Inferring Color Terms

Zhenbang Chen
Massachusetts Institute of Technology
zhenbang@mit.edu

Zhenjia Chen
Massachusetts Institute of Technology
zhenjia@mit.edu

Abstract

Within the context of naming colors, there is a clear balance between accuracy, name specificity, term prevalence, and communicative cost. This paper will detail a probabilistic model for predicting color names that outperforms an approach based strictly on term accuracy. The results provide insight into more general naming schemes for broader concepts beyond color.

1. Introduction

In a general scope, language supplies a communicable representation of concepts and ideas. Often this mapping between representations and concepts is not one-to-one; synonyms and hierarchical definitions means that concepts are often paired with many representations in language. This gives rise to the research question: to what extent is there a tradeoff between specificity, concept frequency, name prevalence, and representational complexity in identifying concepts with language? An example could come from color descriptions, where the concepts of red, crimson, and Nadeshiko pink correspond to increasingly precise bands of the color domain, lower natural occurrence frequencies, and larger word representation size/complexity. Relevant background knowledge about language and idea prevalence is needed to answer this question. This paper explores this broad question through the example problem of color name inference.

2. Related Works

As Berlin and Kay described in their *Basic Color Terms: Their Universality and Evolution*, a popular fundamental problem in cognitive science is the question of how humans compartmentalize and describe the color spectrum [1]. Their work on the evolution of color names in languages supplies excellent domain specific background.

Additionally Charles Kemp and Terry Regier’s research publication, *Kinship Categories Across Languages Reflect General Communicative Principles*, partially inspired the research topic for this paper [2]. Their work describes the interplay between language, communication, and information.

3. Approach and Model

3.1. Behavior

To explore the research question and analyze relationships between phrase specificity, representation costs, and concept frequency, using colors as a detailed example case, our project aims to develop a model that predicts a reasonable natural color name given a specific color value. More specifically, when presented with a visual example of a color, a person may associate a name to describe the color. The name choice could depend on several factors, including the appearance of the color (which could be represented numerically), and the accuracy, preciseness, complexity, and relative frequency in natural language of various name candidates.

Here is a simple example to illustrate the phenomenon we are attempting to model. Consider a person assigning color terms to the three squares in Figure 1.

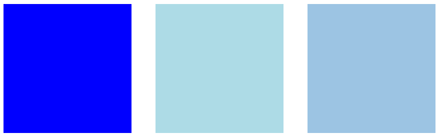

Figure 1. Colored squares of different shades of *blue*. RGB values from left: (0, 0, 255), (173, 216, 230), (155, 196, 226). The right square has the exact coordinates of *pale cerulean* in RGB space.

Given these visual stimuli of color and a choice of a single color phrase to associate with each square, it is likely that a person would choose *blue* for all three squares. This outcome is indicative of the breadth (or impreciseness), simplicity, and prevalence of the term *blue*. Another very plausible naming configuration could be designating the left square as *blue* and the other two as *light blue*. *Light blue* could be a preferable name for certain visual colors since it offers a more precise description of those colors while adding a simple modifier and still being a relatively common phrase.

Conversely, a naming order of *blue*, *light blue*, and *pale cerulean* may be intuitively less likely for a few reasons. Even though the rightmost square has the exact color coordinates of *pale cerulean* in RGB color space, the name's low relative usage frequency and slightly increased representational complexity prevents *pale cerulean* from being a likely choice despite its accuracy. Nevertheless *pale cerulean* is still a more likely choice than *red* due to its superior accuracy and preciseness even though *red* has low representational complexity and occurs extremely frequently in prior probability terms. This illustrates the balance between different cognitive factors and considerations involved in determining an optimal name for describing concepts such as color.

3.2. Model Design

During our brainstorming and design process, we first recognize the components that were essential to consider for the Bayesian model. This included color frequency in natural images and in the English language, which both would affect how likely someone will identify a specific color value as a certain natural color name. It is important for the model to reflect a color term's breadth of description, its accuracy, and its overall prevalence.

3.3. Color Spaces

There are several models to represent color. Common ones include RGB, HCL, and CIELAB. These different color spaces reflect different goals and design choices in their implementation. RGB space is a simple additive defined by the three primary colors. As a color space, RGB can be represented as Cartesian coordinates a three dimensional Euclidean space with the positive axis corresponding to the values of *red*, *green*, and *blue*. Other color models, such as HCL, are more adeptly represented by cylindrical models. One variant of HCL, HSL stands for hue, saturation, and lightness. This color model is designed to be more aligned with human perception as compared to RGB. Similarly, the CIELAB color model is also representative of human vision systems.

