Towards a Computational Model of Human Word-Color Associations

Sonia Krishna Murthy

Advisor: Thomas Griffiths

Submitted to Princeton University
in Partial Fulfillment of the Requirements
for the Degree of Bachelor of Arts

May 2020

Abstract

Computational models of multimodal associations can help us to better understand the ways different domains of human knowledge and experience interact with and supplement each other. Natural language and color are two domains of particular interest, as they are both integral to our experience of the world and are powerful communication devices in their own rights. However, current computational models of human word-color associations attempt to bring the color domain closer to the distributional semantic domain by treating color as a lexical entity like any other target word. My work aims to preserve the rich information contained in human beings' experience of color by maintaining color as a perceptual experience tied to some underlying understanding of word meaning. I first establish a dataset of human color annotations for words that represent varying degrees of abstractness and emotional content. Then, I develop three computational models that are grounded in color data: a distributional semantic model, an image analysis model, and finally, a Bayesian representativeness model. I find that the Bayesian representativeness model is best able to discern meaningful structure in input color data, which allows it to most closely emulate humans' psychological color associations for words.

Contents

1 Introduction & Related Work

Humans' ability to acquire knowledge about the world at massive scales that far exceed their lived experiences has intrigued philosophers and scientists as early as Plato (Landauer & Dumais, 1997). A similar question also arises with regards to humans' natural language abilities: despite only having direct experience with a small subset of an infinite number of possible utterances, humans are constantly able to make sense of spoken and written language without deliberate effort. For decades, this phenomenon has inspired computer scientists to develop models of natural language that capture underlying semantic structures and simulate salient aspects of humans' acquisition of word meaning. While models of natural language, particularly distributional semantic models (Landauer & Dumais, 1997; Griffiths, Steyvers, & Tenenbaum, 2007; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Pennington, Socher, & Manning, 2014), have become more robust in recent years, these models ultimately only derive their semantic information from text, which neglects the breadth of perceptual sources that inform humans' language abilities. This desire to ground natural language models in richer, more human-like sources of semantic knowledge have prompted interest in multimodal approaches to natural language processing (Bruni et al., 2014). However, the project of modeling multimodal associations introduces significant complexity: not only must such models capture the nature of both sensory domains individually, but they must also provide a way of interfacing between their different representations in a psychologically accurate way.

Of particular interest is the domain of color, as it has long been used a means of communication in its own right (Riley, 1995). Color also brings into question the way people process the visual input stream of the world and the emotions and personal experiences they associate with these inputs. As a visual artist, word-color associations intrigue me because of their potential to reveal an intuitive "visual language" of color that is grounded in the most salient aspects of our innate natural language abilities. As a computer scientist, formal representations of something as uniquely experiential as color intrigue me as a way to explore the involvement of human emotion and memory in artificial intelligences.

There are reasons to believe that there exist meaningful connections between words and color, in particular, connections mediated by emotion. The Ecological Valence Theory (EVT) of color preference posits that "a person's preference for a given color is determined by their combined valences (liking/disliking) for all objects and entities that person associates with that color" (Schloss & Palmer, 2017). A different survey asked 40 undergraduate students asking about their favorite colors, the major color they were wearing, their emotional responses to colors, and the reasons for their choices (Hemphill, 1996). Researchers found that only 15% of participants expressed no emotional response to color and the remaining 85% attributed bright colors (white, pink, red, yellow, blue, purple, green) to positive emotions (happy, excited, relaxed), and dark colors (brown, black, gray) to negative emotions (anxious, boring, sad). These findings suggest that color exists in a complex network of memory, emotion, and perceptual experience. However, tapping into this network and directly mapping connections between color and emotion can be difficult, so we must rely on language to articulate these associations. While we cannot easily impart our visual experience of the world directly onto someone else, we can use language to help us inch towards sharing our reality with others. It seems reasonable then that the visual features of our world, of which color is especially important and pervasive, would foster links between color and words.

Because both color and language are so ubiquitous in our daily lives, the potential benefits of developing a comprehensive, psychologically-grounded resource that captures human word-color associations are far-reaching and span many domains. Such a resource is most promising for students with dyslexia or other learning delays. Children remain very visually receptive even when reading and writing is difficult for them. Coloring words based on psychologically-grounded associations of colors and words could help to facilitate their learning and language abilities by triggering both the visual memory and the verbal memory (Ozbal et al., 2011). Marketing and advertising have also used the emotional power of color to capture attention and forge long-lasting connections with brands and products (Labrecque & Milne, 2012). A comprehensive, psychologically-grounded concept-color resource would advance the color theory used by these fields. In the more distant future, such a word-color resource could be used as a kind of translational dictionary between art and language. One could visually represent the gist of a text by creating a "pixel-by-pixel" painting from the words in a passage or conversely, analyze the colors of works of art and use a mapping of words and colors to create a textual companion that "represents" the artwork.

In this work, I begin by collecting and analyzing human color annotations for a set of words to establish a source of color data. I also use this data to determine the existence of color associations for abstract words, or words whose semantics lie in mental concepts, and words devoid of any clear emotional content. I then develop and evaluate three different computational models for representing human word-color associations and predicting a psychologically and perceptually salient color for a given word. The first is a distributional semantic model that aims to produce a mapping between color space and GloVe semantic space (Pennington et al., 2014). A second image analysis model implements a more automatic way of obtaining word-color annotation data with the goal of evaluating other sources that may inform humans' word-color associations. A final Bayesian representativeness model determines which word a particular color point in LAB space should belong to based on its representativeness score (Tenenbaum & Griffiths, 2001)

2 Human Word-Color Annotation Data

2.1 Background

Q2. What colour is associated with *sleep*?

- black
- blue
- brown
- green
- grey
- orange
- purple
- pink
- red
- white
- yellow

Figure 1: Example of survey question presented to participants in Mohammad's (2011) distributional semantic model for measure word-color associations.

Previous studies of word-color associations have attempted to bring the color domain closer to the better formalized semantic domain by treating color as a lexical entity just like any other target word. Ozbal et al.'s (2011) language model method, for example, captured the likelihood of each of the 11 colors modifying the target word based on frequency counts found in the Google Web corpus of English bigrams. Their LSA method used the classic Latent Semantic

Analysis (LSA) (Landauer & Dumais, 1997) framework by treating the colors as words in the word-document input matrix. Word-color associations between the colors and target words were then determined based on spatial proximity within the resulting LSA space. Mohammad (2011) followed a similar data collection procedure by asking participants questions like those in Figure 1. These approaches fail to capture the complexity of an experience of color that is rooted in the way people process the visual input stream of the world and the emotions and personal experiences they associate with these inputs.

My work aims to preserve the rich information contained in human beings' experience of color by maintaining color as a psychological and perceptual experience tied to some underlying understanding of word meaning. I am also interested in understanding the nature of color associations for abstract words, or words whose semantics lie in intangible, mental concepts that are also devoid of any clear emotional content. For concrete words—words that do have a real-world reference—one would expect the color association for the word to reflect something about the visual experience involved: "apple" might be associated with red or green because most people have experience with apples of these colors. However, with abstract words it is less clear why there should be any consistent color associations, yet many would agree that "sad" is blueish/grey and "angry" is red. One hypothesis is that there are strong emotional associations to these abstract words that are mediated by some real-world reference. For example, rain clouds are typically used to depict sadness and our faces become red and flushed when angry. However, do people still agree upon color associations for words without clear emotional content or real-world reference, like "verb" or "noun" or "big"? To answer these questions and create a foundation for future word-color association models, I first create a dataset of human word-color annotations.

2.1.1 CIELab Color Space

While color data is most conveniently obtained using an RGB, hue, saturation, value (RGB-HSV) color space, this color space is known to be a poor proxy for color complexity as perceived by most organisms (Hill, Roger, & Vorhagen, 1997). Additionally, RGB-HSV space is device dependent, meaning that the colors produced in this space depend both on the equipment used and the system settings used therein (Bora, Gupta, & Khan, 2015). To more accurately, consistently, and reproducibly represent the way the human eye perceives color, I interact with the color domain through the three-dimensional CIE L*a*b* color Space (Figure 2).

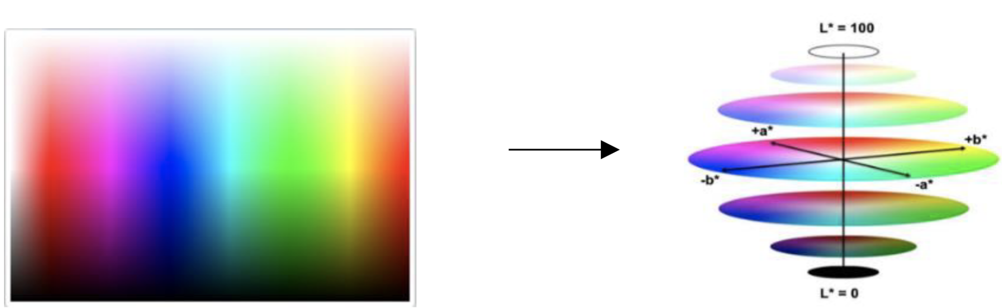

Figure 2: RGB HSV color space (left) vs. the perceptually uniform CIE L*a*b* Color Space (right).

Developed in 1976 by the International Commission on Illumination (CIE), the LAB space intends to allow for "comparisons of differences between object colors of the same size and shape, viewed in identical white to middle-grey surroundings." As a result, Euclidean distances in this space can be used to approximately represent perceived magnitude of color differences (Schanda, 2007). In this space, a* and b* are chromaticity layers, representing red-green and blue-yellow axes, respectively (Bora et al., 2015). L* represents the luminance, or more crudely, the lightness of the stimulus (Schanda, 2007). The rigorous controls of the LAB color space allow for more salient analysis of color data that is crowd-sourced from people using a variety of personal equipment under a variety of settings.

2.2 Methods

Concrete	**Abstract**
dog	big
rain	adjective
lime	religion
universe	noun
cloud	angry
door	justice
berry	sad
tree	verb
sky	existence
	running
	soaring
	fun
	pain
	equality

Table 1: Words for which color annotations were collected, selected based on their degree of abstractness.

The 24 words shown in Table 1 were selected to represent a variety of word-types. About half of these words were chosen for their degree of "concreteness". These are words that have a real-world object reference or for which color is a salient feature of the concept the word refers or relates to. The other half of these words are so-called "abstract" words that are best described only by using other words. Some of these abstract words were chosen based on the concreteness ratings determined by Brysbaert et al. (2013), others were chosen for being emotion-laden (Sutton & Altarriba, 2015), and still others were chosen because their abstractness lay in human constructs or measurements (e.g. "noun" or "big").

To obtain perceptually salient color annotations for these words, 120 participants were given an online survey in which they were shown the 24 words in a random order, one at a time, and instructed to select the color they most closely associated with each word from an RGB-HSV color picker. Participants were also asked to select colors for 12 common color-words (e.g. "red", “orange”, “cyan”, etc.) so as to provide a link between the color and semantic spaces for later analysis. Finally, participants were asked to enter their age and select their gender from “male”, “female”, and “other (please specify)”. To control for the differences across devices as much as possible, participants were also asked to disable any software that altered screen colors (e.g., at different periods of the day, such as to obtain "warmer" or "cooler" colors), to set their screens to a comfortable indoor brightness, and to complete the survey on a laptop or desktop computer.

After converting the hex code color annotations obtained from the survey responses to LAB coordinates, I plotted these responses and the average LAB value of these responses for each word using a Matlab package (Eckhard, 2014).

2.3 Results

Figures 3 and 4 display the 120 color annotations obtained for each word, plotted in LAB space.

2.4 Analysis

2.4.1 Qualitative Analysis

As hypothesized, the LAB spaces of nearly all the concrete and emotion-laden words show strong agreement among their color annotations. “berry”, “cloud”, “door”, “dog”, “tree”, and “lime” all show clustering around the colors of their real-world references. “angry” and “sad” show clustering around reds and blues, respectively, and particularly around darker shades of these colors. Though the emotion and physical senses of “pain” were not disambiguated in the survey, its LAB space clusters around darker, blood-like reds. While “mortal” is considerably more abstract and ambiguous, it too clusters around similar shades of reds and blacks, which could speak to the word’s connection to emotion-laden concepts like fear, death, or pain. On the other hand, the LAB space for “fun” shows exclusively bright colors and most apparently clusters around magenta/pink and yellow/sea green. These connections between brightness and positive valence, and darkness and negative valence corroborate with findings in color-emotion association research (Meier, Robinson, & Clore, 2004).

