well as the size and salience of the perceptual category, all play a role. However, their approach is not a learning model, and so does not address how these factors play out in the acquisition process itself. Belpaeme and Bleys (2005) do propose a model of color term acquisition over a perceptual semantic space, but this work focuses on biases in the emergence of community-wide systems, and does not consider the developmental trajectories of individual learners. The next step is to develop an actual cognitive processing model – one that incrementally learns from input items consisting of a color term and a representation of a color stimulus, and evaluate if and how the proposed factors do indeed affect the developmental patterns of errors it displays.

Ideally, such a computational cognitive model should be subject to further scrutiny: that is, we must also consider whether applying the acquired knowledge in the model shows behavioral effects seen in adults. The linguistic relativity hypothesis points to some interesting hypotheses in the area of color. If the lexical semantic categories of a language can influence perceptual discriminability of those categories, we can expect there to be behavioral differences between speakers of languages with different sets of basic color terms. Given a computational model in which certain factors play a role in matching human developmental patterns, we would want to see whether and how those same factors influence the learned

knowledge and its use in such relativity cases.

For example, in studies on Russian-speaking and English-speaking participants, Winawer et al. (2007) investigated the behavioral effect of having two basic color terms in Russian – *sinij* ('dark blue') and *goluboj* ('light blue') – for a region of color space that has a single basic term in English – *blue*. Their experiments demonstrated that Russian speakers are faster in a non-linguistic color discrimination task when the colors to be distinguished cross their basic color-term boundary, where English speakers (lacking a basic color term distinction) show no such effect. Roberson and colleagues also showed for multiple languages that subjects are able to discriminate color stimuli better around their own lexical category boundaries (Roberson et al., 2005, 2000). Bae et al. (2015) further demonstrate 'perceptual magnet' effects, whereby subjects perceive a color as more similar to the average member of a category (a region labeled by a basic color term) than it actually is.

An interesting recent mathematical model of some linguistic relativity effects is that of Cibelli, Xu, Austerweil, Griffiths, and Regier (2016). In their category adjustment model, two sources of information are combined in a color naming task in response to a color stimulus: the (noisy) perception of the color stimulus and the representation of the color category (a region of color associated with a linguistic term) that the stimulus is nearest to. This approach enables them to account for the category boundary and perceptual magnet effects of Roberson et al. (2005, 2000) and Bae et al. (2015). However, the Cibelli et al. (2016) model has no learning component that incrementally acquires the color category knowledge, so it is unable to explore the effects of various factors on both acquisition and use of color vocabulary knowledge.

To our knowledge, the neural network approach in Colunga and Gasser (1998) constitutes the only previous attempt at an integrated learning and processing model of linguistic relativity: A neural network is incrementally trained on artificial languages and artificial semantic domains to study both the ease of acquisition of lexical semantic categories and the cognitive consequences of the learned knowledge. The model shows interesting effects, such as less coherent categories being harder to acquire, as well as heightened attention in the model to the conceptual dimensions relevant to the category system of a particular language. However,

because it is applied on artificial languages and toy semantics, it remains to be shown whether the demonstrated effects are generalizable to actual lexical semantic systems over real semantic domains.

The model we present in this paper constitutes (to our knowledge) the first computational cognitive processing model to simulate acquisition and linguistic relativity effects when trained on naturalistic inputs and tested on empirical data. Like the approach of Colunga and Gasser (1998), our model has a distributed network architecture and displays similar effects. However, unlike theirs, our model is trained on data that captures the meanings of the actual semantic domain of color and reflects corpus frequencies of color term usages. Moreover, we compare the model output to human data from experimental studies: the child developmental patterns in Bateman (1915) (for English) and in I. Davies et al. (1998) (for Russian), as well as the differential color discrimination effects between English and Russian speakers in Winawer et al. (2007).

The current proposal builds on our earlier work (Beekhuizen & Stevenson, 2015), which used the Generalized Context Model (Nosofsky, 1987) to simulate the developmental English color naming data of Bateman (1915). While this approach showed interesting preliminary results, we subsequently adopted a different learning model – a Self-Organizing Map (SOM; Kohonen, Schroeder, & Huang, 2001). This model framework, which has been previously explored for word learning (e.g., Li & Zhao, 2013),[2] has several advantages for our purposes. In particular, the SOM naturally captures distances among input stimuli as map distances, an advantage for modeling color discrimination. Given that the map layout captures the typology of the semantic space in this way, SOMs are also useful for interpretation and visualization of the learned knowledge. Preliminary results with our SOM model were explored in

[2]Li and Zhao (2013) are focused on other aspects of word learning than we are here. The main differences between their approach and the one we propose below are as follows. (1) In their approach, an input item consists of a word that is always paired with the same semantic representations. In ours, input items consist of words paired with different semantic representations, and one of the purposes is to see how the model captures the range of color hues associated with a single color word. (2) Their model has distinct maps for the word forms and the word meanings, linked with Hebbian connections. Our model is a single map, onto which both word form and meaning are jointly projected, with the goal of seeing how the associated word forms influence the color hue topology of the map. These differences in design reflect the different acquisition phenomena these models are intended to simulate.

Beekhuizen and Stevenson (2016). The present work significantly extends the experiments and analyses with this model over that pilot investigation in two ways. First, here we use a much stricter evaluation criterion for the developmental results: the model is evaluated across all stimulus-response pairs simultaneously, whereas in Beekhuizen and Stevenson (2016) the model was evaluated per stimulus. Second, here we test the model across a range of SOM parameter settings, thus evaluating the robustness of the model to variations in these.[3]

The Computational Model

We approach our two issues of interest – the error pattern in color vocabulary acquisition and the linguistic relativity behavior in color discrimination – by simulating a human learner using a computational category learner. Specifically, as in Belpaeme and Bleys (2005) and Cibelli et al. (2016), we formulate the learning and use of color vocabulary as a categorization task that associates a color term – which is a semantic category label – with a color stimulus – which is a set of semantic features representing the stimulus. While training the model, we can record the errors in its production of color terms in response to color stimuli, and compare the model's error patterns to developmental errors in children. We can also simulate a color discrimination task using the trained model and compare it to adult behavior in such a task. In this section, we first describe the category learning model. We then discuss the two semantic representations of color stimuli that we compare in our experiments – a perceptual representation and a crosslinguistically-derived representation – as well as the two ways we sample input terms in order to explore frequency effects on learning.

Learning color categories

For our categorization model, we use a neural network called a Self-Organizing Map (SOM; Kohonen et al., 2001). Our SOM M is a square matrix of $m \times m$ cells $[c_{11}, c_{12}, \ldots c_{mm}]$; we refer to the dimension of the map with the parameter $size$ – i.e., in our simulations, m is set to

[3]The different approaches to evaluation make differences in the results hard to compare across the pilot and current studies. Foreshadowing the results here, we note that the most robust patterns are found in both experiments, but that the current set-up allowed us to evaluate the model performance more appropriately.

varying values of $size$. Each training input to the model is a vector i of feature values, some of which encode color terms and some of which encode the associated color semantics (i.e., properties of color stimuli). Every cell in the map, c_{jk}, is a vector whose dimensions are the same features, such that color vocabulary (a set of term–color associations) is learned by jointly projecting pairs of term and semantic features from the inputs i onto the SOM in an aggregate representation in the map cells. To effect this, the values in the map cells gradually change in response to the training input. (We use c_{jk}^n to refer to the value of cell c_{jk} at time n when we need to distinguish the cell values over time.) Fig. 2 illustrates a SOM and input item.

At each iteration n of training, an input i is compared to all cells $c \in M$, and is subsequently mapped to the cell to which it is most similar, called its Best Matching Unit, or $BMU(i)$:

$$BMU(i) = \underset{c \in M}{\operatorname{argmin}}\, d_{\text{feat}}(c, i), \tag{1}$$

where d_{feat} is the Euclidean distance between the feature value vectors of c and i.

The feature values of i are then used to update the feature values of $BMU(i)$ as well as (to a lesser extent) those of its neighboring cells in a radius around $BMU(i)$. In this way, over time M will come to display a topology that reflects the similarity among the input items. Assuming input i at time n, all cells c_{jk}^n are updated to yield c_{jk}^{n+1} in proportion to their map distance from $BMU(i)$:

$$c_{jk}^{n+1} = c_{jk}^n + h_{jk} \times (i - c_{jk}^n), \tag{2}$$

Here the vector difference between the input and the cell, $i - c_{jk}^n$, can be thought of as an error signal (the discrepancy between the cell c_{jk}^n and the input i) which is used to "nudge" the features values of c_{jk}^n toward the values of i, according to the weight h_{jk}, which reflects the distance of c_{jk}^n from $BMU(i)$:

$$h_{jk} = \alpha \times \exp\Big(- \frac{d_{\text{map}}(c_{jk}, BMU(i))}{2 \times \sigma^2} \Big). \tag{3}$$

That is, h_{jk} yields the excitation of the cell c_{jk} given a center of activation at the coordinates of $BMU(i)$, taking into account their Euclidean distance in the map grid, given by d_{map}. Here, $\alpha = [0, 1]$ is a learning rate parameter, and σ is the neighborhood radius of $BMU(i)$, which determines the extent of the neighborhood of cells affected by the feature values of i. The settings of α and σ (and the adjustment of σ during learning) are discussed in the section 'Model simulations and parameter settings'.

To test the model, we can simulate the human phenomena of interest, both during and following training, by periodically presenting the model with test inputs. To assess whether the model converges, and to compare its developmental patterns to children's color naming over time, we need to mimic a color naming task by seeing what color term(s) the model associates with each color stimulus at various points in time. To do so, we present the model with a vector s of semantic features representing an unlabelled color chip – i.e., an input without the term-feature vector. We then find $BMU(s)$ using Eqn. (1), but only comparing the semantic features of each map cell c to color stimulus s. The term features of $BMU(s)$ yield a distribution over possible color term responses to the color stimulus s; that is, we can calculate $P(t|s)$ for all terms t as follows:

$$P(t|s) = \frac{\text{value}(t, BMU(s))}{\sum_{t' \in T} \text{value}(t', BMU(s))}, \tag{4}$$

where value$(t, BMU(s))$ is the value for feature t in cell $BMU(s)$. To determine a single best response in the color naming task, we take the term with the highest conditional probability as the response term t_s for s:

$$t_s = \underset{t \in T}{\text{argmax}}\, P(t|s). \tag{5}$$

The ability of people to discriminate two colors is correlated with the distance between the two stimuli – i.e., perceptually closer stimuli are harder to distinguish. Thus, to mimic a discrimination task in our model between two color stimuli s_j and s_k (again, comprised solely of semantic features), we need to extract the model's assessment of their distance according to its learned knowledge. We can consider the distance in the map between their two

Best-Matching Units – i.e., $d_{\text{map}}(BMU(s_j), BMU(s_k))$ – as representing the discriminability of the stimuli within the model. More detail on the methods used to evaluate the model are provided in the respective results sections below.