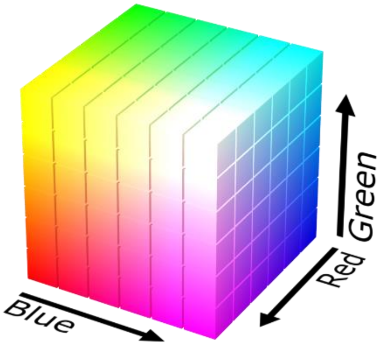

Figure 2. The RGB color space is a cube with similar colors occupying close Euclidean coordinates. Color distance in this space correlates positively with perceived dissimilarity in color. [7]

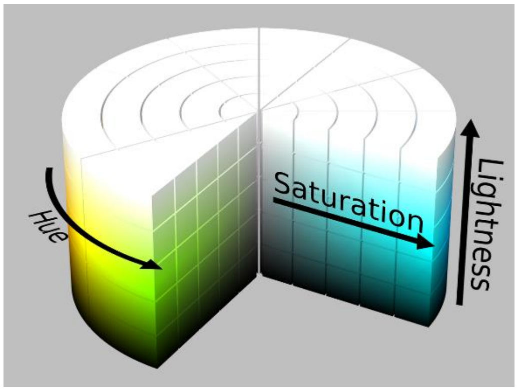

Figure 3. The HSL color space is a cylinder with degenerate HSL values for white and black [7]

Our model chooses to use the RGB space despite some other models being closer to human perception due to its simplicity and the model's natural grouping of similar colors in its color space. Conversely, the HSL model, when projected as a cylindrical space, has an entire plane of hue and saturation values that all represent white and black when the value component is 1 and 0 respectively.

3.4. Color Names in English Corpus

The next critical step was obtaining data on the prior probability of names for colors. Our first attempt involved searching for existing research on the frequency of color term usage in the English language. Unsurprisingly, we came across the classic color cognition study of Berlin and Kay who outlined a set of eleven basic monolexemic color terms present in nearly every modern culture [1]. Their research demonstrated that most modern languages underwent a similar evolutionary trajectory in adopting names for colors. In particular, *white* and *black* always emerged first, followed by *red*, then *yellow* or *green*. These colors were then succeeded by *blue, brown, purple, pink, orange,* and *gray*.

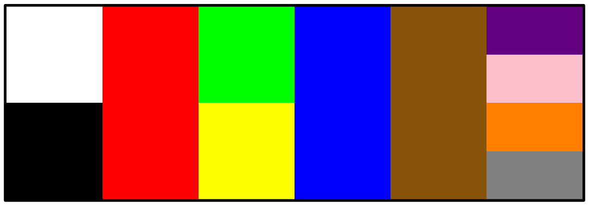

Figure 4. Six stages of the evolution of basic color names according Berlin and Kay's work [1]. In most languages, names emerge in order from left to right.

While this ordering provides some intuitive qualitative insight into the relative prevalence of common basic color terms in language, it is insufficient for building up the prior probabilities for a name prediction model. For frequency statistics on color names, we decided to consult corpus data through the PhraseFinder API. Developed by Martin Trenkmann, PhraseFinder enables searches on the Google Books Ngram Dataset [4][5]. We choose this tool since it allows queries of compound or polylexemic colors names like *dark green* or *yellow orange*. A set of names was compiled from various standard filtered color palette collections, such as HTML color keywords.

Color Names	Occurrences (millions)	Computed Prior	Language Stage
white	99.1	0.2038	1
black	78.7	0.1620	1
red	53.5	0.1101	2
green	37.8	0.0778	3
brown	35.0	0.0721	5
blue	33.5	0.0690	4
gold	31.0	0.0639	~
silver	20.8	0.0427	~
yellow	20.3	0.0419	3
gray	18.1	0.0372	6
orange	9.5	0.0196	6
pink	6.7	0.0138	6
purple	5.9	0.0122	6

Figure 5. Top 13 most frequently used color terms with data from Google Books Ngram Dataset. Excluding *gold* and *silver*, all top names are basic color names. Note that the occurrences ranking generally qualitatively match the language stages from Berlin and Kay's research [?].