The LAB spaces of the present progressive verbs “running” and “soaring” show clustering around the colors one would find in the environments that those actions might most typically take place in. “soaring” clearly clusters around turquoise and sky blue, almost to the point of directly mimicking the “sky” and “cloud” LAB spaces. To a lesser extent, “running” shows clustering around the greens and yellows found in a park or field. If human color associations do exist within a complex network of environment and personal experience (Schloss & Palmer, 2017), then the process of compiling and generalizing all these features of “running” or “soaring” to arrive at a color-association for an abstract state is not trivial. The possible connections between color, environment, and words raise the question of whether athletes and runners who train on red or blue indoor tracks would have different color associations for the word “running” than the casual park jogger.

Other words have a still greater degree of abstractness than these emotion or action words because their semantic value really only exists in relation to other concepts or purely human/social constructs. “equality” and “justice” are two such examples whose LAB spaces interestingly show some consistency among their responses. “equality” appears to cluster around brighter colors,

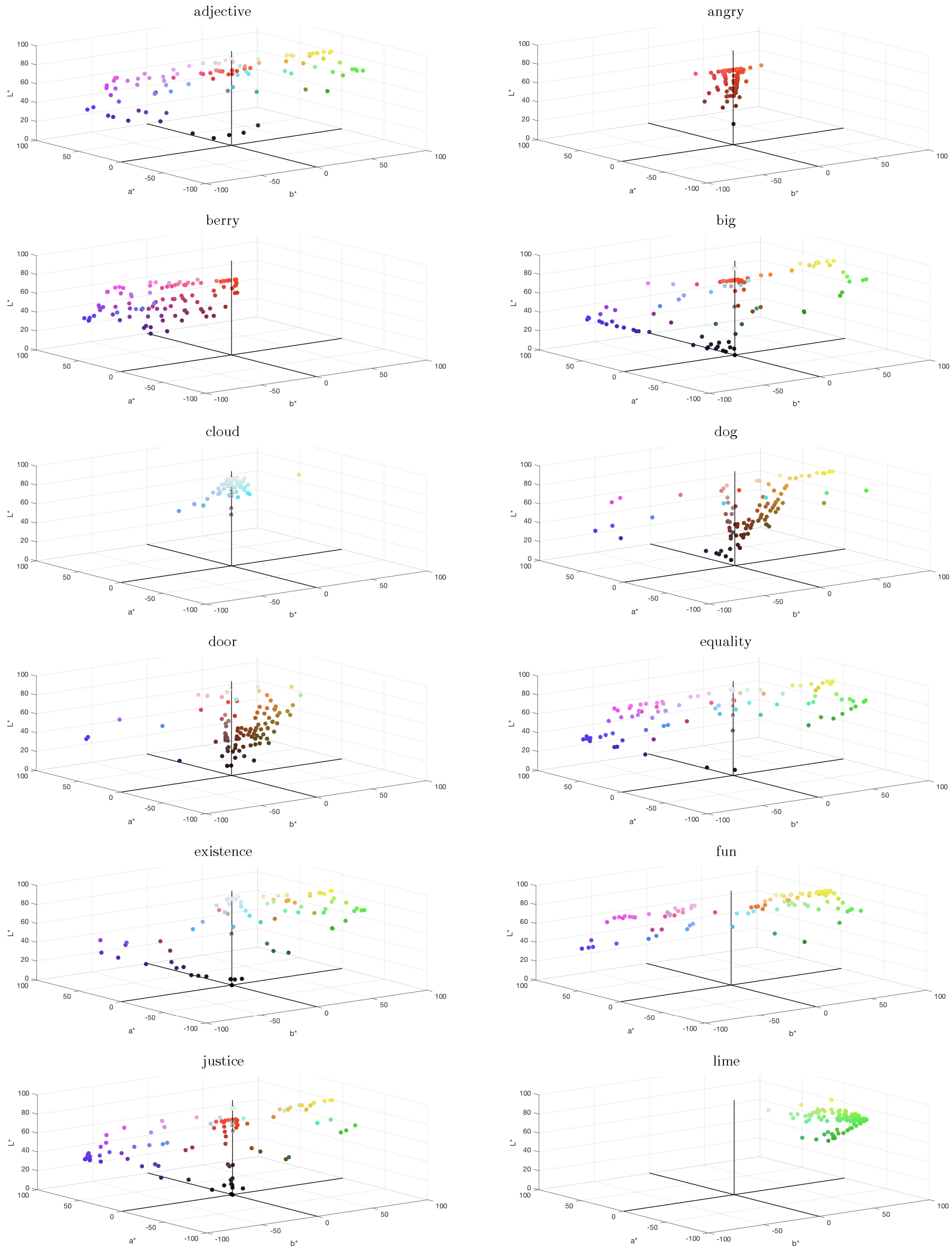

Figure 3: LAB plots of human word-color annotation data.

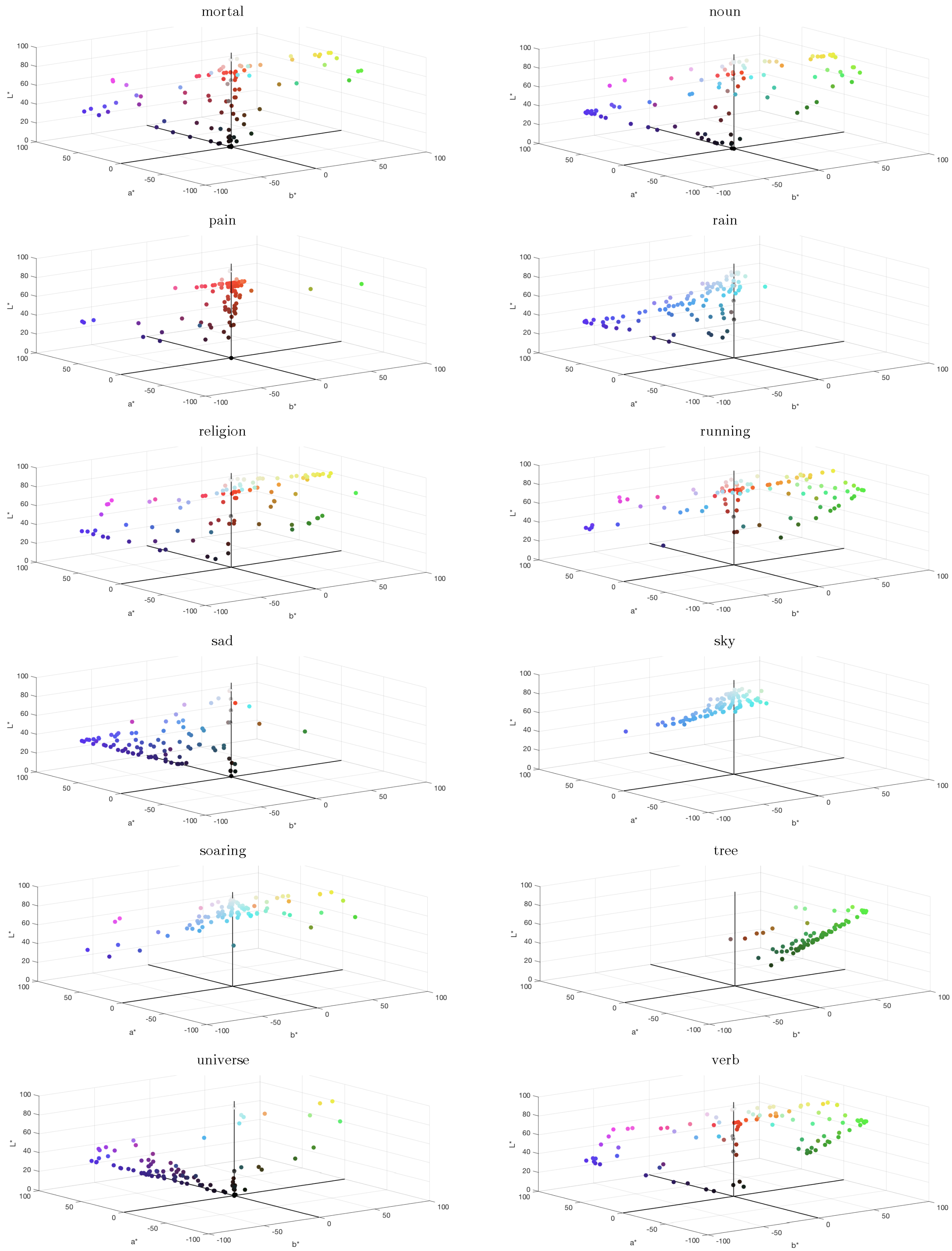

Figure 4: LAB plots of human word-color annotation data.

particularly pinks and purples while "justice" clusters around reds and black. The LAB space of "religion" shows a greater variety of responses, but still some clustering around warmer colors (browns, reds, oranges, and yellows).

Of all the kinds of abstract words, I was most curious about the responses to the part of speech words tested. "noun", "adjective", and "verb" are certainly abstract, human-fabricated ideas like "justice" or "equality," but are unique in that they seem very separate from any emotion or connection to humanity. Each of these parts of speech is made even more abstract by the fact that they encompass hundreds of words across hundreds of different categories and contexts. Despite the unparalleled abstractness of these words, many of the participants I talked to expressed very strong feelings about their color selections for these words. Though the variance in the LAB spaces of these words is quite high, there is some clustering visible around green for "verb" and around pinks and reds for "adjective". It is unclear what might account for these associations. Witthoft & Winawer (2013) found that 11 color-grapheme synesthetes had startlingly similar color-grapheme pairings that were traceable to childhood toys containing colored letters. Given that the population I surveyed was largely comprised of western, college-educated young adults, one hypothesis could be that these associations are due to consistencies in the way western textbooks and grammar lessons use color.

3 Distributional Semantic Model

3.1 Background

Distributional semantics aims to create formal representations of semantic knowledge while assuming the Distributional Hypothesis, which states that words that occur in the same contexts tend to share similar meanings (Harris, 1954). Though popular distributional semantic models take different approaches to the Distributional Hypothesis, they all essentially use co-occurrence statistics of a natural language text to formulate statistical relationships about the semantic environment. Predictive models like word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) use a simple neural network representation to directly predict a word from its neighbors, while count-based models like Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997) and Topic Model (Griffiths, Steyvers, & Tenenbaum, 2007) use co-occurrence counts from a large text corpus to eventually predict words based on spatial proximity. Many of these models produce a spatial representation of semantic similarity such that distances between points in semantic space represent the degree of similarity between word meanings. Another distinction can be made between matrix factorization methods and local context window methods. Matrix factorization methods like LSA use all the co-occurrence information for a large corpus, allowing them to benefit from a large sample size and the repetition therein. However, these input matrices are typically very sparse and result in a spatial representation that isn't complex enough to perform well on word analogy tasks (Pennington et al., 2014). On the other hand, making predictions within local context windows of neighboring words, as word2vec does, doesn't take advantage of this statistical information in a corpus

3.1.1 Global Vectors for Word Representation (GloVe)

Global Vectors for Word Representation (GloVe) (Pennington et al., 2014) draws from the unique advantages of matrix factorization methods and local context window methods. The model is trained on the non-zero entries of a global word-word co-occurrence matrix X, where X_{ij} is the number of times word j occurs in the context of word i and $P_{ij} = P(j|i) = X_{ij}/X_i$ gives the probability that j occurs in the context of i. The ratio of co-occurrence probabilities of i and j with respect to different probe words k, P_{ik}/P_{jk}, reflects the degree of relation between k and i and k and j. Probability ratios close to 1 usefully indicate probe words that are either equally relevant to i and j or irrelevant to both; access to the ratio values helps to further quantify the degree of relevancy. Using this aspect of meaning found in the co-occurrence statistics of a corpus as a starting point, GloVe addresses a variety of other computational and semantic concerns to develop into the following weighted least squares regression model

$$J = \sum_{i,j=1}^{V} f(X_{ij}) \left(w_i^T \tilde{w_j} + b_i + \tilde{b_j} - \log X_{ij} \right)^2$$

where J is the function to be minimized, V is the size of the vocabulary, w_i is the word vector for word i, $\tilde{w_j}$ is the context word vector for word j (essentially the transpose of w_j), and b_i and $\tilde{b_j}$ are the "bias" terms for the word vector w_i and context word vector $\tilde{w_j}$, respectively. This bias term is introduced to account for the fact that, for word-word co-occurrence matrices, the distinction between a word and a context word is arbitrary. The weighting function $f(X_{ij})$ is non-decreasing and relatively small for large values of X_{ij}, so that neither rare nor frequent co-occurrences are overweighted.