The term and semantic feature spaces

Formally, each cell of the model and each training input to the model is a concatenation of two vectors, a term-feature vector representing a color term and a semantic-feature vector representing the semantic properties of a color chip. Each cell of the SOM thus represents a learned association between a set of semantic-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 basic color terms in a language. To represent the uth term t_u of T within an input item, the uth feature is set to a value $a \in (0, 1]$, and all other term features set to 0. For example, in a language with four basic color terms, the term vector for input term t_2 is $[0, a, 0, 0]$. The parameter a reflects the relative importance of term features (vs. semantic features) in training. The term-feature vector of each cell of the SOM will come to hold a distribution over terms; see the example cell in Fig. 2. As shown in Eqn. (4), the term vector of $BMU(s)$ can be normalized to arrive at a probability $P(t|s)$ for each term t given a color stimulus s.

Turning to the semantic features of a color chip, recall that one of our goals in this paper is to compare two different semantic representations with regard to whether they enable the model to match human behavior in color acquisition and discrimination. First we consider a semantic representation based on perceptual properties of color. The $L^*a^*b^*$ space is a color appearance model thought to encode the human perceptual dissimilarity between colors (Fairchild, 1998), and so we adopt this as our perceptual semantic representation. The 3-dimensional $L^*a^*b^*$ space specifies values for the dimension of luminance (L^*), and for two chromatic dimensions – a RED–GREEN axis (a^*), and a BLUE–YELLOW axis (b^*).[4]

[4] Recall that in Eqn. (1), we take the Euclidean distances over map and input feature vectors as a first step in learning. A reviewer noted that problems with Euclidean distances over the $L^*a^*b^*$ color space have been identified, as in Komarova and Jameson (2013). While we are aware of such (and other) issues with this color appearance space, $L^*a^*b^*$ was intended to be a geometric space in which Euclidean distances are meaningful,

While such a perceptual representation might be presumed to be adequate for a perceptual domain like color, there is suggestive evidence that linguistic categorization plays a role in how we process colors (Brown & Lenneberg, 1954). For our second semantic representation, we thus consider a method that draws on lexical semantic categorizations. Specifically, we adopt the method of Beekhuizen et al. (2014), who propose an approach for creating a crosslinguistic vector space for a semantic domain. The resulting semantic space is intended to reflect the set of cognitive biases for that domain in human language learners – i.e., the underlying tendencies in carving up a semantic domain into categories labeled by lexical items – following the Typological Prevalence Hypothesis (Gentner & Bowerman, 2009). For the domain of color this would mean that if many languages place a lexical boundary in the same region of the color space, this region, for some reason, attracts semantic categorization boundaries, or forms a natural location for such a boundary because it lies between two salient color areas.

As proposed by Beekhuizen et al. (2014), we use crosslinguistic elicitation data over situations in a semantic domain to form the basis of a vector-based semantic representation of those situation meanings. In this case, the situations S are 330 Munsell color chips, for which the World Color Survey (Kay et al., 2009) provides elicitation data in 110 languages, with around 25 participants per language.[5] Using this data, for each language l in the sample L, we can form a conditional probability distribution $P_l(t|s)$ over the terms $t \in T_l$ (the basic color terms in l) for every semantic stimulus $s \in S$ (each color chip). The probability distribution $P_l(\cdot|s)$ is based on the relative frequency of the term responses in l to color chip s over all informants. Each color chip s can then be represented as a vector that is the concatenation of these observed conditional probability distributions for all languages in the data; see Figure 3.

---

and remains a standard in cognitive science (e.g., cf. its use in Bae et al., 2015; Regier, Kay, & Khetarpal, 2007; Wagner, Dobkins, & Barner, 2013; Webster & Kay, 2012).

[5]The 110 languages in the sample are spoken in non-industrialized cultures and are unwritten. Following Kay et al. (2009), we assume that the sample is sufficiently broad to capture the range of color distinctions made in language. While we acknowledge that the salience of certain discriminations in the color space may vary as an effect of culture (and more generally, of the ecology the speakers of a language inhabit), we expect that this is not a major issue here. To preview our results, we find that the Self-Organizing Map learns the color term systems of two languages spoken in industrialized cultures when using our feature space derived from the color discriminations made in the 110 languages of the World Color Survey.

Given the World Color Survey data, this approach yields 2339 semantic features (one for each color term in each language) for a color chip s. The intuition is that two color chips, s_i and s_j, are more similar the more they are referred to with the same terms by subjects within each sampled language – i.e., the more similar $P_l(\cdot|s_i)$ is to $P_l(\cdot|s_j)$, for all languages $l \in L$. If the use of terms for two color chips is more similar, the vectors (the rows in Fig. 3) for these colors are more similar too. For example, according to this intuition, color chips s_1 and s_2 are more similar to each other than to chip s_n in Fig. 3.

It would be possible to use each of these two semantic spaces – $L^*a^*b^*$ and the crosslinguistic approach – to represent the Munsell chips used as input stimuli in our model. However, the two representations have a widely different number of feature dimensions (3 vs. 2339). As such, a direct comparison would not isolate the effects of the information in the feature spaces from the effects of dimensionality. To achieve comparable semantic spaces, with feature vectors of the same dimension, we cast the two feature spaces into a common format. For each of the semantic representations, we use a distance matrix between all pairs of color chips in S (chips in the Munsell set) to create comparable vector spaces. Intuitively, rather than encoding a color chip s_i directly as its corresponding vector in the semantic space, we encode it as a vector of distances to every other color chip s_j, as calculated in the semantic space.

Specifically, every entry $d_{i,j}$ in the $|S| \times |S|$ distance matrix D for a feature space contains the Euclidean distance between color chips s_i and s_j in that feature space. It is these matrices that we use as semantic features in the model: a color chip s_i is represented as row $D_{i,\cdot}$ – a vector of $|S|$ elements – from the distance matrix over the underlying semantic space. In this way, each color chip s is represented not as an absolute location in an n-dimensional semantic space, but as a relative location in an $|S|$-dimensional space – i.e., relative to all other color chips in S.

We call the new (comparable) semantic feature spaces obtained in this way `perc` and `xling`, for the $L^*a^*b^*$ perceptual space and the crosslinguistic elicitation space respectively.[6] To give

[6]Because the `perc` and `xling` spaces have the same dimensionality, we can use the two feature spaces simultaneously by taking the average over the two matrices. In pilot experiments, we found that this combined space performed as good as (but not better than) the best of the two individual spaces, so we do not report the results here.

an impression of the topology of the two spaces, we show in Fig. 4 a 3-dimensional projection of the color categories for each language in each feature space. The projection is given by a Multidimensional Scaling (MDS) of the category centroids of the training items in each language. As we can see, the `perc` space is more evenly spread out over the geometric space, whereas the `xling` space shows a more uneven spread due to particular crosslinguistic naming tendencies. For example, BROWN, BLACK, PURPLE, and GREY are all close together in `xling` space because many languages label many of the chips subsumed under these four labels with a single term. On the other hand, WHITE and GREY are far apart in the `xling` space, despite being perceptually relatively similar (as can be seen in the `perc` space), because very few languages label the two with a single term. A key goal of our experiments is to explore how these different semantic spaces influence the performance of the model.

Training input and the role of term frequency

As noted earlier, it has been proposed that color term frequency plays a role in errors in color vocabulary acquisition (Yurovsky et al., 2015). Thus, in addition to exploring any differential effects that arise from the two semantic spaces for encoding color stimuli, we also investigate the role of term frequency in our model. We do this by varying the sampling of inputs presented to the model in training. Specifically, we train the model for language $l =$ English or $l =$ Russian by sampling input items as a pair of a color term $t \in T_l$ (basic terms from l) and a color stimulus $s \in S_l$ (encoded by our `perc` and `xling` semantic features). The terms T_l are the basic color terms in each of English and Russian, shown in Table 1. Our stimuli S_l are the subset of Munsell color chips for which we have adult elicitation data in l – color terms from multiple informants – to serve as the "ground truth". For English, $|S| = 211$ and $|T| = 11$ (from Berlin & Kay, 1969); for Russian, $|S| = 49$ and $|T| = 12$ (from I. Davies & Corbett, 1994). (In the case of I. Davies and Corbett 1994, who provide the values in Yxy color space for their stimuli, we find the Munsell chip closest to each Yxy value in their data.) We use the distribution $P(t, s) = P(s|t)P(t)$ to sample inputs from T and S. We estimate the conditional probability distributions for $P(s|t)$ from the adult elicitation data in English

(Berlin & Kay, 1969) and in Russian (I. Davies & Corbett, 1994). We extract the counts of each Munsell chip s labeled by a given term t, and obtain $P(s|t)$ by normalizing over all s to sum to 1 for each t.[7]

In order to study the effect of color term frequency, we vary how we estimate $P(t)$. As one estimation, we use the relative term frequency over all basic 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 from the spontaneous spoken section of the Russian National Corpus.[8] We also estimate $P(t)$ using a uniform distribution, reflecting a learner that is unaffected by color term frequency. By comparing the corpus-based $P(t)$ to a uniform $P(t)$, we can see which learning and relativity effects are due to a naturalistic frequency distribution, and which to the structure of the semantic feature space. The two sampling options are called `corpus` and `uniform`. The relative term frequencies and the uniform $P(t)$ per language are given in Table 1.

Model simulations and parameter settings

As we are interested in how well the two semantic feature spaces and frequency samplings explain empirical developmental patterns and color discrimination effects, we test the match of the model performance to human behavior in four conditions, comprising every combination of `features = {perc, xling}` and `sampling = {corpus, uniform}`. We test these combinations under a number of settings of the free parameters in

[7] Note that this sampling approach does not address how frequently various color stimuli occur in the world or in discourse; it simply tells us for some color term the likelihood of various color chips being labeled with that color term. At this point we know of no way to estimate a sampling of colors people refer to: one might use estimates of colors 'in the real world' (Burton & Moorehead, 1987; Hendley & Hecht, 1949; Howard & Burnidge, 1994), but this does not give us estimates of colors in the child's world, nor the discourse prominence of these colors.

[8] The Russian National Corpus (RNC) is available at `http://www.ruscorpora.ru/en/index.html`. Note that the source of the frequency data turns out to have little effect on the results: running our experiments on frequency data from the two publicly-available Russian CDS corpora on CHILDES (MacWhinney, 2000) or from adult spontaneous spoken English in the COCA corpus (M. Davies, 2008-) yielded identical significance patterns for convergence and development experiments, and the same qualitative pattern for the discrimination experiment. This suggests, foreshadowing our results, that any differences between child-directed and adult language (e.g., the terms for BLACK and WHITE in both languages being more frequent in adult language than in child-directed language) do not change the results as much as using a uniform frequency distribution over terms.

the model to ensure robust effects given a range of their values.