After aggregating frequency data for 416 color names, we found that all eleven of Berlin and Kay's basic color terms appear in the 13 most frequently used color phrases. Additionally, the occurrences rankings generally qualitatively match their proposed language development stages. While some names have definitions outside of color (e.g. *gold*, *orange*, and *lavender*), this provides additional assurance that the corpus data roughly reflects the usage frequency of color terms explicitly in the context of describing color. We also ignored more extreme examples of color terms that were very likely to be used in contexts outside of color, such as *cream*, *olive*, and *rose*.

$$P(c_i) = \frac{o_i}{\sum_j^n o_j}$$

Equation 1. Prior probability of each color term c_i defined as the quotient of its occurrences o_i and the sum of occurrences for all color terms.

For each of the 416 color terms, we now define its prior probability as simply the quotient of its occurrences and the sum of all color name occurrences. Interestingly, we found that the top 13 colors names comprised 92.6% of all color term usage in the corpus.

3.5. Color Frequency Data

Following the prior data, the next crucial component of the Bayesian model is the likelihood probability distribution. For our model, we decided to sample RGB color values from natural images. More specifically, the RGB frequency is aggregated across a 100,000 image subset of the Places dataset and used to compute a probability for each RGB value [6]. The dataset was chosen for its varied content that is reasonably representative of color in the real world.

Figure 6. Example images of the Places dataset subset [6]

Given this probability distribution and the corpus frequency data, we define a "radius" for every color. This radius is the model's representation of that color's domain. A larger radius for a certain color indicates that particular color has a high breadth and less precision. From this radius, a likelihood function for each color name is then computed as a 3d gaussian centered on that color's canonical RGB value (e.g. (0,0,0) is the mean for white). The 3x3 covariance matrix of the gaussian is defined as scalar matrix where the elements are a positive function of the aforementioned radius. A larger radius necessarily means a larger covariance.

The radius is dependent on the RGB value probability distribution and the color corpus probability. It is defined as the radius of the sphere centered on a color's RGB value such that the total RGB value probability enclosed by the sphere is equal to that color's corpus probability.

Given a radius for each color, the model then has to make a choice of the relationship between said radius and the elements of the covariance matrix for that color's normal probability distribution function. To this end, we experimented with several different functions including linear, polynomial, logarithmic, and sigmoidal relationships.

After testing, we found that a simple linear function works well, especially for general cases. Choices for this relationship affect final predictions by changing how much larger radii impact the probabilities. Logarithmic and sigmoidal relations essentially caps the influence of extremely common colors such as white and black.

Intuitively, a high corpus probability for a certain color can be explained as either due to that color's wide breadth or its frequency in the real world. In other words, a color name may be common because it covers a large range of RGB values or because the values that it occurs frequently in nature.

By defining the radius this way, the likelihood incorporates both the RGB probability and the corpus probability in a way that is reflective of that intuition. For example, colors that have high corpus frequency but low RGB probability in the vicinity of its canonical RGB value would have a large radius, which corresponds to a larger covariance of its gaussian. This means it probability distribution is more spread out, implying more impreciseness.

3.6. Representation Cost

Representation cost of color refers to some metric of name complexity such as name length or word/syllable count. Ideally, the model should penalize complex color names such as *pale cerulean* in favor of simpler names such as *light blue* even if the name with the higher representation cost is more accurate in terms of actual RGB value.

After gathering corpus data for color name frequency, it is clear that most names are already monolexemic or bilexemic. Color names of three or more words are rare in practice and effectively ignored by the corpus data aggregation. This means that complex color names are naturally less common in the corpus and therefore have a lower prior probability. In this way, representational complexity is already encompassed in the relative frequency statistics.

3.7. Combining Components

After compiling or addressing all necessary components, we finally combine them into a model for inferring natural color names. The model's premise is simple: given an observation of a color, predict a ranked set of corresponding color terms that a person would assign to that color. Formally, we apply a Bayesian inference model using the compiled data on color term prior probabilities and their associated likelihood probabilities.

$$P(c_i \mid v) = \frac{P(v \mid c_i) \times P(c_i)}{P(v)}$$

Equation 2. Posterior probability of a color term c_i given a specific color value v.

As previously discussed, we first formulate the prior probability of each of the 416 color terms from the relative frequency data aggregated from the English corpus. Intuitively, this is the probability of the color term without considering an observed color value.

For an observed RGB color value, *v*, the likelihood of that value for a given color term conveys the probability that the name would give rise to that specific color value. This amounts to a probability distribution sample from the multivariate normal distribution previously defined for each color term.