GloVe performs significantly better than both flavors of word2vec and SVD-based methods on both semantic and syntactic word analogy tasks. This indicates the presence of the same kind of meaningful linear substructure responsible for word2vec's success, while utilizing the valuable co-occurrence statistics of the corpus to improve computational efficiency. In this project, I use the GloVe model pretrained on the "glove-wiki-gigaword-100" data, which is a combination of the Wikipedia 2014 dump and the Gigaword 5 corpus, trained on 6 billion tokens, with 400,000 tokens considered unique.

3.2 Methods

To analyze the relationship between this color space and the semantic space that the 24 tested words exist in, I first obtained the GloVe cosine similarities s between each of the 25 words and each of the 12 color-words (Table 2). According to the GloVe model, the cosine similarities (or Euclidean distance) between two word vectors provide a way of measuring the linguistic or semantic similarity of the words (Pennington et al., 2013). Using each word's semantic relatedness to the color-words, I then calculated the semantically-predicted location w' for each word in two ways:

$$w' = \frac{\sum_{color\text{-}words} s(word, color\text{-}word) v(color\text{-}word)}{\sum_{color\text{-}words} s(word, color\text{-}word)} \tag{1}$$

$$w' = \frac{\sum_{color\text{-}words} e^{(\gamma * s(word, color\text{-}word))} v(color\text{-}word)}{\sum_{color\text{-}words} e^{(\gamma * s(word, color\text{-}word))}} \tag{2}$$

where v is the mean color value of the color-word, obtained by averaging the color values for each color-word. In equation (2), the hyperparameter γ is tuned so as to minimize the sum of the distances d over all 25 tested words, allowing me to determine the significance of s to the calculations of w'. I tested the following values of γ: 0.000001, 0.00001, 0.0001, 0.001, 0.01, 0.1, 0.5, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0, 6.5, 7.0, 7.5, 8.0, 8.5, 9.0, 9.5, 10.0, 11.0, 12.0, 13.0, 14.0, 15.0, 16.0, 17.0, 18.0, 19.0, 20.0, 21.0, 22.0, 23.0, 24.0, 25.0. The final color-semantic distance d between the actual mean location (color value) w of a word as given by the human annotations and its semantically-predicted location w', is

$$d = w - w'$$

Word	**Color-word**	**Cosine similarity** s
dog	red	0.43856516
	orange	0.31241626
	yellow	0.39498812
	green	0.42506728
	blue	0.4526073
	purple	0.32258704
	pink	0.4341574
	brown	0.40182555
	black	0.44863808
	white	0.41592243
	cyan	-0.040190637
	magenta	-0.013375642

Table 2: Example of GloVe cosine similarities s between the tested word "dog" and each of the 12 color-words.

3.2.1 Evaluation

In addition to conducting visual analyses of the LAB plots produced by the model, I evaluated my results by performing the following statistical analysis to establish the chance color-semantic distance d (i.e. what d would have been if there was no relationship between an average color location and a semantically-predicted location). I obtained a distribution over the distance between the average color location and semantically-predicted location by randomly pairing the average human color annotation w of a tested word with the semantically predicted location w' of any other tested word. Using this, I determined the value of d such that 95% of the permutation distances were above this value. Any of the words whose actual distances were below this value could then be interpreted to be statistically significant—that is, less than 5% probability by chance. I used this same evaluation method with different, model-appropriate definitions of w' for the image analysis and Bayesian representativeness models detailed later in this paper.

3.3 Results

Figure 5 shows a sample visualization of the distance between average human color annotation and color location predicted by the distributional semantic model. Figures 6, 7, and 8 present the results of the statistical analysis of computing the semantically-predicted location w' using equation (1) (i.e. without exponentiating the word/color-word similarity value or multiplying by hyperparameter γ). Figures 10, 11, and 12 present the results of the statistical analysis of computing the semantically-predicted location w' using equation (2) (i.e. by exponentiating the product of the word/color-word similarity value and the hyperparameter γ that minimized d).

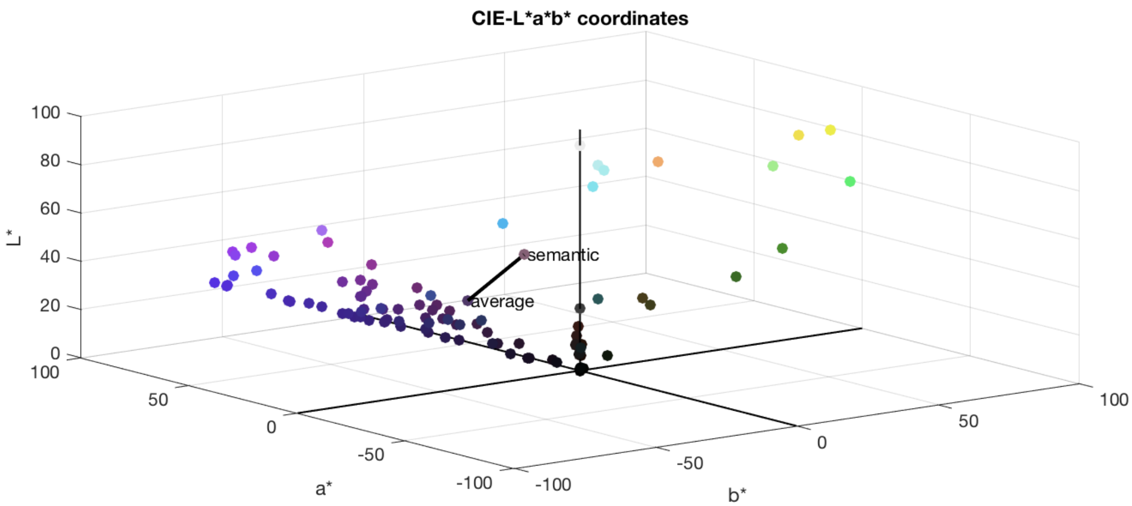

Figure 5: Example of a LAB plot from Section 2 with additional points for the average of the human color annotations (labeled "average") and the color location as predicted by the distributional semantic model (labeled "semantic") for the word "universe". The distance d between these two points forms the basis of further statistical analysis.

3.4 Analysis

3.4.1 Statistical Analysis

Plotting the distances d between average color value of the human annotations and semantically predicted location for each word shows that "berry" and "big" perform the best of all the words, suggesting a particularly strong connection between the color domain and semantic domain for these words. This is puzzling because while "berry" is a concrete word that one would expect to have consistent color associations and references for, "big" is an extremely abstract word. The LAB space for "berry" does show agreement among its color annotations, with clustering around reds, purples, and pinks. However the LAB space for "big" has high variance, meaning that its average color location would tend towards the origin, which is close to its inconclusive semantically predicted location, and thus, may explain the statistical significance for this word.

Considering this color-semantic distance for "lime"—a concrete word with a very familiar real-world object reference—in conjunction with its LAB space seems most troubling for the hypothesis that a distributional semantic model is capable of correctly modeling word-color associations. Few words' LAB spaces cluster as strongly around a single color as that of "lime", yet its semantically-predicted location is so far away from this cluster that it results in the largest distance between the color and semantic domains. With the exception of "big" and

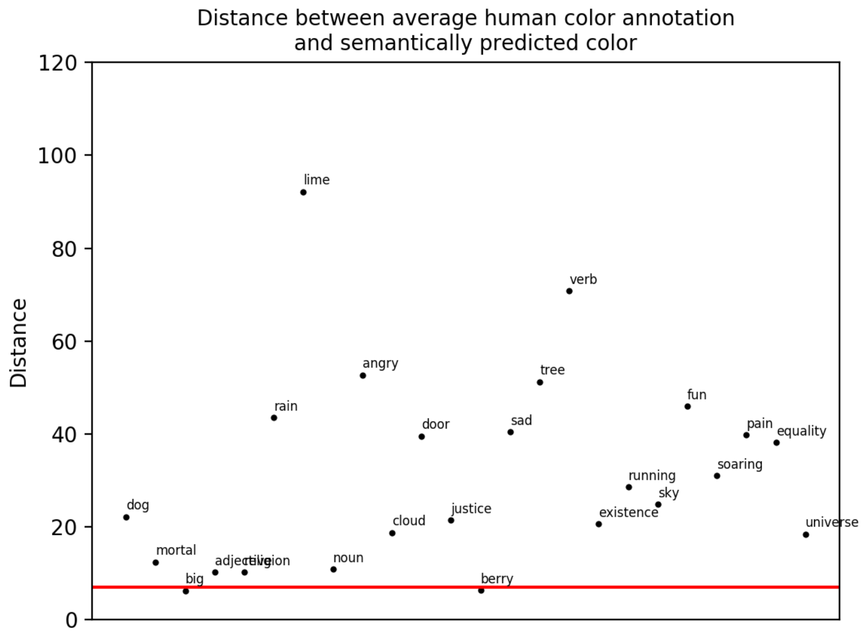

Figure 6: Distances between the average color location w and semantically-predicted location w' for each word.

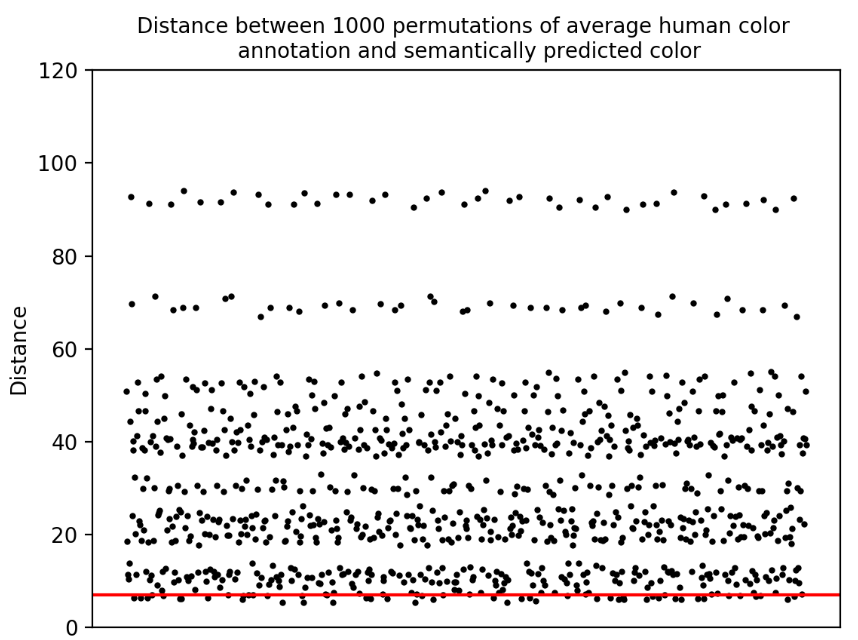

Figure 7: Scatter plot of distribution of distances obtained from randomly permuting w and w' across all the words tested.

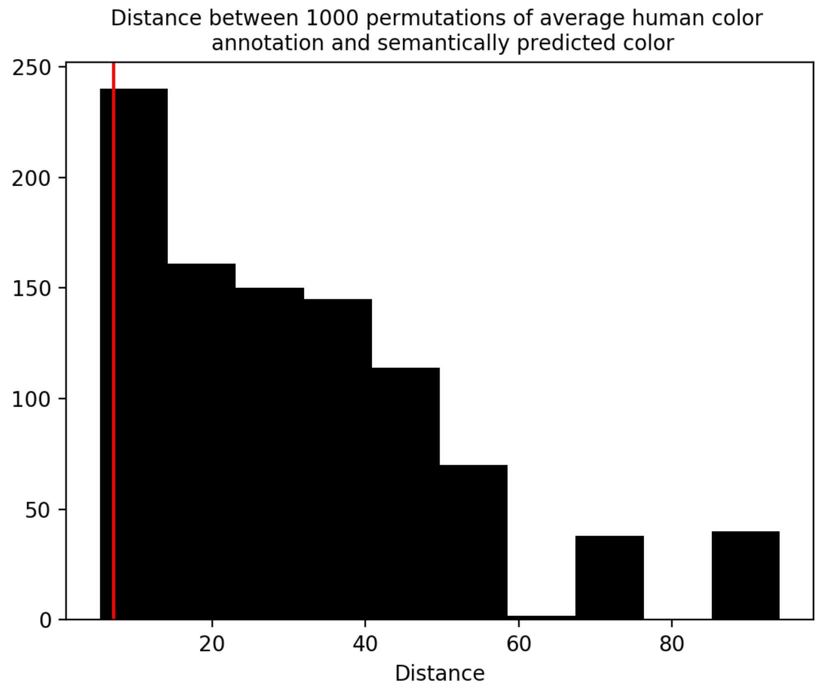

Figure 8: Histogram of distribution of distances obtained from randomly permuting w and w' across all the words tested.