The free parameter σ (see Eqn. (3) on page 11), and how it changes over the course of training, dictate the neighborhood radius that determines the "spread" of an input's influence on the cells of the SOM. We set σ to a reasonably large value to start (5.0), and then decrease it gradually (by steps of 0.1) to a minimal value (0.001) based on how well the map cells are able to match the inputs (i.e., the decrease of σ is modulated by the error signal between input i and its Best-Matching Unit $BMU(i)$; Li and Zhao 2013). A large initial value for σ and a small decrease were chosen to avoid all-too-quick convergence in the model, since slower learning is necessary to study developmental effects.[9]

The SOM has three other free parameters – a [strength of term features], α [learning rate], and $size$ [number of rows and columns of the SOM] – for which there is no strong motivation to set them at a specific value. To assess how robustly the model performs on combinations of `features+sampling`, we look at a range of values of the model parameters a, α, and $size$ for each combination. Specifically, we run 30 simulations for each of the 27 combinations of $a = \{0.1, 0.3, 0.5\}$, $\alpha = \{0.1, 0.3, 0.5\}$, and $size = \{7, 10, 12\}$, whose combinations were found to reliably yield a good performance on simulating the adult naming pattern for both languages (cf. Appendix A).. We report both the average performance of a `features + sampling` combination over these 27 free parameter combinations, as well as the range of performance scores.

Before turning to our main results in the next two sections, we first briefly summarize simulations of the model undertaken to confirm that it converges to knowledge of color categories that matches adult color naming; details can be found in Appendix A. Under all combinations of `features + sampling`, for both languages, the model achieves mean accuracies of 94%–97% for its most probable term matching the modal adult color name response. There are only very small (1%–2%) differences in performance between the

[9] A reviewer raised the question whether the setting of σ and the learning rate α interact, which we agree is a possibility, but does not seem to be an issue here. The high initial value of σ was chosen to allow the map to be very elastic initially, and become more fixed as its mapping error decreases. Pilot experiments revealed that setting initial σ much lower prevented the model from displaying substantial overextension patterns. Given this high value, none of the settings of α significantly affected the model's ability to simulate the developmental pattern (as presented in the section on development).

semantic feature spaces and the input sampling methods, attesting that: (a) both `perc` and `xling` are adequate as the basis of a semantic representation for learning color word associations; and (b) given sufficient input, the model learns these associations under either `corpus` or `uniform` term sampling. Because the model stops showing improvements in its fit to adult color naming after 25K inputs, we use 30K inputs as representing the "adult" state of the model in our discrimination experiments. Figure 5 gives two visualizations of trained Self-Organizing Maps, one for English and one for Russian. We can see that most colors cover a single contiguous region on the map and that similar colors are located adjacently on the map, indicating that the SOM indeed captures the topology of the training data on its two-dimensional grid.

Model Results: Acquisition and Developmental Errors

The first question we explore with our computational model, posed in the Introduction, is: What drives the difference in ease of acquisition of various lexical semantic categories? We explore the color naming behavior of our model over the time course of learning, and compare it to the developmental error patterns in children. In particular, we explore the possible semantic and frequency factors that influence the overextension of one color term for another, by considering the operation of our model under different settings of `features+sampling`.

Empirical data on developmental errors

We use our model to simulate the color term acquisition patterns in two developmental color naming studies (in English and in Russian), for which detailed over- and underextension data was available.

We take our English data from Bateman (1915), who presents a cross-sectional experiment on 591 English-speaking children in five age groups (151 6-year-olds, 160 7-year-olds, 179 8-year-olds, 72 9-year-olds, and 29 10- and 11-year-olds). Each child was presented with eight color chips corresponding to eight of the eleven basic color terms of English (cf. Table 1) –

specifically, representing the eight color categories BLACK, WHITE, RED, BLUE, YELLOW, GREEN, PURPLE, and ORANGE. (Three other color categories are associated with basic color terms in English: BROWN, GREY, and PINK.) For six out of the eight colors, few or no overextension errors were found by Bateman, whereas for PURPLE and ORANGE, children more frequently responded with an incorrect term. Fig. 6 gives the probability of the response terms given these two color chips (i.e., the proportion of children in an age bin labeling a chip with each of the indicated terms). As we can see, the error pattern for ORANGE is somewhat haphazard, but there is a clear underextension of the term *orange*. For PURPLE, it was found that *blue* is the primary term being overextended to this chip. From Bateman (1915), we have access to the counts of all color term responses to each of the eight color stimuli he uses, across each of the five age groups.

I. Davies et al. (1998) carried out a similar cross-sectional experiment in Russian: 200 Russian-speaking children (80 3-year-olds, 80 4-year-olds, and 40 5-year-olds) were each presented with 12 color chips, one for each of the 12 basic color terms in Russian (see Table 1). The color chips were selected to be good examples of color categories that occur widely across languages, with the exception of DARK BLUE and LIGHT BLUE, for which Russian has two basic color terms (*sinij* and *goluboj*, respectively). In these cases, the chips that were thought to be good representations of these two Russian color categories were used. We extract data for our study from I. Davies et al. (1998), who report only the proportion of children giving the dominant term, as well as the most-frequent error per chip per age group; see Fig. 7 for some examples.[10] Notable overextensions were found in the blue-purple range (Figure 7, top row), with *sinij* 'dark blue' being used for the LIGHT BLUE and PURPLE chips at 3 years old more than the correct terms (*goluboj* 'light blue' and *fioletovyj* 'purple', respectively). The term *goluboj* was also overextended to DARK BLUE, but to a lesser extent than *sinij* to LIGHT BLUE. Several other chips displayed significant overextension patterns (Figure 7, bottom row): PINK, ORANGE, and BROWN. For the remaining chips (WHITE, BLACK, GREEN, RED, YELLOW, and GREY) few overextension errors were found.

[10]The bars in Fig. 7 do not add up to 1 because the remaining probability mass is assigned to other terms and/or failures to respond.

Evaluating development in the model

Our goal is to see how well the model's term responses to test stimuli mimic the above patterns of child color term responses. We do so by looking at the probability $P(t|s)$ (as calculated using Eqn. (4), page 12) for all basic terms t in response to a test stimulus s.

The test stimuli for evaluating development were chosen to reflect as closely as possible the descriptions of color chips shown to children by Bateman (1915) for English and by I. Davies et al. (1998) for Russian. Bateman only verbally describes his 8 test stimuli as being good exemplars of the 8 color categories he studies. For our model test stimuli, we took the (multiple) Munsell chips identified by Berlin and Kay (1969) as the central members of each of those 8 English color categories. To test the model on a particular color category, we found the predicted $P(t|s)$ for each of the chips in the corresponding set, and took the average over them as the model prediction for that category. (For example, if five chips $s_1, \ldots, s_5$ were identified as central members of BLUE, we averaged the $P(t|s)$ values for each of $s_1, \ldots, s_5$ to give values of $P(t|s)$ for the hypothetical BLUE chip used by Bateman.) I. Davies et al. (1998) report the values of their 12 test stimuli in the Yxy color space; we convert each into $L^*a^*b^*$ values, and then identify the closest Munsell chip as the corresponding test stimulus s for which we extract $P(t|s)$. Thus, S_{test} in each language consists of the semantic features s for 8 sets of Munsell chips (English) and 12 Munsell chips (Russian).

We now need to compare the model's output of $P(t|s)$ on all test stimuli in S_{test} to the patterns of child responses in the human data. Directly comparing the numerical distributions can mask important ordering differences (e.g., one distribution $P(\cdot|s)$ might be numerically closer to the child distribution of responses, but get wrong a critical ordering of one term favored over another). To ensure that we assess the quality of ordering of the model responses, we compare the *rankings* of term–stimulus pairs in the behavioral data and in the model. The ranking in the behavioral data is given by the number of children giving a particular color term t in response to a color stimulus s. The model ranking is given by taking all term–stimulus (t, s) pairs for our test stimuli and ranking them according to their $P(t|s)$ values, from high to low. We compare these two rankings using the non-parametric test Kendall τ_b.

Next we must consider at which point in training to perform this assessment of the model's responses. We cannot know *a priori* how to map the children's ages from the color naming studies onto corresponding amounts of model training. Rather than arbitrarily picking a time point, we search for an alignment of the children's term rankings in the different age groups to the rankings of the model in various consecutive bins of model time. Specifically, we test various sizes of consecutive sequences of input data, beginning at various times, and find the alignment of ages to test points in the model that maximizes the fit of the model rankings to the empirical data. By ensuring, for every simulation, that we have found the best fit of the model to the human data, we know that we are fairly comparing the optimally fitting results for all conditions (`features = {perc, xling}` and `sampling = {corpus, uniform}`) under each parameter setting.

More precisely, given A age groups in the data ($A = 5$ for English and $A = 3$ for Russian), we need to find A consecutive bins of test data in the model whose rankings best fit the child rankings. The best alignment to a series of bins is given by the alignment for which the τ_b between the ranking for each age group and the ranking of the corresponding series of test points is highest. The development score $\mathbf{score_D}$ is then defined as the τ_b for the best alignment, averaged over the A age groups.[11]

The difference between the `features` and `sampling` settings are evaluated with a two-way ANOVA with the $\mathbf{score_D}$ as the dependent variable, and `features` and `sampling` as independent variables. All significant differences are reported and all reported differences are significant, to $p \leq .001$.

While we know of no reasonable "baseline" performance to compare our results to, we can consider what an error-free system would do – i.e., a model always predicting the correct term with a probability of 1. Since the goal of the developmental evaluation is to consider whether the model is making errors that match those of children at different ages, it is interesting to see how well a system making no errors would match children's behavior.

[11] Designing a transparent evaluation procedure for (partially incomplete) developmental naming data is both important and very difficult. Several methods we tried masked some of the patterns we were interested in. The current procedure, while lacking an assessment of precise quantitative fit, was adopted because it allows a direct inspection of the critical ordinal patterns in the data.

Results on development

Considering the results for English in Table 2 (left side), we find that (although significant) the differences between the feature spaces and between the sampling procedures are relatively small (a difference of 5% between the best and worst combinations of `features` + `sampling`). Moreover, under no combination of `features` + `sampling` does the aggregated mean $\mathbf{score_D}$ of the model surpass the error-free learner. This is primarily due to two things. First, there are generally few child errors to predict for English, making the error-free learner a strong baseline to surpass. Second, the errors that do occur in the child data are rather haphazard: with the exception of *blue* for PURPLE (see Fig. 6b), no other color stimulus has an error response that is dominant over other error responses (cf. ORANGE in Fig. 6a). A $\mathbf{score_D}$ that surpasses the error-free learner would require the model to predict the presence of haphazard overextensions, while also predicting the absence of non-attested overextensions.

On the other hand, we do find indicators of a difference between the feature spaces and between the sampling methods on the English data. First, `xling` has a significantly higher $\mathbf{score_D}$ than `perc`, and `corpus` has a significantly higher $\mathbf{score_D}$ than `uniform`. Second, considering all the simulations for each `features` + `sampling` setting, we find that many more `xling` + `corpus` simulations surpass the error-free learner than simulations using other model settings. Specifically, whereas 21% of the `xling` + `corpus` simulations scored higher than the error free learner, only 10% or fewer of the other combinations did so (10% of `perc` + `uniform`, 9% of `perc` + `corpus`, and 3% of `xling` + `uniform`).