A key tunable parameter in our model is the value of the scalar covariance matrix used for the likelihood distributions. As mentioned earlier, to create a more accurate model, we experimented with different functions that mapping a color terms' computed radii to its covariance matrix. Intuitively, this mapping should be an increasing function over radius. We experimented with logarithmic, polynomial, sigmoidal, and linear relations. Ultimately, we found that a linear function performed reasonably well.

Lastly, after computing the posterior probability of each color term hypothesis, we return the ranked results as the model's predictions for color names given an RGB color value.

4. Evaluation and Results

4.1. Color Predictions

For evaluating our model, we compare its top predictions against the most common human responses when identifying color. For the survey, we gathered 7 responses after asking participants to identify the color of six different colored squares. The respondents were allowed and encouraged to name multiple colors for each square.

Overall, the top choices for each color generally appeared among top predictions from our model. Specifically, In five of six tests, our model predicted the same top color term as the survey responses. In total, 75% of the top-2 survey responses for each square also appeared in the top-2 term predictions for each color value. Considering the top-3 responses for each square, this accuracy is roughly 71%. For comparison, a simple nearest neighbor approach for the same colors yields 17%, 17%, and 18% accuracy for top-1, top-2, and top-3 metrics. Intuitively, the predictions of nearest neighbor algorithm represents the results when only considering the accuracy of each term in isolation.

Model	Top 1	Top 2	Top 3
Bayesian	83%	75%	71%
Nearest Neighbor	17%	17%	18%

Figure 7. Results of the color model compared to the nearest neighbor

4.2. Limitations

This model of color interpretation has several limitations stemming from the initial choice of an RGB color space, the symmetrical probability distribution for likelihood, and the natural ambiguity of color names in the corpus.

The RGB color space is not designed to be aligned with human visual systems. This means that colors perceived to be similar, such as *blue* and *cyan*, are actually quite far apart in RGB space. As a consequence, when given a color close to *cyan*, the model rarely presents *blue* as a top prediction despite it being a frequent choice when surveyed.