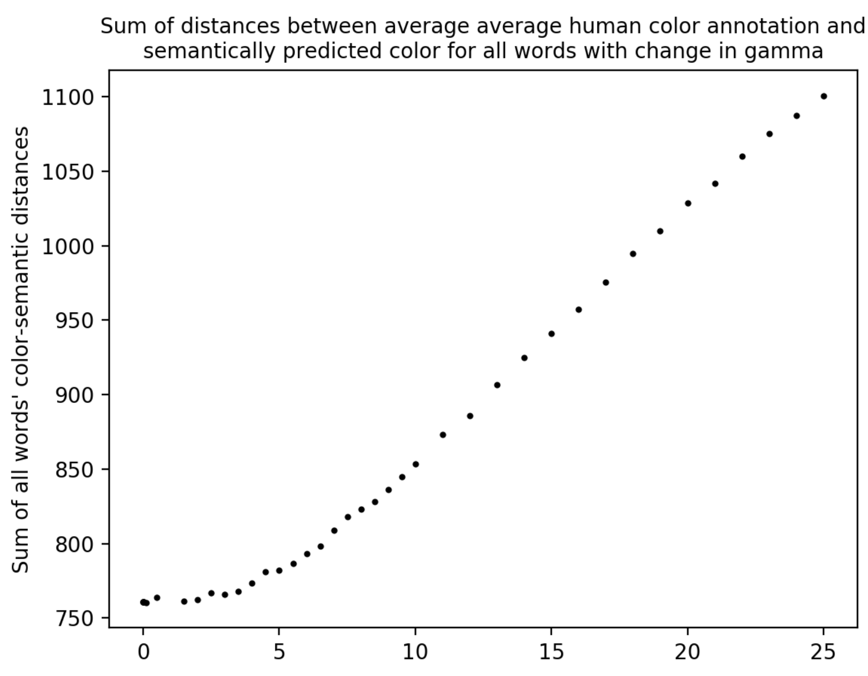

Figure 9: Sum of distances d over all 24 words, for each value of γ tested. $\gamma = 0.10$ was chosen as it minimized the sum of the distances d over all 24 tested words.

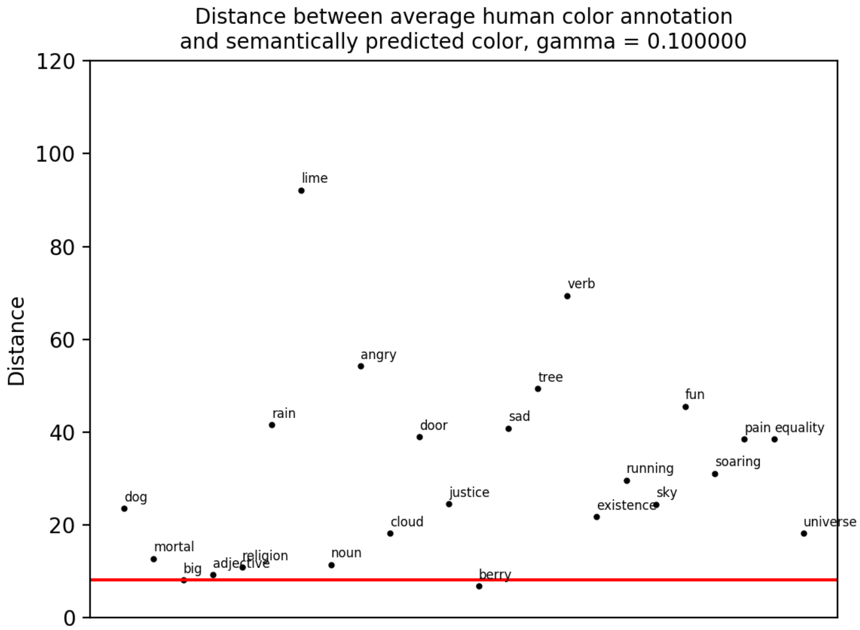

Figure 10: Distances between the average color location w and semantically-predicted location w' for each word for optimal value of γ.

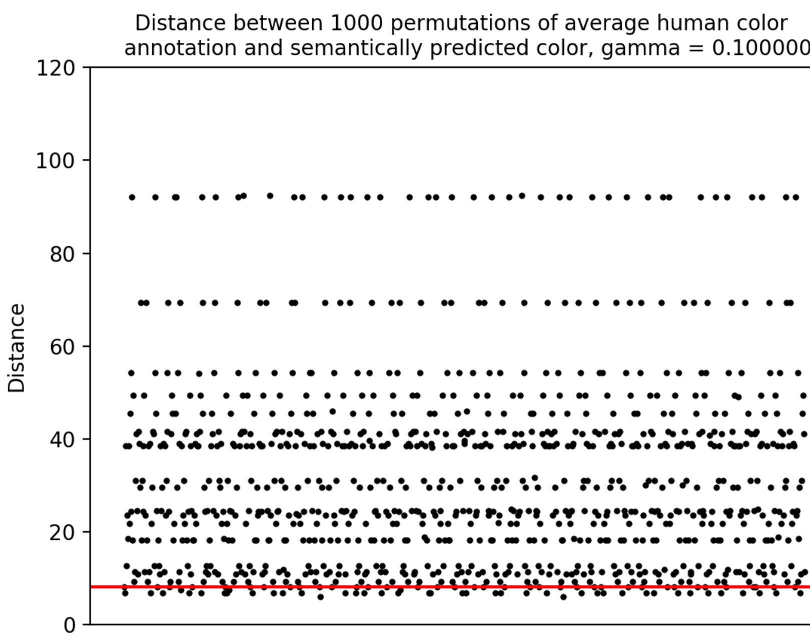

Figure 11: Scatter plot of distribution of distances obtained from randomly permuting w and w' across all the words tested for optimal value of γ.

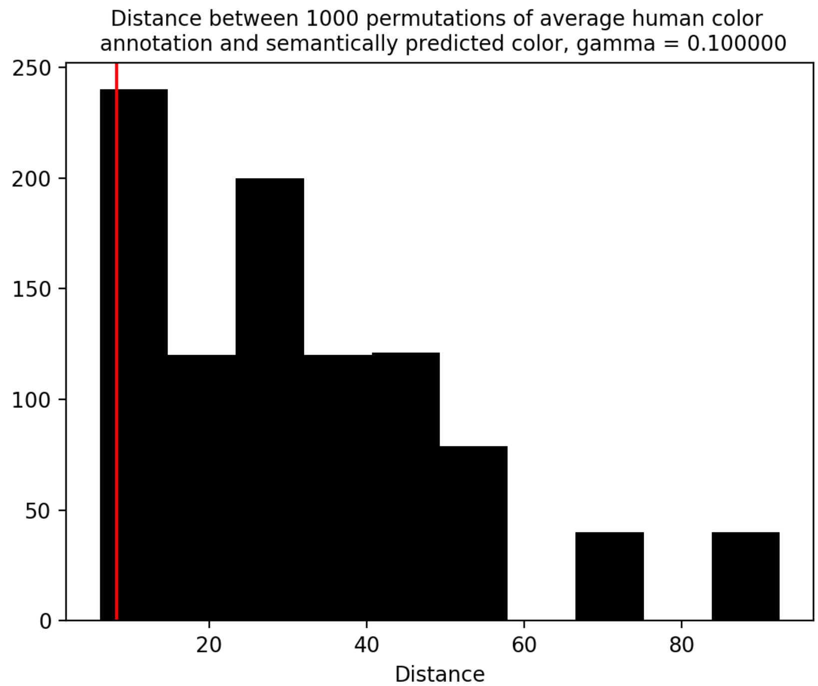

Figure 12: Histogram of distribution of distances obtained from randomly permuting w and w' across all the words tested for optimal value of γ.

"berry", the tested words all have statistically non-significant distances between their average color location and semantic location. This is puzzling because among these are emotion-laden words like "sad", "pain", and "angry", and concrete words like "dog", "tree", and "lime", all of which have LAB plots that clearly cluster around a single color. Even when there appears to be word-color agreement in the color domain, this apparently does not translate to the same word-color agreement in the semantic domain.

These findings are maintained in the results of choosing γ as to minimize the color-semantic distance: "big" and "berry" still outperform the rest of the words, while many concrete and emotion words have statistically non-significant color-semantic distances. Larger values of γ suggest that the information found within the cosine similarity between words and color-words is important to the semantically-predicted location of a word. However, values of γ closer to 0 suggest that there is not much information in the semantic domain, as essentially eliminating the similarity metric would not change the semantically-predicted location w'. I found $\gamma = 0.10$ to be optimal, which leads me to believe that the information derived from the distributional semantic model is not useful to the predictive abilities of this model.

3.5 Discussion

These results prompt the need for some improvements to this model in order to capitalize on the theoretical potential for grounding automatic word-color associations in distributional semantics. As a first step, improving the design of the data collection survey might help to either reduce, or make more meaningful, the variance in the color data obtained. Many of the words presented were highly abstract and quite difficult to arrive at color associations for.

Some participants mentioned that they ended up choosing #ffffff (white) or #000000 (black) for those words that did not elicit any particular color association. It is likely that many others impulsively selected a color for certain words without having any serious associations, only because the survey did not allow for non-responses. In the future, allowing subjects to indicate that a word did not elicit any particular color association or to indicate on a scale from 1 to 5 how strongly they felt about the color they selected for each word might help determine the most salient color associations.

Other concerns and opportunities lie in my method of interfacing with the semantic domain. In using the GloVe cosine similarity metric, the calculation of the semantically-predicted location w' was ultimately derived from how frequently a tested word co-occurred with each of the color-words, within the corpus that the semantic model was trained on—namely, seven distinct international sources of English Newswire and Wikipedia (Pennington et al. 2013). Thus, the color-semantic distances I have obtained could say less about a semantic basis for psychological word-color associations, and more about the sparsity of words and color-word co-occurrences within the language registers represented in the GloVe training data. It is possible that a training corpus comprised of fictional stories, poetry, or creative writing might be less sparse with respect to words and color-word co-occurrences because of the illustrative, metaphorical language featured in these kinds of texts.

Finally, my two methods of computing the semantically-predicted location w' raise concerns about the effectiveness of distributional semantic models at large, for this application. Initially, I computed w' using the raw cosine similarities which included negative values. Repeating these analyses having exponentiated the product of the cosine similarities and different values of γ showed that $\gamma = 0.10$ minimized the sum of the color-semantic distances across all words. However, this optimization did not change number of words or words with statistically significantly color-semantic distances, it only lowered every word's color-semantic distance. Even with these optimizations and theoretical potential, it is questionable whether distributional semantic models provide a suitable foundation for representing word-color associations in practice. The relatively small value of γ I obtained suggests that the cosine similarity metric is not significant in the calculation of the semantically-predicted location. Further, plotting these semantically-predicted locations in LAB space shows little variation across the different words. Nearly every word's predicted location clusters around the origin, likely because the distribution over the color-words is uniform.

4 Image Analysis Model

4.1 Background

The above discussed shortcomings of the distributional semantic model led me to reexamine not only model structure, but also the kind of data that any model of word-color associations would require to be successful. One hypothesis is that the "agreed-upon" color associations for words might be grounded in more than just the psychological associations I relied on in collecting the human word-color data in Section 2. As online pictures and videos have become more ubiquitous in our lives, perhaps concepts that were once more abstract or mental are now more grounded in the visual references found in online content.

To test this, I conduct an analysis of the colors contained in the top Google image results for each of the 24 words tested. The goal of this model is two-fold. First, I better understand how the human subjects' color annotations relate to the colors found in online images and whether color data for certain abstract words should be supplemented by information from their visual references. Secondly, I address the issues of small sample size, labor intensive process, and lack of strong clustering in the human color annotations by implementing an automatic method of obtaining comprehensive and consistent word-color association data, which will hopefully provide a stronger foundation for other work in this domain.