Turning to Russian, recall that the observed errors in children are much more prominent and consistent than in English (see Fig. 7). Importantly, in our model results on Russian (Table 2, right side), we find that all mean $\mathbf{score_D}$ values for all four combinations of settings surpass the error-free learner. Moreover, the results across the `features` + `sampling` settings show a similar pattern as in English, but more pronounced: the magnitude of the difference between the best (`xling` + `corpus`) and worst (`perc` + `uniform`) combinations is much greater than in English (here 14%), and there are consistent and stronger differences between

both features and sampling, with xling performing better than perc by 4–7%, and corpus performing better than uniform by 7–10%.

Analysis of Developmental Results

In a by-item analysis of model output in both languages, we find that most differences in performance between the features + sampling settings were found for color categories that display systematic overextension in children, e.g., the use of *blue* for PURPLE in English, and the use of *sinij* ('dark blue') for PURPLE and for LIGHT BLUE in Russian. In other words, the locus of the better performance of the xling + corpus combination is its ability to capture the asymmetric error patterns of interest.

In order to understand why this difference in performance arises, we identified two factors that underlie the model's pattern of errors. The most important factor is the confusability of color categories: If two categories are close together in a feature space, it is more likely that a stimulus belonging to one of those categories could be misclassified as the other (as for instance Bartlett 1978 has argued for color terms in perceptual space). Semantic confusability explains the difference in performance between the perc and xling feature spaces. (See Fig. 4 on page 52 for the positioning of the color categories of each language in each semantic space.) For English, the BLUE and PURPLE test items are closer together – more confusable – in xling than in perc. Similarly for Russian, the DARK BLUE and LIGHT BLUE test items are closer together in xling than in perc. Moreover, the xling space displays less confusability than the perc space between pairs of categories for which we find no overextension errors. For example, in Russian the model trained on the perc space predicts an (unattested) pattern of overextensions of *rozovyj* ('pink') to PURPLE as well as *fioletovyj* ('purple') to PINK, because the PURPLE and PINK categories are relatively close to each other in the perc space. This pattern of mutual overextensions is absent for the model trained on the xling space, in which the two colors are further apart and hence much less confusable. Semantic confusability alone can lead to "mutual overextensions" – e.g., *sinij* ('dark blue') being used for some LIGHT BLUE stimuli, and *goluboj* ('light blue') being used for some

DARK BLUE stimuli (see Fig. 7a and Fig. 7b). However, children notably make asymmetric overextensions – as the figures show, *sinij* is overextended more to LIGHT BLUE than *goluboj* is to DARK BLUE. Moreover, there are cases where there is little or no "mutual" overextension – e.g., *sinij* is common for PURPLE but not *fioletovyj* for DARK BLUE (see Fig. 7c). Similarly the overextension of *blue* to PURPLE in English is asymmetric. These asymmetries are where the second factor – that of frequency – comes in (cf. Yurovsky et al. 2015): i.e., relative frequency of terms can induce a directionality over potential overextensions that arise from semantic confusability. For example, the higher frequency of *sinij* ('dark blue') over *fioletovyj* ('purple') can mean that when the DARK BLUE and PURPLE regions are confused in the feature space, the term *sinij* ('dark blue') is overextended, but not the term *fioletovyj* ('purple').

This frequency factor plays out in our model as follows. Under `sampling = corpus`, the model initially does not assign a stable location on the map for stimuli associated with infrequent terms: that is, there is an initial asymmetry of representational strength between frequent and infrequent terms. Only when a fuller range of stimuli has been seen does the learner carve out a set of cells on the map for the infrequent terms. For example, in English, when the model encounters far fewer examples of PURPLE than BLUE, this leads to the representational asymmetry just described. Combining this asymmetry with the fact that the categories BLUE and PURPLE are very similar to each other and hence confusable, we find that the model overextends *blue* to PURPLE, but not *purple* to BLUE. Frequency can also block overextensions from happening. For example, in Russian, the (unattested) overextension of *rozovyj* ('pink') to RED under `uniform` sampling is mitigated using `corpus` sampling: the frequency of *krasnyj* ('red') is much greater than that of *rozovyj*, hence RED has a stronger initial representation than PINK. Because of this, in both feature spaces when using `corpus` sampling, the model predicts an overextension of *krasnyj* to PINK, but not *rozovyj* to RED. Since such representational asymmetry does not arise when sampling from a uniform distribution over terms, `sampling = uniform` fails to capture either the asymmetric overextension effects or their blocking.

Taken together, these results indicate that the `xling` feature space captures between-category

confusability better than the `perc` space, and that frequencies under `corpus` sampling appropriately drive the asymmetry of the resulting overextensions (while `uniform` cannot). Thus our model points to two interacting factors as the source of asymmetric overextensions in children – one involving the semantic representation of the color categories, and one involving the frequency of the terms used to refer to them.

Model Results: Discrimination Task/Linguistic Relativity

The second question we explore with our computational model, posed in the Introduction, is: How does acquiring a lexical semantic system of categories influence cognitive processing in that domain? We examine the effect that having a category boundary in the color space facilitates the discrimination of stimuli that are located across the category boundary from each other. As in development, we explore the mechanisms that lead to the observed behavior, and whether and how the semantic space and input sampling influence the results in our model.

Empirical data on color discrimination

As noted above, Russian has two basic color terms in the region covered by English *blue*: *sinij* 'dark blue' and *goluboj* 'light blue'. Winawer et al. (2007) ask whether this situation affects behavior in color discrimination: The expectation is that if linguistic relativity holds, the linguistic distinction between two stimuli (*sinij* for DARK BLUE and *goluboj* for LIGHT BLUE) would help a Russian speaker to discriminate between the corresponding color stimuli in a non-linguistic task (compared to an English speaker, for whom they are both *blue*). Winawer et al. (2007) presented adult monolingual speakers of Russian and English with triplets of a stimulus color chip, an identical target chip, and a different distracter; Fig. 8 shows examples of such stimulus-target-distracter triplets. Participants were asked to decide which of the target or distracter was identical to the stimulus, and response latency was measured. On each trial, the distracter was either perceptually more similar ('near') or less so ('far') from the stimulus chip (according to a color appearance model). Critically, the distracter was also either 'within'

the same category of DARK BLUE or LIGHT BLUE as the stimulus, or 'across' the category boundary, based on a separate per-participant labeling task. Note that Russian speakers labeled the stimuli with their basic color terms of *sinij* and *goluboj*, while English speakers, lacking this basic category distinction, used the multi-word terms of *dark blue* and *light blue*.

First considering the 'near' vs. 'far' cases, participants in both languages were slower at picking the target chip when it was 'near' the distracter compared to when it was 'far.' This confirms that color discrimination in this task is more difficult when stimuli are perceptually more similar, for both English and Russian speakers.

Now turning to the critical 'across'- vs. 'within'-category cases, differences arose between the English and Russian speakers. For English speakers, there was no difference in response latency to 'across' vs. 'within' cases, whether overall or separated into the 'near' and 'far' stimuli. Although English speakers labeled the stimuli differently as *light blue* and *dark blue*, this difference does not affect their speed in non-linguistic color discrimination. By contrast, Russian speakers picked the target faster when it was located 'across' their basic category boundary from the distracter compared to when it was 'within' the same category. That is, a target-distractor pair labelled by different basic color terms was easier to discriminate than a target-distractor pair where each was labelled by the same basic term. When broken down by 'near' and 'far' stimuli (holding the stimulus distance constant), it was found that this effect held only in the 'near' cases. Intuitively, this indicates that Russian speakers treated 'across' target-distracter pairs as more distinctive than 'within' pairs – even though they were both equally close perceptually – showing an influence of their linguistic categorization using the basic terms *sinij* and *goluboj*. Interestingly, this effect could be isolated to the perceptually 'near' cases, suggesting that linguistic categorization only plays this facilitatory role in harder tasks.[12]

[12]Winawer et al. (2007) tested participants under three conditions: with a verbal interference task, with a spatial interference task, and with no secondary task. We report in the text the results of the no interference condition, since our model cannot simulate the secondary tasks of the other conditions. The results for English speakers was the same across all three conditions. For Russian speakers, the results when performing the spatial interference task were the same as in the no-interference condition, but when performing a verbal secondary task their linguistic category advantage was eliminated. Taken together, this patterns shows that the observed effect in the no-interference condition is a linguistic influence on non-linguistic color discrimination.

For our simulations of color discrimination, we adapt the stimuli from the experiment of Winawer et al. (2007), and compare our model trained on English vs. Russian in terms of discriminability of those stimuli, assessing performance on perceptually 'near' vs. 'far' and 'across'- vs. 'within'-category pairs.

Evaluating discrimination in the model

As noted earlier, we can mimic the human color discrimination task in our model by determining the distance between colors in the model's learned representation, reflecting the fact that greater distance between two stimuli leads to easier discriminability by humans. Specifically, we take the state of the model after having processed $30,000$ input items to correspond to adult organization of the color terms, and we take the distance between a pair of stimuli in the learned map of the model to correspond to the perceived difference between them, and correspondingly the degree of ease or difficulty in discriminating them. This operationalization follows naturally from the nature of learning in the SOM, as its acquired representation attempts to be a faithful projection of our semantic and term features onto a 2-dimensional grid. Although map distance cannot be directly interpreted as reaction time, we can evaluate the various feature settings by their qualitative fit to the observed behavioral pattern.

We convert the 20 stimuli from the color discrimination task of Winawer et al. (2007) into our representation of color semantic features as follows. Consider the `perc` feature space; the procedure is identical for the `xling` space. We first take the stimuli at each end of the scale in the Winawer et al. data set (i.e., the darkest and lightest blue), find the nearest Munsell chip for each, and encode them in the `perc` feature space. Call these stimuli s_1 and s_{20}. For the remaining 18 stimuli between these, we linearly interpolate the values of s_1 and s_{20} in the `perc` feature space. This procedure yields a vector $S_{\text{disc}} = [s_1, ..., s_{20}]$ of color stimuli, where all distances between stimuli s_i and s_{i+1}, for $1 \leq i < 20$, are equal. Following Winawer et al., we consider two stimuli s_i and s_j to be 'near' if $j = i + 2$, and 'far' if $j = i + 4$.

In order to decide if pairs of a target and a distracter are in the same category (i.e., both DARK

BLUE or both LIGHT BLUE), Winawer et al. (2007) established per-participant category boundaries by asking each participant, following the discrimination experiment, to label the twenty stimulus color chips as *sinij* ('dark blue') or *goluboj* ('light blue') for Russian, and *dark blue* or *light blue* for English. For a Russian-trained model, we can analogously retrieve the best term label t_s for each of the stimuli in S_{disc} (using Eqn. (5) on page 12), but we cannot do this for an English-trained model, since it is not provided different terms for the two blues. According to Winawer et al., the observed category boundaries for Russian and English hardly differ, and so we use as the English boundary the mean location of the Russian category boundary under the same combination of model settings. A target-distracter pair (s_t, s_d) is now considered 'within'-category if s_t and s_d are on the same side of the category boundary we set, and 'across'-category if on opposite sides.