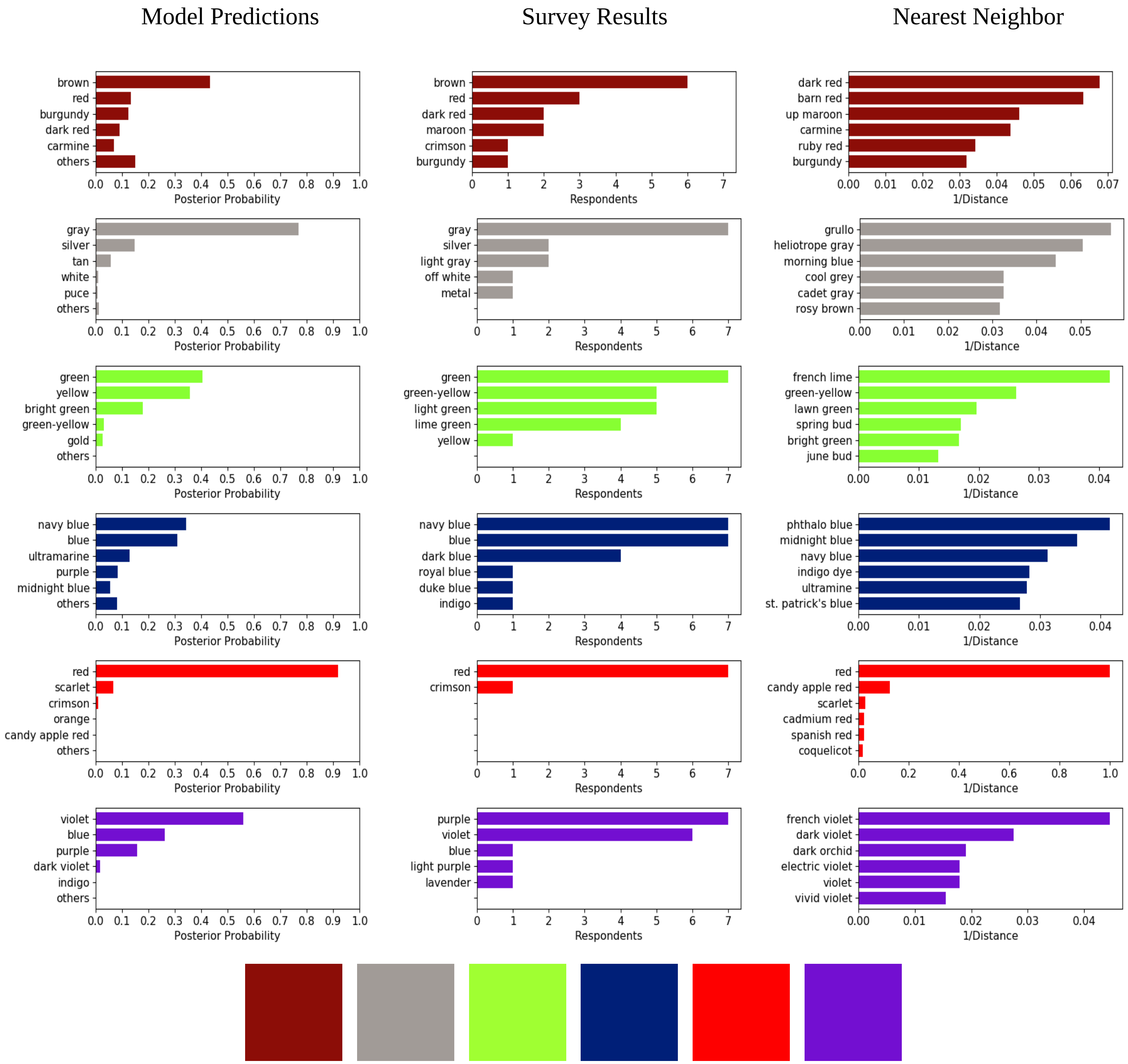

Figure 8. The results of our model on six color values. Tested color values in RGB from left: (140, 13, 7); (161, 155, 151); (135, 255, 50); (0,31,120); (254,0,0); (115, 15, 209). The left column shows the posterior probabilities of color terms produced by our model when given these color values. The middle column shows the responses of 7 people that were presented with the visual color and asked to identify the color. They were allowed and encouraged to name multiple colors. The right column shows the naive color term predictions of a simple nearest neighbor ranking as a baseline comparison.

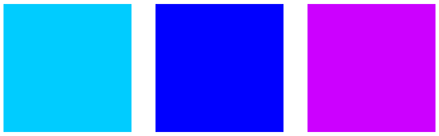

Figure 9. These color squares have RGB values of (0, 200, 255), (0, 0, 255), and (200, 0, 255). The left and right squares are equidistant from *blue*, the middle square, in RGB space.

Another limitation of the model lies in the symmetric nature of the 3d gaussians that comprise the likelihood probability distribution. In practice, there are colors that are significantly different from each other despite being equally close in RGB space. Consider Figure 9, which show two different color squares alongside *blue* in the middle. Although the left and right color squares are the same distance from *blue* and therefore have the same likelihood probability in this model, the left square is perceived to be more similar to *blue* than the right square.

Among the colors name, a few are overrepresented in the corpus. These names have multiple meanings, which allows them to be used in a context that does refer to color at all. Words such as gold and silver may be used both as colors and as minerals. This means that these colors have inflated frequencies in the corpus and therefore inflated probabilities in their prior probability distributions. As a result, these color names are more likely to be predicted by the model in spite of lack of similarity between the RGB value and the color name.

5. Conclusion

Concept naming is a popular and fundamental area of research in cognitive science that provides insight into the interplay between language and ideas. Based on the results for color naming, the application of names to concepts appears follows a probability driven model where there is a tradeoff between accuracy, precision, and term prevalence. This could be further extended to other contexts of identification, such as for occupations or for persons.

6. References

[1] Berlin, Brent and Paul Kay. Basic Color Terms: Their Universality and Evolution. *University of California Press*, 1969.

[2] Kemp, Charles, and Terry Regier. Kinship Categories Across Languages Reflect General Communicative Principles. *Science Vol 336*, 1049-1054, 2012. https://science.sciencemag.org/content/sci/336/6084/1049.full.pdf

[3] Loreto, Vittorio, Animesh Mukherjee, and Francesca Tria. On the Origin of the Hierarchy of Color Names. http://cse.iitkgp.ac.in/~animeshm/hierarchy_evolang.pdf

[4] Trenkmann, Martin. PhraseFinder. Accessed Nov, 2019. https://phrasefinder.io/about

[5] Davies, Mark. Corpus of Contemporary American English. Accessed Nov, 2019. https://www.english-corpora.org/coca/

[6] Zhou, Bolei, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. http://places2.csail.mit.edu/index.html

[7] Datumizer. Wikimedia. Accessed Dec, 2019. https://commons.wikimedia.org/wiki/User:Datumizer