4.2 Methods

Using the GoogleImageCrawler package from PyPi, I collected the top 8 image results for each of the 24 words and extracted the colors from these images using the color conversion packages found in OpenCV. After experimenting with a few different methods of identifying the predominate colors in each image, including obtaining all the colors from each image and then identifying those colors that made up $x\%$ of the pixels in the image, I found k-means clustering to be the most effective method of arriving at a reasonable number of representative colors. Selecting the top 15 color-clusters for each of the top 8 images returned for each word, as identified in RGB, and then converting these values to the LAB space gave the final mapping of the distribution of colors for each word.

4.2.1 Evaluation

I evaluated these results as described in Section 3.2.1, this time with w' equal to the average of the image analysis color points for each word.

4.3 Results

Figures 13 and 14 display the top 120 color values obtained for each word, plotted in LAB space. Figures 15, 16, and 17 show the results of the evaluation method described in Section 3.2.1.

4.4 Analysis

4.4.1 Qualitative Analysis

In general, the results of this model exhibit less variance than the human color annotations, which may suggest that this data is better suited as a foundation for predictive models than the human responses that show minimal agreement, especially for abstract words. However, one immediate issue is that across all plots, a significant portion of the color points are uninformative shades of grey found in the backgrounds of the images. This is likely because usually only a small proportion of an image contains the content of the search, and since the images were used in their original state without any attempt to isolate the main objects whose colors we care about, this method was unable to maximize the amount of truly representative color data obtained for each word. With that said, for the remainder of the qualitative analysis for this model, I disregard the greyish-black points that are not unique to any word's LAB space.

The LAB spaces for the concrete words "lime", "cloud", "door", "sky", and "tree" cluster around the expected colors and corroborate the hypothesis that the human color annotations for

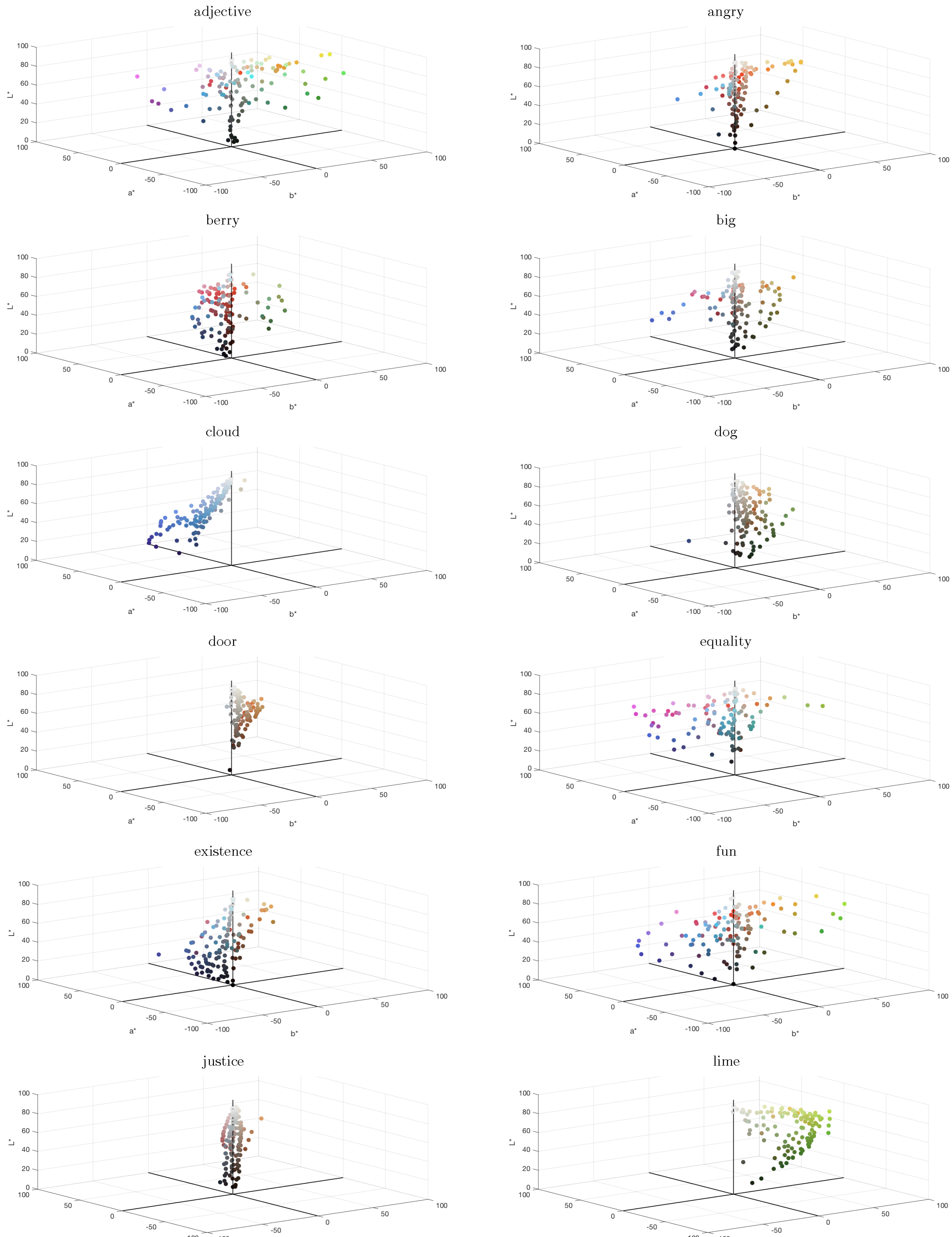

Figure 13: LAB plots of image analysis data.

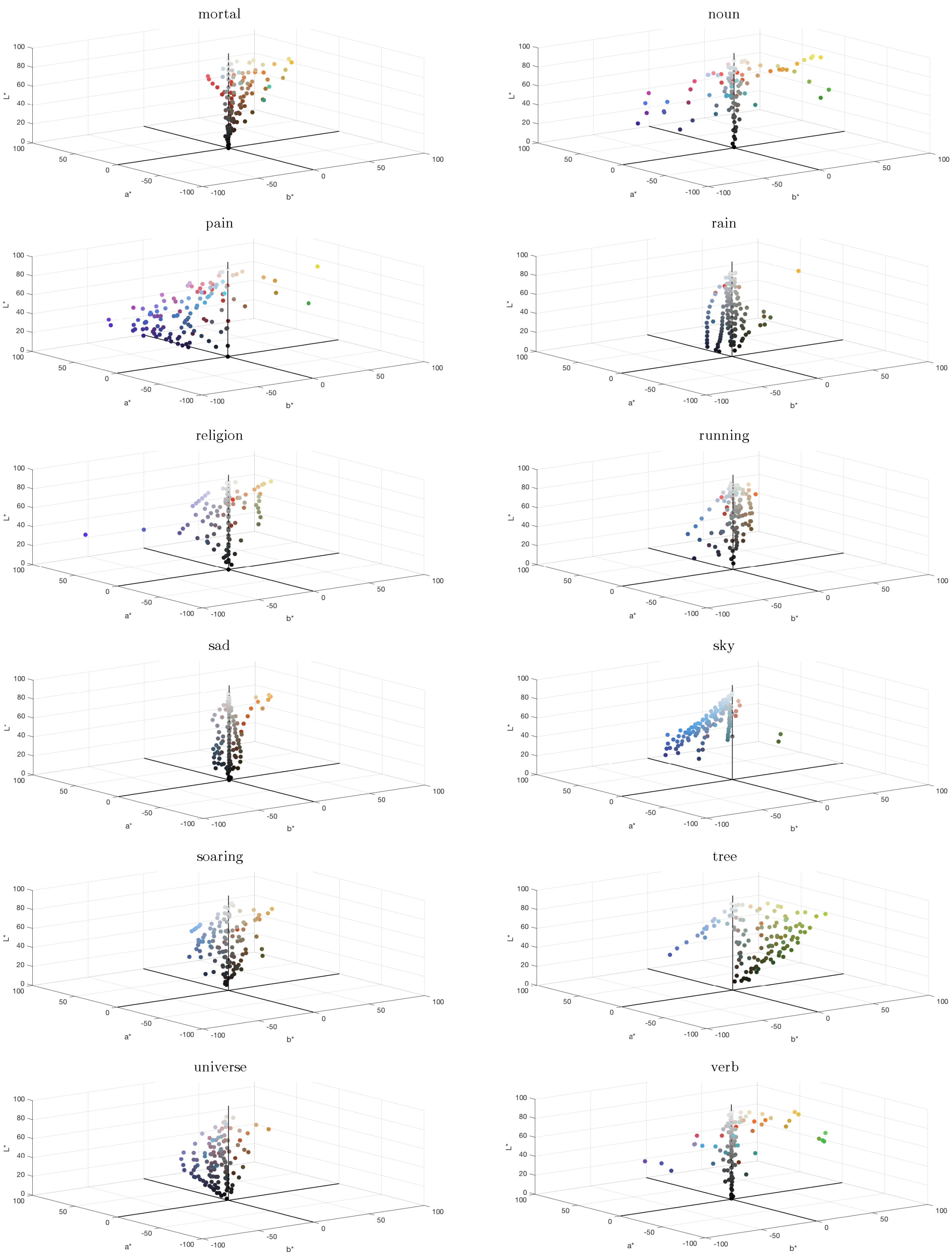

Figure 14: LAB plots of image analysis data.

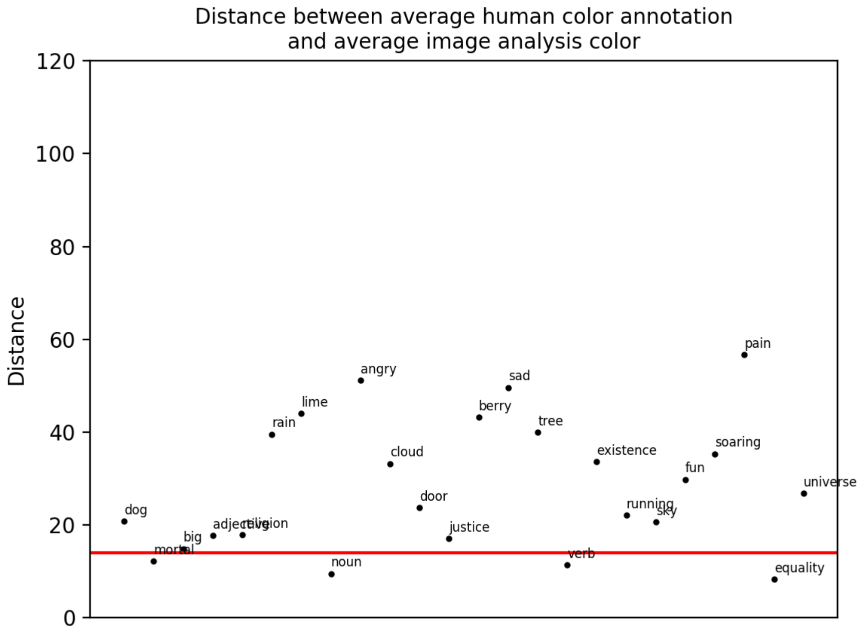

Figure 15: Distances between the average human color annotation w and average image analysis color w' for each word.

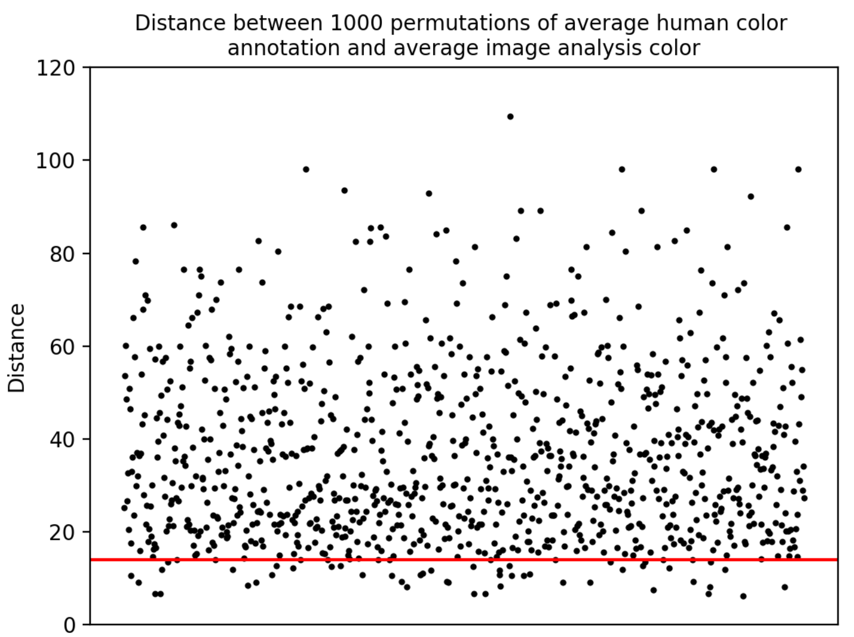

Figure 16: Scatter plot of distribution of distances obtained from randomly permuting w and w' across all the words tested.