We can now find the *model's perception* of the distance between any two of these stimuli s_i and s_j as the Euclidean map distance d_{map} in the SOM between their Best-Matching Units – i.e., $d_{\text{map}}(BMU(s_i), BMU(s_j))$. The greater the distance in its map, the easier for the model to discriminate the target from the distracter (corresponding to faster discrimination in the human data). Importantly, we adopted this operationalization of discrimination because it directly draws on the learned representation in the map. (The use of Euclidean distance was motivated by the fact that this metric is used elsewhere in the SOM, cf. Eqn. (3).) Analogously to Winawer et al., the map distances for the 8 'near' and 8 'far' (s_t, s_d) pairs closest to the category boundary were calculated from the model for each simulation. We take the mean $d_{\text{map}}(s_t, s_d)$ value in each of the 4 conditions ('within'–'across' $\times$ 'near'–'far') given a particular combination of model settings, for all simulations that are valid (see below).

We compare the different conditions of the experiment – 'far' vs. 'near', 'across' vs. 'within' – using a log-transformed ratio between the two (mean) distances (i.e., the average over the simulations using all model parameter settings). For example, the 'far'-'near' ratio is defined as the ratio between the mean d_{map} for 'far' pairs and the mean d_{map} for 'near' pairs. We use the $\log_2$ ratio, as this is easier to interpret: e.g., a log ratio of 0 for 'far'-'near' means that there is no difference between 'far' and 'near' distances; values above 0 mean that the 'far'

distances are greater, whereas values below 0 mean that the 'near' distances are greater.

Results on discrimination

Before turning to the numeric results, we note that we found a substantial difference between the semantic feature spaces in the number of valid simulations that we could use in this analysis. Specifically, a total of 15% of the simulations in Russian did not have the category structure over the test stimuli required to simulate the Winawer et al. (2007) experiment.[13] We had to omit such invalid simulations from our analysis. Importantly, more than twice as many `perc` simulations were invalid as `xling` simulations (21% of `perc` vs. 9% of `xling`). This shows a clear advantage of the `xling` semantic space, with which the model more often learns a category space with the expected properties. On the other hand, we found no differences in validity of simulations, nor even in the match to the behaviorial data, when using the two input sampling procedures in the model. Thus we pool the `corpus` and `uniform` settings for each semantic feature space in our reported results.

We describe here how we simulate each of the three main findings of the Winawer et al. (2007) experiment with reference to the log ratios from the valid model simulations. These log ratios are shown in Fig. 9, split out over feature space and language.

First, Winawer et al. (2007) find that 'far' pairs of stimuli are easier to discriminate than 'near' pairs in both languages; that is, people are faster at discriminating stimuli that are perceptually more distant. To achieve this effect in our model, we expect the map distances, d_{map}, for the 'far' pairs to be greater than the distances for the 'near' pairs in both languages – that is, the ratio of 'far' to 'near' should be distinctly higher than 0. Looking at Fig. 9a, we find that the far:near ratio is well above 0 under all circumstances – for both `perc` and `xling`, and for English and Russian. Because we use $\log_2$ ratios, the mean log ratio values of around 1 mean that 'far' pairs are about 2 times as far from each other on the map as 'near' pairs. Because 'far' stimuli pairs are, by design, about twice as far as 'near' stimuli pairs in both feature spaces, this confirms that the model captures the distance in the underlying semantic spaces

[13] That is, in these simulations it was not possible to identify a single category boundary between *sinij* ['dark blue'] and *goluboj* ['light blue'] within the list of test stimuli; rather, there were two or more category switches.

appropriately.

Second, Winawer et al. (2007) find that 'across' pairs are easier to discriminate than 'within' pairs in Russian, but not English. This is the main finding in their paper, which is attributed to the existence of a basic category boundary in Russian but not in English. To model this, we expect the log ratio of d_{map} values for the 'across' pairs of stimuli to 'within' pairs of stimuli to be substantially higher than 0 in Russian (indicating faster discrimination) but not in English. Fig. 9b shows that the across:within ratio for Russian in the model is greater than 0, whereas the ratio is much closer to 0 for English. [14]

Finally, Winawer et al. (2007) note that the above cross-category advantage for Russian holds for 'near' pairs but not for 'far' pairs. That is, the greater ease of discrimination when there is a category boundary between stimuli only holds for the harder ('near') discrimination tasks. In our model, we expect the across:within ratios of the 'near' stimuli to be higher than 0 in Russian, but not those of the 'far' stimuli. As with the across:within ratio overall in English, we expect there that neither 'near' nor 'far' pairs will show a ratio higher than 0. Figures 9c and 9d illustrate that the model captures this effect to some extent. On the one hand for both the 'near' and 'far' pairs, the across:within log ratio is greater than 0, which is contrary to the prediction for the 'far' pairs. On the other hand, the Russian across:within log ratios are higher for the 'near' pairs than for the 'far' pairs, showing that for the 'near' pairs (but less so the 'far' pairs) the model discriminates the 'across' pairs more readily than the 'within' pairs. For both 'near' and 'far' pairs in English, the ratios are close to 0, indicating no difference.

Analysis of Discrimination Results

We find that the model is able to simulate the language-specific effects found in Winawer et al. (2007). We found no difference between the `sampling` options, indicating that while

[14] A reviewer raised the question to what extent the *size* parameter influences the d_{map} values. The across:within ratio for Russian is the only place where we find such an effect: the ratio is on average greater for maps of $size = 7$ than for maps of $size = 10$, which are in turn greater than maps of $size = 12$. This reflects the fact that same-category stimuli are more often projected onto the same Best-Matching Unit in smaller maps than in larger ones, thus making the ratio on average greater. However, this quantitative difference did not change the qualitative pattern under any of the settings of $size$: in all cases the log-across:within ratio for Russian was significantly greater than both 0 and the English log-across:within ratio.

`corpus` sampling leads to a better developmental match, once the model sees enough input to converge on adult knowledge, the structure of this knowledge is not affected in a relevant way by how frequently the stimuli occurred in learning. We also found that for the `features` settings, both semantic spaces showed the crucial differences between Russian and English. The model displays the critical effect of this experiment because in learning, the semantic and term features are jointly projected onto the underlying map representation. This enables the model to display language-specific topologies: the area of the map representing a particular portion of the semantic feature space (here, DARK BLUE and LIGHT BLUE) will differ between two languages according to how the terms carve up the underlying space into lexical categories (here, *sinij* and *goluboj* vs. *blue*). Our earlier Fig. 5 illustrates this: here we can see that the twenty test stimuli are spread out over fewer cells in English ($n = 2$) than in Russian ($n = 3$), and the map distances across DARK BLUE and LIGHT BLUE are thus larger in Russian than in English. The ability of the `xling` space to capture this effect, together with its better match (over `perc`) to child developmental patterns, support the use of a crosslinguistic semantic space for modeling both language acquisition and linguistic relativity data.

Discussion

Our proposed computational model simulates both developmental errors in learning the vocabulary of a semantic domain, and differential effects across languages in the use of such learned semantic knowledge. To our knowledge, this is the first unified model of both child acquisitional patterns and adult linguistic relativity effects in an actual semantic domain (cf. Colunga & Gasser, 1998). Using the domain of color as a testbed, we explore various factors that may influence the observed human behavior: properties of the semantic representational space, the frequency of terms in the input, and the mechanisms by which these interact. In particular, we consider these factors in exploring the following questions with our model: (1) why some linguistic categories – i.e., the associations between a term and a portion of the semantic space – are harder to learn than others, and (2) how learning a language-specific set of lexical categories affects processing in that semantic domain.

In investigating the first question, we focus on children's asymmetric overextension errors – e.g., use of the term *blue* for both BLUE and PURPLE, but not (or less) use of the term *purple* for BLUE. Such errors are revealing about the developing semantic category structure. Our model learns color categories and their name distributions by jointly projecting associated color-term+color-stimulus pairs onto the cells of a neural network (called a Self-Organizing Map Kohonen et al., 2001). Examination of the model's developmental trajectory reveals two factors at play in asymmetric overextension errors: semantic confusability and term frequency. When the color stimuli for two color terms are close in the semantic representational space, they are projected onto close areas of the model's representational map. When terms are sampled uniformly, these two (adjacent or overlapping) areas are similar in strength of representation, leading to interchangeability of the two terms. For asymmetric overextension to occur, one term must be dominant in frequency, enabling it early on to be more strongly represented in the map; only later, given sufficient input, will the less frequent term carve out its own robust representation. Consequently, when color terms are input in proportion to their corpus frequencies (rather than uniformly), our model shows a better match to observed overextension patterns in children. Our model thus provides a mechanistic explanation for the interplay between semantic closeness of the learned categories (e.g., Bartlett, 1978; Pitchford & Mullen, 2003) and term frequency (Yurovsky et al., 2015) in developmental error patterns for color terms.

The second question we explore concerns linguistic relativity – how the learned lexical categories in a domain may influence semantic processing in that domain. Here we consider evidence that having a lexical boundary between two parts of the color space (e.g., two basic color terms, *sinij* for DARK BLUE and *goluboj* for LIGHT BLUE, in Russian) affects the speed of processing stimuli across that boundary compared to a language which does not have the boundary (e.g., a single basic term *blue* for both DARK BLUE and LIGHT BLUE in English). We find in our model that the joint projection of a color-term and color-stimulus together onto the distributed map representation enables the language-specific lexical categories to 'warp' the semantic space. That is, the same portion of the semantic space of stimuli (in this example,

the BLUE portion including DARK BLUE and LIGHT BLUE) will be allocated differently onto the map representation in the model, depending on the need (or not) to make a basic color distinction within that space. Our model then explains linguistic relativity effects (at least, category effects typical of color processing; e.g., Cibelli et al. 2016; Winawer et al. 2007) as arising from an acquisition mechanism that yields language-specific topologies for the same semantic domain.

A clear next step for this research would be to bring these two aspects of investigation closer together, by exploring the developmental trajectory of linguistic relativity effects. A well-known body of work involves children's acquisition of the tight-fit/loose-fit distinction in Korean, and the comparisons of sensitivity to this semantic property across Korean- and English-speaking children (Bowerman & Choi, 2001; McDonough et al., 2003). Modeling such data in our approach would shed further light on the acquisitional mechanisms and resultant knowledge structures discussed above, and their developmental time-course. However, while the model here has the virtue of using a simple architecture that enables identification of general mechanistic underpinnings of its behavior, it is limited to learning words from individual presentations. By contrast, studying the acquisition of relational knowledge, such as that expressed by verbs and adpositions (Gentner & Bowerman, 2009; Majid, Boster, & Bowerman, 2008; McDonough et al., 2003; Saji et al., 2011), will be better situated within a model of word-learning in context. Our plan is to extend our work on cross-situational word learning (e.g., Beekhuizen, Fazly, Nematzadeh, & Stevenson, 2013; Nematzadeh, Beekhuizen, Huang, & Stevenson, 2017) to explore the acquisition of relational semantic domains and associated linguistic relativity behavior.