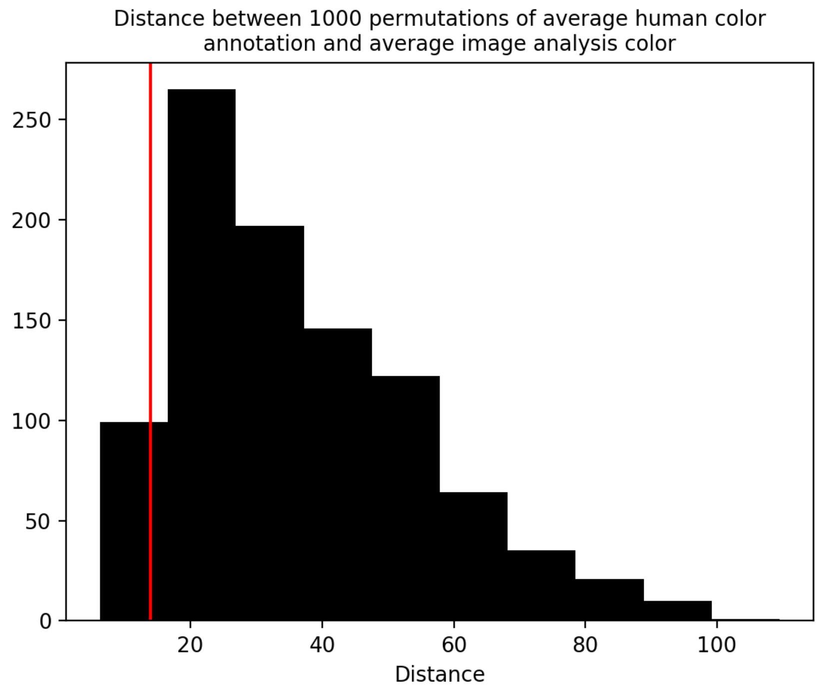

Figure 17: Histogram of distribution of distances obtained from randomly permuting w and w' across all the words tested.

these words are rooted in their visual references. For other words, comparing the image analysis LAB plots to their human response counterparts reveals much more interesting information about where humans' color associations may be rooted. We can see that image data for the emotion-laden word "angry" contains a cluster of bright reds that corroborates with the results of the human annotation data. This suggests that, despite being an abstract word with no direct reference, its psychological color association aligns with popular visual references of the word rather than a mental understanding of the word or the way this word is used in language. As another example, the image data for "berry" appears to be tied to the red of strawberries and raspberries, which is the most distinctive color of the word's real world references (other berries are essentially black and this is a color shared across the images for all words). However, the human color annotations for this word cluster around pinks and purples much more than red, which suggests that, despite being a concrete word, the psychological color association for "berry" is tied to a more abstract understanding of the word's usage.

4.4.2 Statistical Analysis

Plotting the distances between average color value of the human annotations and the average image analysis color for each word shows that the model's results for "mortal", "noun", "verb", and "equality" are statistically significant. This seems to reveal a problem with the model, rather than provide any indication that the model is successfully finding color associations for such abstract words. Because so much of the color data in the LAB plots for this model was comprised of shades of grey, it is likely that the average color value for each word was skewed towards these uninformative background colors. Simultaneously, the high variance in the human color annotation data for many of the words, especially these abstract ones, resulted in averages that were also greyish colors close to the origin. The combination of these two issues results in distances that are low, but not because the model has arrived at a unique or meaningful color association for the word.

Further, as in the distributional semantic model, the results for many concrete words like "lime", "tree", and "berry" are not only statistically insignificant but also have distances that are among the highest of all the words tested. This is especially troubling given that these words have both predictable image references that contain the expected colors (as shown in their LAB plots) and are some of the only words for which the human color annotations showed strong clustering and thus, a unique average color association. The emotion-laden words "angry", "sad", and "pain", for which the human data also show strong color associations, perform even worse than the concrete words with the largest distances overall.

5 Bayesian Representativeness Model

5.1 Background

While it appears that the previous image analysis model is inadequate as a means of predicting color associations for words, as a dataset it may serve as a more robust foundation for other modeling approaches. The model described in this section approaches the goal of representing word-color associations from the opposite direction than the previous distributional semantics and image analysis models. Rather than trying to predict a color (or color cluster) from a word, this model determines which word a particular color point in LAB space should belong to based on the representativeness scores (Tenenbaum & Griffiths, 2001) of points in LAB space. This score formalizes what makes an observation (in this case, a color) a good example of a category (word).

5.2 Methods

First, I aggregated all the color points obtained from the image analysis task in Section 4 for all 24 words and 1000 points randomly sampled from LAB space. This provided a sample of a total of 3880 points that are representative of the distribution of points in color space for the selected words.

Treating this task of classifying points in color space as a supervised machine learning problem, I estimated a Gaussian for each word using the color points belonging to that word, as determined by the image analysis task. These Gaussian distributions act as a separate binary classifier per word that a point could be classified as, and in each, the aim is to find $p(w|x)$, where w is the word and x is the point to be classified, by applying Bayes' rule to $p(w|x)$:

$$p(w|x) = \frac{p(x|w)p(w)}{\sum_{i=1}^{n} p(x|w_i)p(w_i)}$$

The posterior, $p(w|x)$, is the conditional probability of a word label given the color point observed; the likelihood, $p(x|w)$, is the conditional probability of a color observation given the color label; the prior, $p(w_i)$, is the probability of a word label class occurring and can be set to incorporate into the model knowledge about how the relative frequency/rarity of a word or concreteness/abstractness of a word might affect the classification of a color point as that word. The likelihood is modeled as a normal/Gaussian distribution given by the following equation:

$$p(x|w) = \frac{1}{\sqrt{(2\pi)^3|\Sigma|}} exp(\frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)) = \mathcal{N}(x|\mu,\Sigma)$$

where $|\Sigma|$ is the determinant of the covariance matrix $\Sigma \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T$ and the mean μ is equal to $\frac{1}{N}\sum_{i=1}^{N} x_i$, where x_i is the point being sampled.

Having calculated the parameters of the Gaussian for each word, I used these distributions to calculate the representativeness score $R(x, w_{current})$ (Tenenbaum & Griffiths, 2001)

$$R(x, w_{current}) = \frac{p(x|w_{current})}{\sum_{other-words} p(x|w_{other})}$$

for each of the 3880 color points aggregated above. $R(x, w_{current})$ describes how representiative the color point x is of the current category (word) $w_{current}$. Since these probabilities are very small, I addressed the issue of underflow by computing $R(x, w_{current})$ in log space:

$$\begin{aligned}\frac{p(x|w_{current})}{\sum_{other\text{-}words} p(x|w_{other})} &\approx \log\left(\frac{p(x|w_{current})}{\sum_{other\text{-}words} p(x|w_{other})}\right)\\ &= \log p(x|w_{current}) - \log \sum_{other\text{-}words} p(x|w_{other})\end{aligned}$$

and calculating the second term by applying the LogSumExp (LSE) function:

$$\begin{aligned}&LSE(p_o) = \log(p_o)\\ &logmax = max(LSE(p_o))\\ &LSE(p_o) = LSE(p_o) - logmax\\ &logsum = logmax + log\sum exp(LSE(p_o))\end{aligned}$$

where $p_o = p(x|w_{other})$. Using these scores, I selected the 120 most representative points for each of the 24 words and plotted them in LAB space.

5.2.1 Evaluation

I evaluated these results as described in Section 3.2.1, this time with w' equal to the average of the most representative color points for each word.

5.3 Results

Figures 18 and 19 display the 120 most representative color points obtained for each word, plotted in LAB space. Figures 20, 21, and 22 show the results of the evaluation method described in Section 3.2.1.

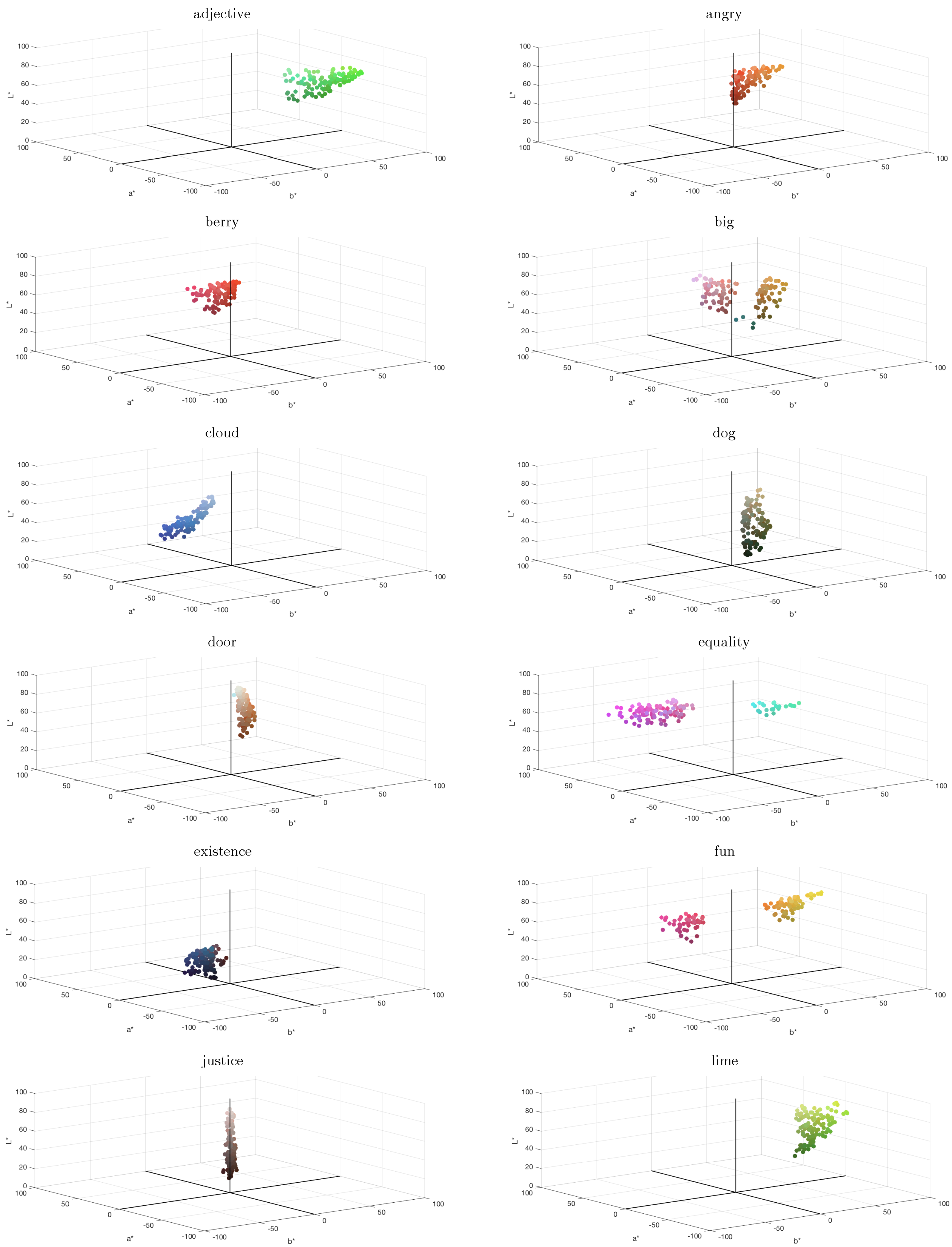

Figure 18: LAB plots of Bayesian representativeness model results.

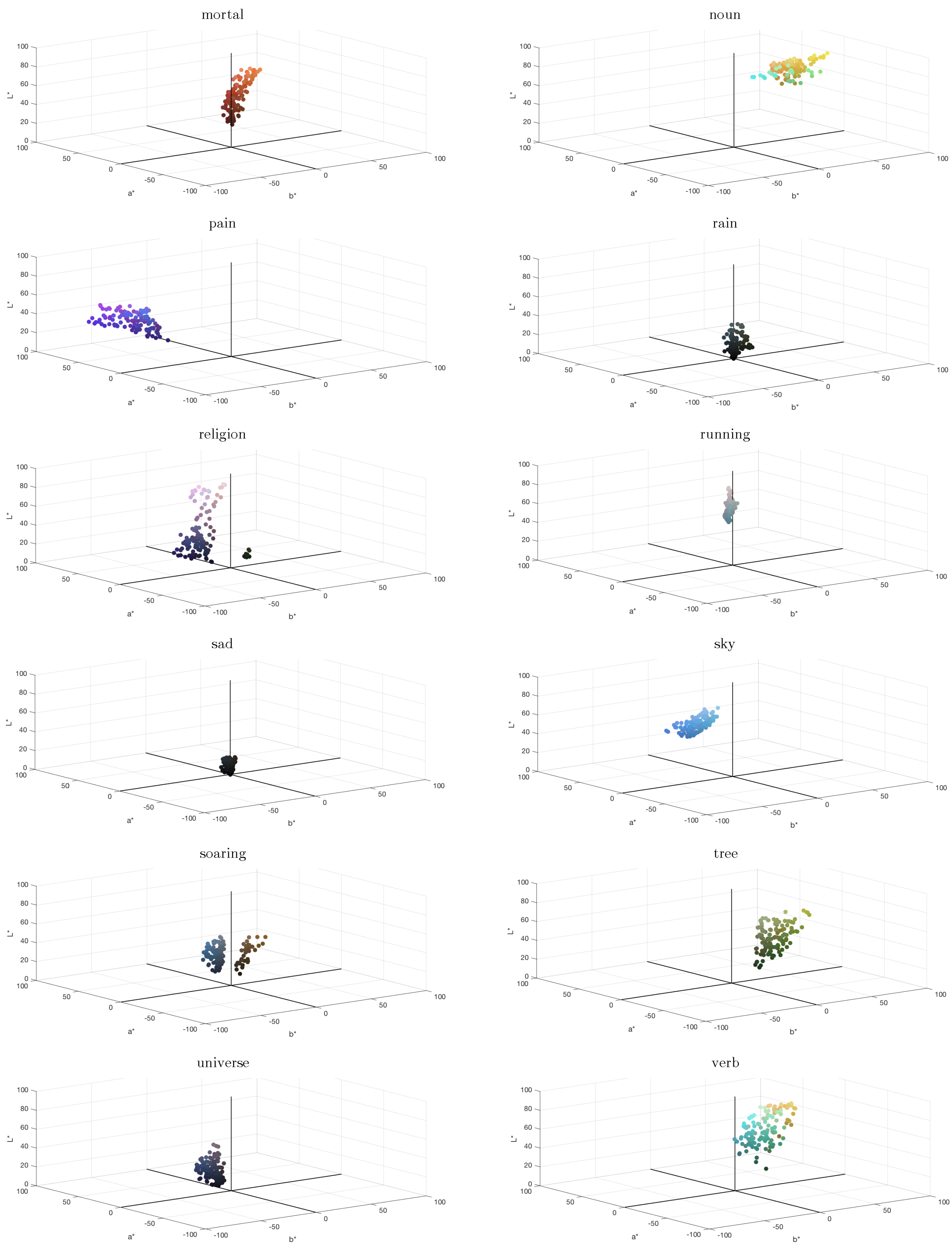

Figure 19: LAB plots of Bayesian representativeness model results.

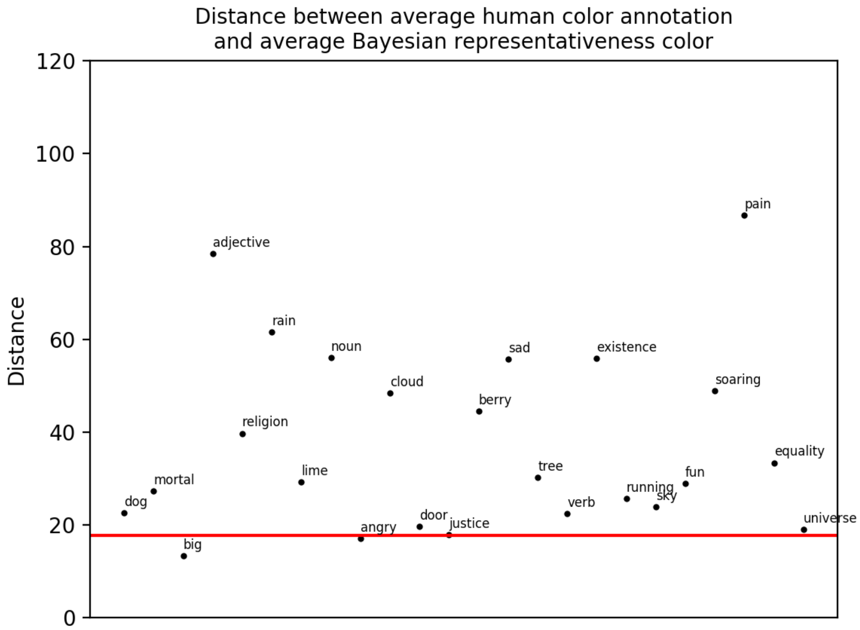

Figure 20: Distances between the average human color annotation w and average Bayesian representativeness color w' for each word.

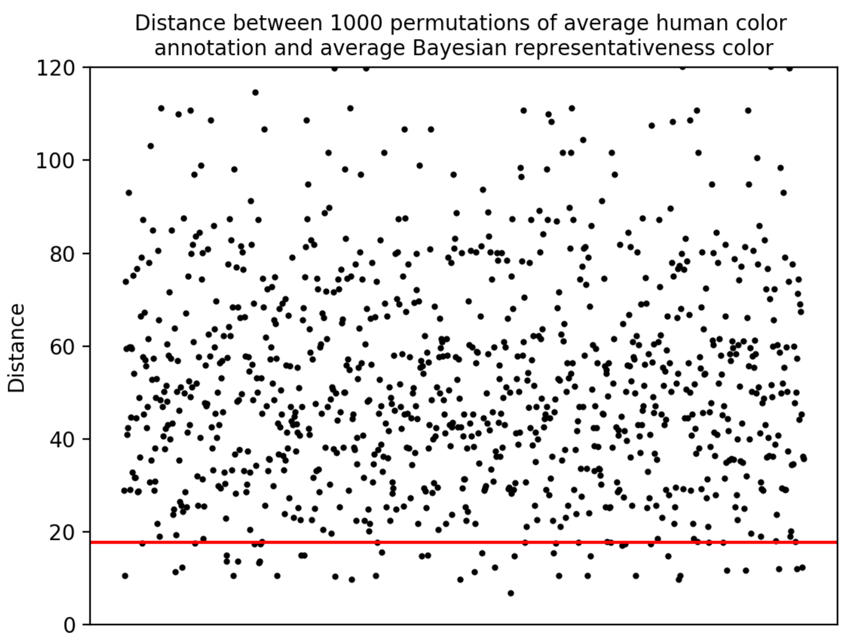

Figure 21: Scatter plot of distribution of distances obtained from randomly permuting w and w' across all the words tested.

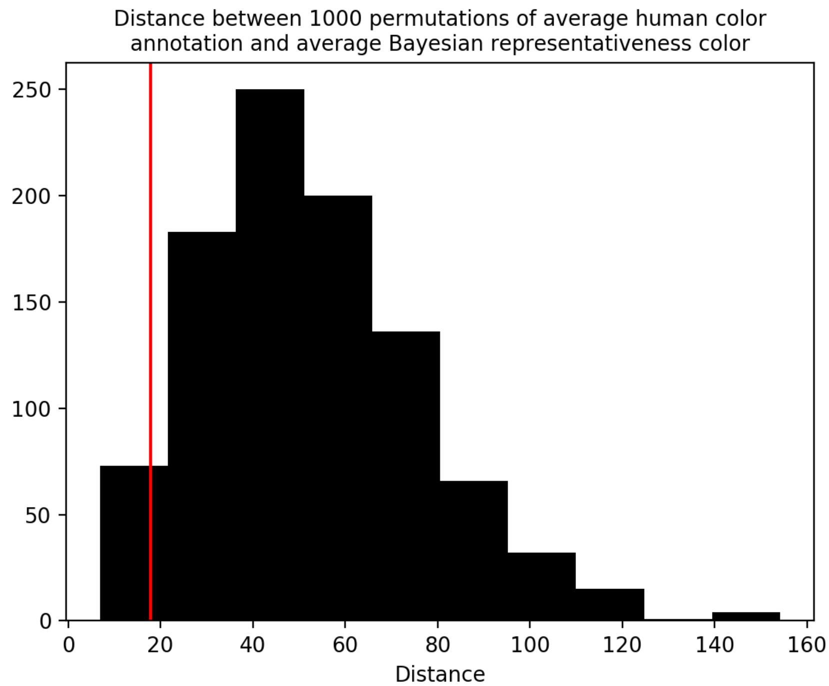

Figure 22: Histogram of distribution of distances obtained from randomly permuting w and w' across all the words tested.

5.4 Analysis

5.4.1 Qualitative Analysis

A qualitative analysis of the LAB plots for this model shows promising results. For each word, we can observe distinct clusters of color points. These LAB plots are also largely free of the uninformative "background" colors (e.g. greys or black) found in the image analysis data that was used as input to this model, meaning that the Bayesian model has done a good job of retaining the points that are uniquely representative of each word and discarding those that are not.

The results for the concrete words "lime", "sky", "tree", and "door" provide preliminary confirmation that this model works as desired, as the LAB plots for these words agree with the colors of their prototypical object references and the human color annotations collected (bright green, blue, dark green, and brown, respectively). The definitive clustering for "adjective", "noun", and "verb" is more interesting, as these words describe purely human constructs and are among the most abstract words tested. While the human color annotations for these words exhibited high variance, the Bayesian model's results clearly cluster around bright green, golden yellow, and turquoise, respectively. Comparing this to the image analysis LAB spaces for these words suggests that the results for this model may be an undue exaggeration of the image data, as the majority of the LAB spaces for these words were truly grey and it seems like the Bayesian model picked up on the small clusters of a few points that happened to exist in otherwise sparse data. Because of the high variance in the human annotation data, it is difficult to determine whether the results of this model on these words reflect human associations. However, the human color annotations for "verb" seem to roughly cluster around the same greenish-blues that this model outputs for "verb".

Most promising are the results for the abstract words "equality" and "fun". This model shows that the most representative points for the word "equality" are primarily bright pinks and magenta with a smaller cluster of blue. These are colors that are traditionally associated with female and male and that are tied to the most commonly addressed subset of equality, gender equality. The most representative points for "fun" are divided into a cluster of bright, saturated pinks and a cluster of orange-yellows–colors that are typically associated with the positive emotions that are inherent to the word "fun" (Hemphill, 1996). These results on abstract words that truly lack a visual reference in a way that some of the other abstract words (e.g. "angry", "sad", or "pain") do not, show that this model is able to produce color annotations that are grounded in some understanding of word meaning. These results are especially impressive given that the image data used for these words showed a decent amount of variance, which suggests that the model was able to find structure among the data even when there visually appeared to be none. The results for these words are also unique in that they are the only words for which the model determined more than one representative color. This seems to be a desirable feature for a model of word-color associations, as one would expect certain words to be associated with more than one color (e.g. generally bright or dark colors) and might get misleading color annotations from a model that is only able to predict a single color.

Still, some words that exhibit strong agreement in their human color annotations do not have these psychological associations reflected in the results of this model. While the black clusters returned for "rain" and "sad" are not necessarily amiss as these words have darker, more negative connotations, they do not corroborate with the human color annotations, which both strongly clustered around shades of blue. This is also odd because "rain" is a concrete word tied to water which is typically represented as being blue, and "sad" has long been tied to the color blue through some of our most popular linguistic metaphors (e.g. "I'm feeling blue"). Another disagreement is seen in the results for the word "pain", which this model finds is best represented by shades of dark bluish-purple. However, "pain" is an emotion-laden word for which the human color annotations show strong associations with the color red. One would also expect this word to be associated with red through images of blood or the heart, as this how humans experience pain. It is more difficult to say what one would expect the color annotations for the abstract words "running" and "soaring" to be. Regardless, my results show that this model was generally less effective at predicting colors for these words, as the most representative points for "running" are mostly grey with a few light blues and for "soaring" are mostly black with a hint of sky blue. This is a little odd given that these are words that are easily represented in images and should include significant amounts of the environment the activity is taking place in, which is what I hypothesized might inform the human color associations for these words.