Consideration of a range of semantic domains also raises the issue of having an appropriate semantic representational space. Indeed, one of our goals in pursuing the present work was to further explore the possibility of deriving an effective distributional semantics from crosslinguistic elicitation data, as we had proposed for the domain of spatial relations (Beekhuizen et al., 2014). In line with Anderson (1982) and the Typological Prevalence Hypothesis of Gentner and Bowerman (2009), our approach assumes that crosslinguistic

agreement in naming semantic situations can reveal the similarity structure among those situations in an underlying semantic space. Intuitively, for example, the more languages there are that label two color chips with the same color term, the more similar those chips are assumed to be, and the closer they are located in the derived vector-based space. An advantage of this approach is that crosslinguistic data may be available for a domain in which an accurate and well-motivated semantic space is otherwise difficult or not currently possible to elaborate (cf. our work on indefinite pronouns, Beekhuizen et al. 2017).

A further advantage of exploring this approach in the domain of color is that there is also a well-accepted perceptual space for encoding color ($L^*a^*b^*$, Fairchild 1998), which enables us for the first time to directly compare an existing, well-understood semantic representation (which we encode here as `perc`), with our crosslinguistic approach (here called `xling`). We find in our model simulations that `xling` performs better than `perc` in accounting for developmental errors, because it better captures the observed semantic confusion between colors. For example, Russian children overextend *sinij* [‘dark blue’] to *goluboj* [‘light blue’] more than might be expected by the perceptual discriminability of the two captured in `perc`, but DARK BLUE and LIGHT BLUE are much closer together in `xling`. On the other hand, for the linguistic relativity simulations, we find that the model trained using either `perc` or `xling` achieves the observed effects. These results together support the idea that our crosslinguistic approach to encoding a semantic space may provide a practical and principled alternative when an existing semantics is not otherwise available, and in fact, may capture cognitive influences – which have shaped the world’s languages – that other approaches do not.

A deeper question remains regarding the crosslinguistic space, namely why some pairs of color chips are categorized with the same term within many languages, whereas others, with the same perceptual distance between the pair, are not. Apparently, factors other than our biologically-grounded system of color perception – such as correlations of (properties of) objects with particular perceptual values in the color space – are relevant to subjects’ naming of color chips (Mitterer, Horschig, Müsseler, & Majid, 2009; Saunders & Van Brakel, 1997;

Wierzbicka, 2005). Our method does not directly address the question of what factors determine the underlying similarity space revealed by crosslinguistic elicitation data, let alone whether these factors are biological, social, or ecological. Crucially, however, by letting languages 'speak for themselves', our method is agnostic with respect to the sources and allows for the derivation of a semantic similarity space without any commitment to a set of underlying features. On the other hand, this approach could be deployed as a discovery procedure for previously unseen dimensions of variation (cf. Majid et al., 2008), and thus has the potential to contribute to the question of the ultimate causes of the similarity space.

References

Anderson, L. B. (1982). The 'perfect' as a universal and as a language-specific category. In P. J. Hopper (Ed.), *Tense-aspect: Between semantics and pragmatics* (pp. 227–264). Amsterdam: John Benjamins.

Bae, G., Olkkonen, M., Allred, S., & Flombaum, J. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General*, *144*, 744–763.

Bartlett, E. J. (1978). The acquisition of the meaning of color terms: A study of lexical development. *Recent Advances in the Psychology of Language: Language Development and Mother–Child Interaction*, 89–108.

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

Beekhuizen, B., Fazly, A., Nematzadeh, A., & Stevenson, S. (2013). Word learning in the wild: What natural data can tell us. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society.*

Beekhuizen, B., Fazly, A., & Stevenson, S. (2014). Learning Meaning without Primitives: Typology Predicts Developmental Patterns. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society.*

Beekhuizen, B., & Stevenson, S. (2015). Perceptual, conceptual, and frequency effects on error patterns in English color term acquisition. In *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning.*

Beekhuizen, B., & Stevenson, S. (2016). Modeling developmental and linguistic relativity effects in color term acquisition. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society.*

Beekhuizen, B., Watson, J., & Stevenson, S. (2017). Semantic typology and parallel corpora: Something about indefinite pronouns. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society.*

Belpaeme, T., & Bleys, J. (2005). Explaining universal colour categories through a

constrained acquisition process. *Adaptive Behavior*, *13*(4), 293–310.

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.

Bowerman, M. (1993). Typological perspectives on language acquisition: Do crosslinguistic patterns predict development? In *Proceedings of the Twenty-Fifth Annual Child Language Research Forum* (pp. 7–15).

Bowerman, M., & Choi, S. (2001). Shaping meanings for language: universal and language-specific in the acquisition of semantic categories. In M. Bowerman & S. C. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 475–511). Cambridge: Cambridge University Press.

Brown, R. W., & Lenneberg, E. H. (1954). A study in language and cognition. *Journal of Abnormal and Social Psychology*, *49*(3), 454–462.

Burton, G., & Moorehead, I. R. (1987). Color and spatial structure in natural scenes. *Applied Optics*, *26*(1), 157–170.

Cibelli, E., Xu, Y., Austerweil, J. L., Griffiths, T. L., & Regier, T. (2016). The Sapir-Whorf hypothesis and probabilistic inference: Evidence from the domain of color. *PloS one*, *11*(7), e0158725.

Colunga, E., & Gasser, M. (1998). Linguistic relativity and word acquisition: A computational approach. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (pp. 244–249).

Davies, I., & Corbett, G. (1994). The basic color terms of Russian. *Linguistics*, *32*(1), 65–90.

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.

Davies, M. (2008-). *The corpus of contemporary american english (coca): 560 million words, 1990-present.* (Available online at https://corpus.byu.edu/coca/)

Dougherty, J. (1978). On the significance of a sequence in the acquisition of basic color

terms. In B. Blount & M. Sanches (Eds.), *Sociocultural Dimensions of Language Change* (pp. 133–48). New York: Academic Press.

Everett, C. (2013). *Linguistic relativity: Evidence across languages and cognitive domains*. Berlin: De Gruyter/Mouton.

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.

Gumperz, J. J., & Levinson, S. C. (1996). *Rethinking linguistic relativity*. Cambridge: Cambridge University Press.

Harkness, S. (1973). Universal aspects of learning color codes: A study in two cultures. *Ethos*, *1*(2), 175–200.

Hendley, C. D., & Hecht, S. (1949). The colors of natural objects and terrains and their relation to visual color deficiency. *Journal of the Optical Society of America*, *39*(10), 870–873.

Howard, C. M., & Burnidge, J. A. (1994). Colors in natural landscapes. *Journal of the Society for Information Display*, *2*(1), 47–55.

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

Kay, P., Berlin, B., Maffi, L., Merrifield, W. R., & Cook, R. (2009). *The World Color Survey*. Stanford: CSLI Publications.

Kohonen, T., Schroeder, M. R., & Huang, T. S. (2001). *Self-organizing maps* (3rd ed.). Berlin: Springer.

Komarova, N. L., & Jameson, K. A. (2013). A quantitative theory of human color choices. *PLoS ONE*, *8*(2), e55986.

Li, P., & Zhao, X. (2013). Self-organizing map models of language acquisition. *Frontiers in*

Psychology, *4*(828).

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah: Erlbaum.

Majid, A., Boster, J. S., & Bowerman, M. (2008). The cross-linguistic categorization of everyday events: A study of cutting and breaking. *Cognition*, *109*(2), 235–250.

Malt, B. C., Sloman, S. A., & Gennari, S. (1999). Knowing versus naming : Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, *262*, 230–262.

Mauri, C. (2008). *Coordination Relations in the Languages of Europe and Beyond*. Berlin: Mouton De Gruyer.

McDonough, L., Choi, S., & Mandler, J. (2003). Understanding spatial relations: Flexible infants, lexical adults. *Cognitive Psychology*, *46*(3), 229–259.

Mitterer, H., Horschig, J. M., Müsseler, J., & Majid, A. (2009). The influence of memory on perception: It's not what things look like, it's what you call them. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1557.

Nematzadeh, A., Beekhuizen, B., Huang, S., & Stevenson, S. (2017). Calculating probabilities simplifies word learning. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society.*

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology*, *13*(1), 87–108.

Pinker, S. (2007). *The Stuff of Thought: Language as a Window into Human Nature*. New York: Viking.

Pitchford, N. J., & Mullen, K. J. (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., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, *104*(4), 1436–1441.

Roberson, D., Davidoff, J., Davies, I., & Shapiro, L. (2005). Color categories: Evidence for the cultural relativity hypothesis. *Cognitive Psychology*, *50*, 378–411.

Roberson, D., Davies, I., & Davidoff, J. (2000). Color categories are not universal:

Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, *129*, 369–398.

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.

Saunders, B. A., & Van Brakel, J. (1997). Are there nontrivial constraints on colour categorization? *Behavioral and Brain Sciences*, *20*(2), 167–179.

Shatz, M., Behrend, D., Gelman, S., & Ebeling, K. (1996). Colour term knowledge in two-year-olds: Evidence for early competence. *Journal of Child Language*, *23*, 177–199.

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

Theakston, A., Lieven, E., Pine, J., & Rowland, C. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, 127–152.

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

Webster, M. A., & Kay, P. (2012). Color categories and color appearance. *Cognition*, *122*(3), 375–392.

Whorf, B. L. (1956). *Language, Thought and Reality*. Cambridge: MIT Press.

Wierzbicka, A. (2005). There are no "color universals" but there are universals of visual semantics. *Anthropological Linguistics*, *47*(2), 217–244.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, *104*(19), 7780–7785.

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

2775–2780). Austin,TX: The Cognitive Science Society.

Appendix A

Model Convergence

We evaluated whether the model is able to obtain an adult level of understanding of the color terms. In prior experimentation, we determined that the model, under any combination of parameter settings, stopped improving in its color term categorization at the latest after around $25,000$ input items. Because of this, we decided to consider $30,000$ iterations to be the 'adult' state of the model. After having trained the model on $30,000$ input items in a given language, we test its color naming behavior for the complete set of training color chips S. To mimic color naming in adults, we extract the most probable term t_s produced by the model for each of the color stimuli $s \in S$, using Eqn. (5) (repeated here as Eqn. ()):

$$t_s = \underset{t \in T}{\operatorname{argmax}}\, P(t|s).$$

We then compare these responses to the adult judgments for these stimuli, and assess the degree of model convergence with adult linguistic behavior using $\mathbf{score_C}$:

$$\mathbf{score_C} = \frac{|S_{\text{correct}}|}{|S|},$$

where S_{correct} is the subset of all stimuli, S, for which $t_s = t_{\text{correct}}$, and t_{correct} is the modal adult response for the given chip. Differences between conditions of interest (`features` and `sampling`) are evaluated using a two-way ANOVA with the $\mathbf{score_C}$ as the dependent variable, and `features` and `sampling` as independent variables. All significant differences are reported and all reported differences are significant to $p = .001$ or less. Table 3 shows the average convergence scores per language for the four combinations of `features = {perc, xling}` and `sampling = {corpus, uniform}`. Under all conditions, the model's predictions match those of adults in $> 90\%$ of the cases, outperforming the naive baseline of always guessing the most frequent term ($\mathbf{score_C} = .20$ for English and .22 for Russian). Despite all conditions doing well, we found small but

significant differences between the settings: the $\mathbf{score_C}$ values were higher for `perc` than for `xling`, and higher for `uniform` than for `corpus`. We failed to find a source of these differences in the error patterns: under all conditions, the model made the same types of errors, just in slightly different quantities. This, combined with the fact that the differences between conditions are rather small ($.02$ for both English and Russian), leads us to believe that the variation does not reflect interesting properties of the feature spaces or sampling effects.

Because we are testing on the same chips as used in training, it could be suggested that this is not a fair method for assessing learning in the model. Indeed, normal practice is to train on one set of data and test on an unseen set. However, several factors led us to test convergence with adult behavior in this way. First, we assume that generally adults are not being asked to label "unseen" colors in a color naming task; rather they are seeing colors they have likely been exposed to with various labels and generating which they think is the best label for that color. Second, our model is not simply memorizing the association of color chips to their dominant labels – it is unable to. Because every input to the model affects the representation of more than one cell in the map, the model cannot directly store exemplars, as in some learning models. Finally, because of the small size of the Russian data set (49 stimuli), holding out stimuli during training make the categorization problem unrealistically difficult, as some color categories have only a few example stimuli labelled with that color term. Hence, we believe the reported set up is a reasonable approach for simulating adult color naming experiments given the available data.

Nonetheless, to assess the ability of our model to generalize to unseen data despite the small data sets, we ran the simulations again in a leave-one-out cross-validation procedure. For each stimulus $s_i \in S$, the model was trained on input items sampled from the set of all stimuli except s_i (i.e., trained on $S \setminus s_i$). In particular, we trained the model for each of the four combinations of `features` and `sampling`, with one set of model parameters that we found to perform particularly well on the experiment with non-held-out data ($\alpha = .5$, $a = .5$, and $size = 12$). For each left-out stimulus s_i, we ran five simulations. We do not report the full results of all these follow-up experiments here, but simply summarize. Crucially, we found

that for English, the model reached convergence scores of $\geq .90$ under any combination of feature set and sampling method. For Russian, however, the scores dropped to around $.70$ (with a reversal of the pattern of better performing feature sets, and a larger difference: `xling` scored on average $.73$, versus $.69$ for `perc`). We believe the lower performance in Russian is due to the data scarcity: many of the color categories are only represented by 3 or 4 exemplars, which makes the acquisition of a generalizable color category more tentative.

Table 1
Basic color terms for English and Russian with corpus frequencies

English			Russian			
color term	abbre-viation	relative frequency	color term	abbre-viation	gloss	relative frequency
red	re	.20	*belyj*	bel	'white'	.22
blue	bu	.19	*chernyj*	che	'black'	.22
green	gn	.16	*krasnyj*	kra	'red'	.20
yellow	ye	.13	*zelenyj*	zel	'green'	.09
white	wh	.09	*sinij*	sin	'dark blue'	.06
orange	or	.07	*goluboj*	gol	'light blue'	.05
pink	pi	.05	*seryj*	ser	'grey'	.04
black	ba	.05	*zheltyj*	zhe	'yellow'	.04
brown	br	.04	*rozovyj*	roz	'pink'	.04
purple	pu	.02	*korichnevyj*	kor	'brown'	.02
grey	gy	.01	*oranzhevyj*	ora	'orange'	.01
			fioletovyj	fio	'purple'	.01
uniform $P(t)$		.09	uniform $P(t)$			.08

Table 2
Results for development: Mean $\mathbf{score_D}$ *(worst–best) across all simulations*

	English		Russian	
	`corpus`	`uniform`	`corpus`	`uniform`
`perc`	.75 (.62-.87)	.74 (.54-.90)	.74 (.68-.79)	.64 (.56-.72)
`xling`	.78 (.66-.89)	.73 (.64-.83)	.78 (.66-.83)	.71 (.61-.77)
error-free learner	.80		.63	

Table 3
Results for convergence: Mean $\mathbf{score_C}$ *(worst–best) across all simulations*

	English		Russian	
	`corpus`	`uniform`	`corpus`	`uniform`
`perc`	.96 (.89-.99)	.97 (.90-.99)	.96 (.88-.99)	.96 (.93-.98)
`xling`	.95 (.88-.99)	.96 (.90-.99)	.94 (.88-.99)	.94 (.89-.98)
frequency baseline	.20		.22	

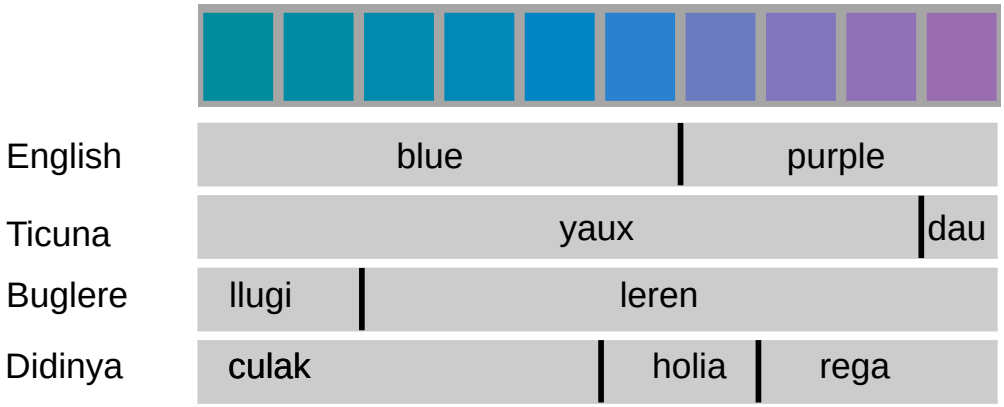


Figure 1. Categorization of 10 Munsell color chips (F25-F34) in four languages. Data from Berlin and Kay (1969) and Kay et al. (2009).

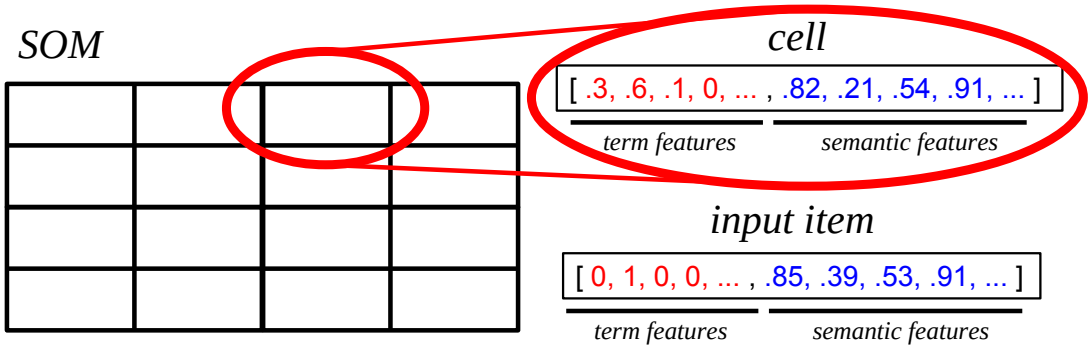
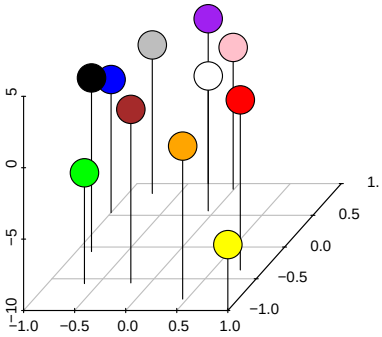


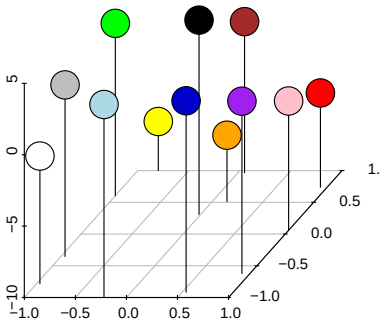
Figure 2. An example of a Self-Organizing Map and an input item.

	Color terms in all languages														
	l_1					l_2					l_n				
Color chips	t_1	t_2	t_3	t_4	t_5	t_1	t_2	t_3	t_4	$\dots$	t_1	t_2	t_3	t_4	t_5
s_1	.8	.2	0	0	0	0	1	0	0		0	.9	0	0	.1
s_2	1	0	0	0	0	0	.7	0	.3		0	.6	0	0	.4
$\vdots$										$\ddots$					
s_n	0	0	0	0	1	0	.2	.8	0		0	0	.6	.4	0

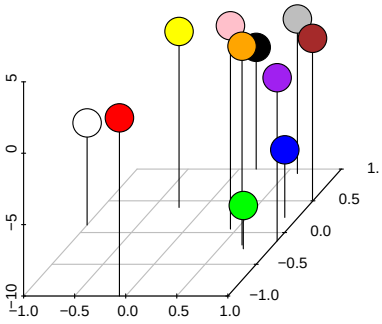
Figure 3. An example of the concatenated probability distributions $P_l(\cdot|s_i)$ for all terms t in all languages l, where each s_i is a color chip.



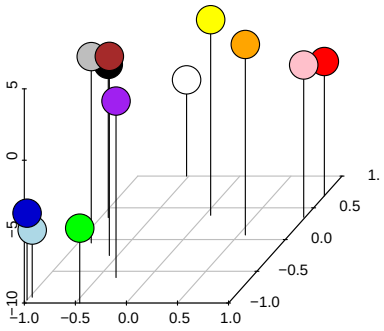
(a) English in the `perc` space



(b) Russian in the `perc` space

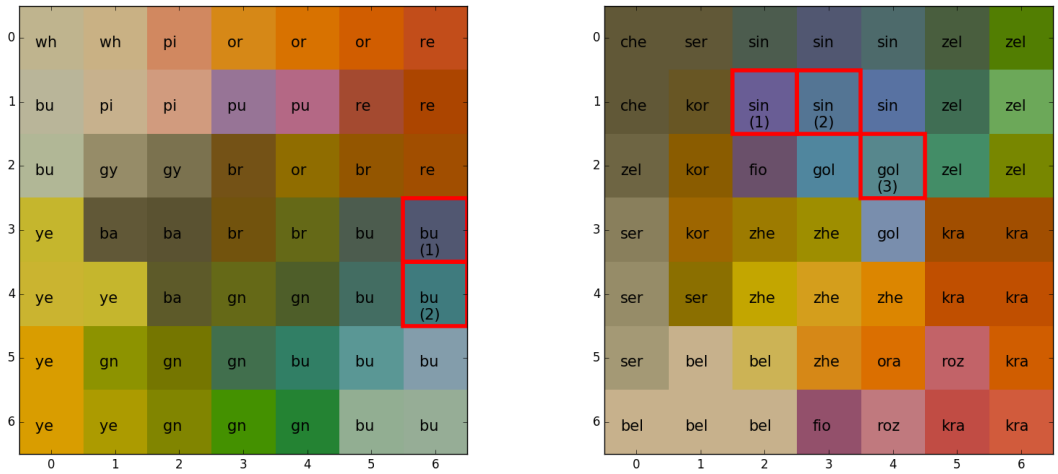


(c) English in the `xling` space



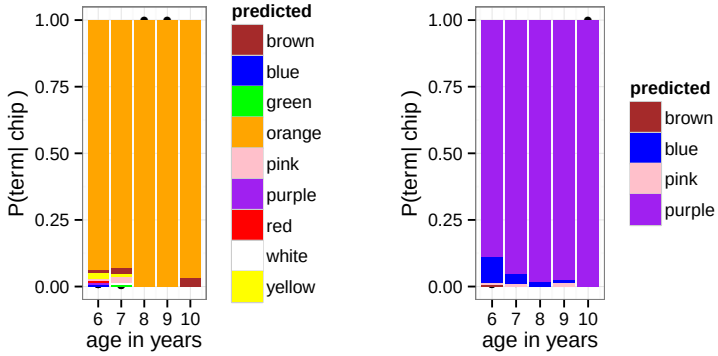
(d) Russian in the `xling` space

Figure 4. Category centroids of training stimuli for the two languages in the two feature spaces, as visualized by a 3-dimensional Multidimensional Scaling (MDS) solution.



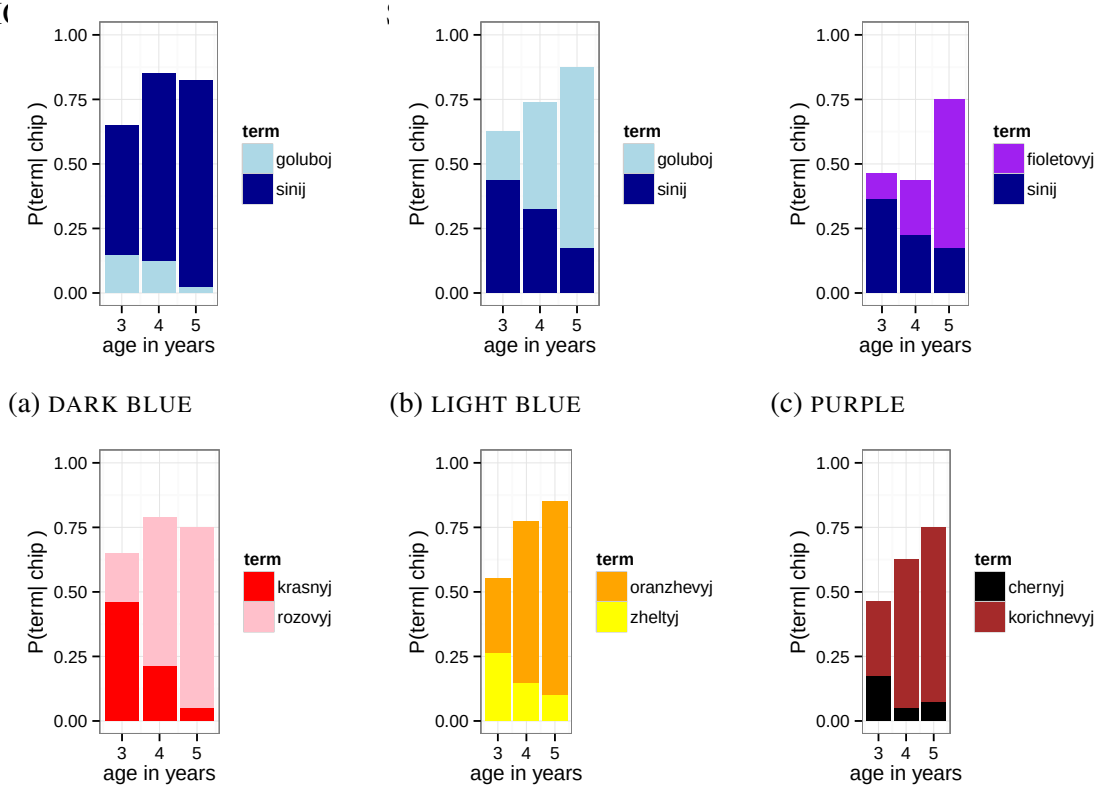
(a) English (b) Russian

Figure 5. Two trained 7×7 Self-Organizing Maps. The displayed color in each cell is an approximation in RGB of its feature space coordinates; the two or three letter codes represent the most likely color term for the cell (see Table 1 for codes). Red squares around a cell and numbers in parentheses indicate the cells onto which the stimuli of the discrimination experiment are projected.



(a) ORANGE (b) PURPLE

Figure 6. Probability of color term responses to ORANGE and PURPLE, based on counts reported in Bateman (1915).



(a) DARK BLUE (b) LIGHT BLUE (c) PURPLE

(d) PINK (e) ORANGE (f) BROWN

Figure 7. Probability of color term responses to six color stimuli; based on counts reported in I. Davies et al. (1998).

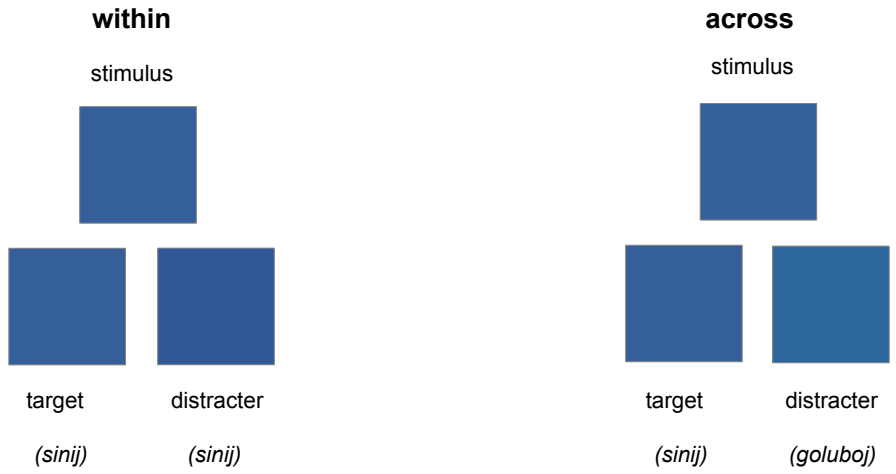
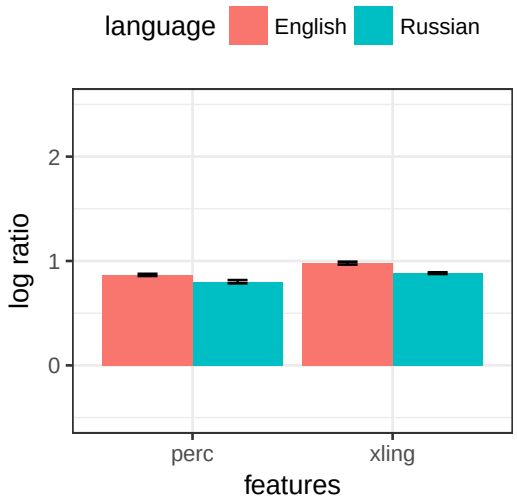
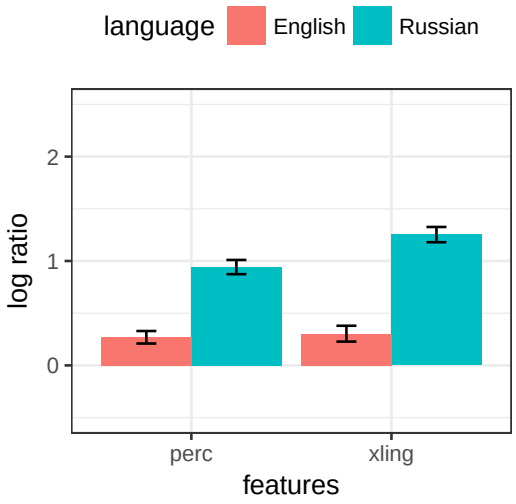


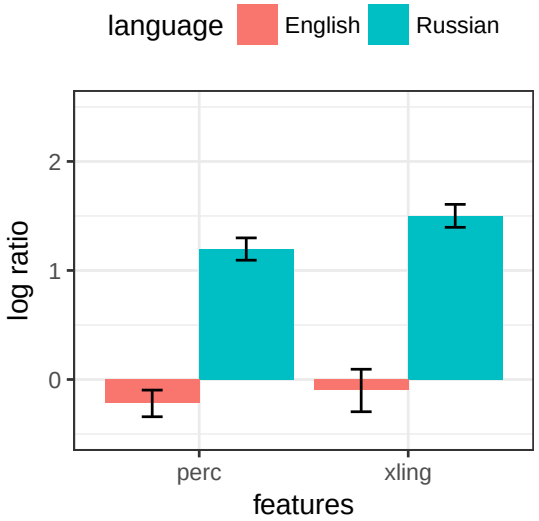
Figure 8. Two examples of a triplet with a stimulus, an identical target, and a different distracter from the Winawer et al. (2007) color discrimination experiment. Both distracters are at the same distance ('near') from the stimulus; the distracter in the left example is 'within' the same category (both are named as *sinij* in Russian), while the right one is 'across' the category boundary (the stimulus/target is named *sinij*, the distracter as *goluboj*)



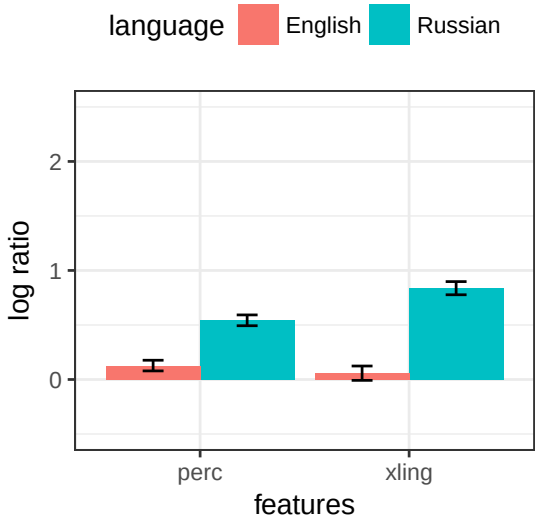
(a) Log-ratio of ‘far’ to ‘near’ distances per language and feature set.



(b) Log-ratio of ‘across’ to ‘within’ distances per language and feature set.



(c) Log-ratio of ‘across’ to ‘within’ distances per language and feature set for ‘near’ pairs.



(d) Log-ratio of ‘across’ to ‘within’ distances per language and feature set for ‘far’ pairs.

Figure 9. Model predictions for the two main conditions and interaction in the Winawer et al. (2007) experiment: results are shown as log ratios, where a value of 0 indicates no difference between the compared values.