5.4.2 Statistical Analysis

The distances between the average human color annotations and the average Bayesian representativeness color show that the model's output for the words "big", "angry", and "justice" are statistically significant, with "universe" and "door" very nearly so. For "big" and "justice" this seems to be indicative of the same issue found in the image analysis model, where the combination of high variance in the human color annotations for these words and the most representative points for this model clustering around the origin (especially for "justice") trivially results in low distance values. However, the statistical significance of "angry" is among the first promising quantitative results. Both the LAB spaces for this model and the human annotation data strongly cluster around the same shades of red, showing that this model's prediction aligns with the psychological color association. Further, we can see that some of the largest distances are for abstract, human constructs like "adjective" and "noun", which is more reflective of the high variance in the human color data than previous models' results. Still, the results for a number of concrete words are not statistically significant, despite the fact that my qualitative analyses show promising results. I address these concerns in the following section through a comparative analysis of the results of the three models presented in this paper.

6 Comparative Analysis & Discussion

Figure 23 summarizes the results of the three models described in this paper. At first glance, we can see that color predictions produced by the distributional semantic model have the least variation across the tested words and are all mostly the same shade of pinkish-brown. The results of the image analysis model show slightly more variation in color output, but is still limited to darker shades of neutral colors. Only with the Bayesian representativeness model do we finally see more distinctive, saturated color predictions. For the concrete words "lime", "angry", "berry", "cloud", and "sky" these color predictions even match the colors of the real world references that one would expect to ground these words' color associations. This qualitative assessment alone shows much promise for the use of Bayesian representativeness.

Though the models fared poorly according to their individual statistical analyses, comparing the distance between the average human color annotation and the average color of each model's response for particular words further suggests that my Bayesian representativeness model is successful as a computational model for human word-color associations. In the results for the concrete word "lime" we can see that the distance between human annotation and the model's prediction goes from around 90 for the distributional semantic model, to around 45 for the image analysis model, and finally down to around 30 for the Bayesian representativeness model. We see similar results for the word "angry", with around 55 for the distributional semantic model, around 50 for the image analysis model, and less than 20 for the Bayesian model. For "tree", the distributional semantic model has a distance of around 50, the image analysis model a distance of around 40, and the Bayesian model a distance of around 30. Even when the distributional semantic model achieved a reasonable color prediction for a word, as in the case of the word "door", the Bayesian model was far superior in its distance to the average human color annotation ($\tilde{4}0$ vs. $\tilde{2}0$). The Bayesian model performs better even for the abstract words "fun", "justice", "running", and "verb". The fact that this model consistently produces lower distances for concrete words whose human color annotations show strong clustering, means that this model is in fact capturing humans' color associations for these words. Further, the Bayesian representativeness model still improves upon the performance of the image analysis model for all the above mentioned words. This means that selecting colors from reference images alone is not enough to model human word-color associations. The Bayesian representativeness model is able to discern valuable structure within the input data that makes its output a better representation of the human associations we seek to model.

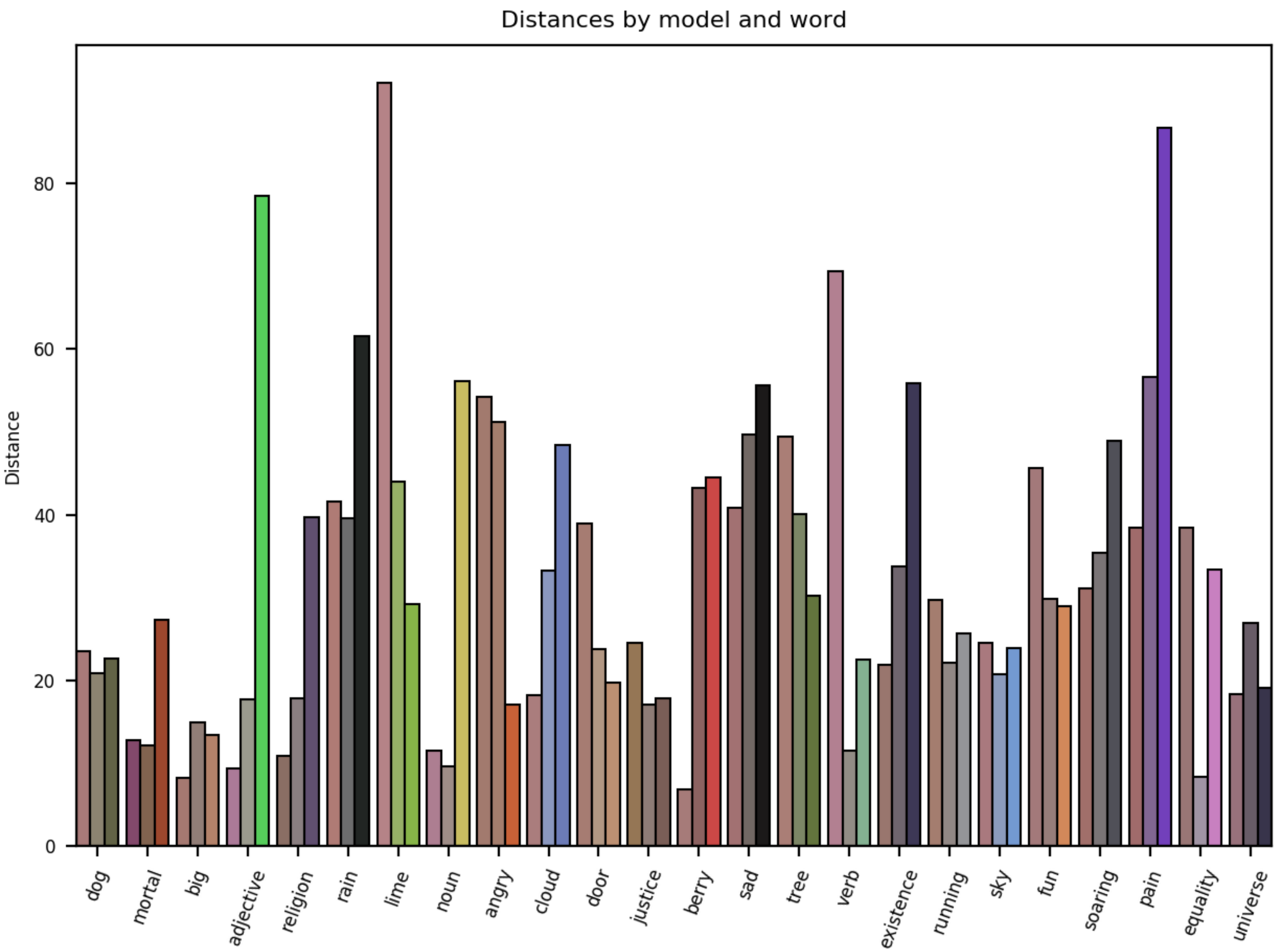

Figure 23: Multi-bar plot of the distances between the average human color annotation and the average color of each model's response. In each group of bars for each word, the left-hand bar shows the results of the distributional semantic model, the middle bar shows the results of the image analysis model, and the right-hand bar shows the results of the Bayesian representativeness model. The bars are colored with the average color produced by the model for that word.

This comparative analysis also suggests that the problem in my individual statistical analyses may lie in having evaluated the models against the human color annotations I collected. Of course, human color annotations should provide the ultimate comparative baseline, as a good computational model of human word-color associations should emulate humans' representation of these modalities. However, it seems that a significant issue with this dataset was the limited number of responses I was able to collect through my networks. This small sample size may have led to especially high variance in the responses for many of the tested words, resulting in average color annotations that were not very useful as a comparative baseline for my models' performances. One solution would be to conduct the statistical analyses against each of the human color annotation responses, rather than their averages. An alternative human evaluation procedure, where humans rate the results of each model, would also help to verify the qualitative analyses detailed in this paper. However, I believe efforts to obtain a much larger set of human word-color annotations would be most useful. In the future, in addition to addressing the areas of improvements covered in Section 3.5, I plan to use Amazon Mechanical Turk to obtain a greater number of color annotations for a larger set of words. Additionally, this would allow me to study how color associations might differ across different populations, as in this work, I was only able to survey western, college-educated adults.

7 Conclusion

In this work, I develop a computational model for human word-color associations that is grounded in humans' visual experience of color, rather than a lexical proxy for color. To do this, I first establish a preliminary dataset of human color annotations for a set of words that represents varying degrees of abstractness. I find that while image analysis provides a more automatic, less labor-intensive alternative to obtaining color data, it does not sufficiently represent humans' color associations. Future work to develop a larger set of human color annotations through crowd sourcing, and perhaps supplementing this with color data from images, will provide the foundation necessary to develop more robust models of human word-color associations.

Of the distributional semantic, image analysis, and Bayesian representativeness models that I implement in this paper, I find that the Bayesian representativeness model does the best job of producing unique color annotations for a variety of word types and outperforms the distributional semantic and image analysis models in terms of proximity to average human color association for these words. While distributional semantics has been the standard for previous models of human word-color associations, my results show that the Bayesian representativeness model is better able to discern meaningful structure in input color data that allows it to more closely emulate humans' color associations for words.

8 References

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, 1, 238-247.

Bora, D. J., Gupta, A. K., & Khan, F. A. (2015). Comparing the Performance of L*A*B* and HSV Color Spaces with Respect to Color Image Segmentation. *International Journal of Emerging Technology and Advanced Engineering*, 5(2), 192-203.

Bruni, E., Tran, N., & Baroni, M. (2014). Multimodal Distributional Semantics. *Journal of Artificial Intelligence Research*, 49, 1–47.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2013). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3), 904-911. doi:10.3758/s13428-013-0403-5

Eckhard, T. (2014). Plot Lab color coordinates. MATLAB and Statistics Toolbox Release R2013a.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in Semantic Representation. *Psychological Review*, 114, 211–244.

Harris, Z. S. (1954). Distributional Structure. *WORD*, 10(2-3), 146-162. doi:10.1080/00437956.1954.11659520

Hemphill, M. (1996). A Note on Adults' Color–Emotion Associations. *The Journal of Genetic Psychology*, 157:3, 275-280. doi: 10.1080/00221325.1996.9914865

Hill, B., Roger, T., & Vorhagen, F. W. (1997). Comparative analysis of the quantization of color spaces on the basis of the CIELAB color-difference formula. *ACM Transactions on Graphics (TOG)*, 16(2), 109-154.

Ibraheem, N. A., Hasan, M. M., Khan, R. Z., & Mishra, P. K. (2012). Understanding Color Models: A Review. *ARPN Journal of Science and Technology*, 2, 265-275.

Iliev, R., Axelrod, R. (2017). The Paradox of Abstraction: Precision Versus Concreteness. *Journal of Psycholinguistic Research*, 46, 715–29.

Kousta, S., et al. (2011). The Representation of Abstract Words: Why Emotion Matters. *Journal of Experimental Psychology*, 140(1), 14–34.

Labrecque, L., & Milne, G. (2012). Exciting red and competent blue: the importance of color in marketing. *J. of the Acad. Mark. Sci.*, 40(5), 711–727.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211–240.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *NIPS*, 3111–3119.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *CoRR.*

Mohammed, S. (2011). Colourful Language: Measuring Word–Colour Associations. *In Proceedings of the ACL 2011 Workshop on Cognitive Modeling and Computational Linguistics (CMCL).*

Ozbal, G., Strapparava, C., Mihalcea, R., & Pighin, D. (2011). A Comparison of Unsupervised Methods to Associate Words with Colors.

Pennington, J., Socher, R., & C. Manning, C. D. (2014). GloVe: Global vectors for word representation.

Plataniotis, K. N., & Venetsanopoulos, A. N. (2000). Color Image Processing and Applications. Berlin: Springer.

Riley, C. (1995). *Color Codes: Modern Theories of Color in Philosophy, Painting and Architecture, Literature, Music, and Psychology.* University of New England Press.

Schanda, J. (2007). Colorimetry: Understanding the CIE system. Hoboken, New Jersey: Wiley.

Schloss, K. B., & Palmer, S. E. (2017). An ecological framework for temporal and individual differences in color preferences. Vision Research, 141, 95-108.

Sutton, T. M., & Altarriba, J. (2015). Color associations to emotion and emotion-laden words: A collection of norms for stimulus construction and selection. *Behavior Research Methods*, 48(2), 686-728. doi:10.3758/s13428-015-0598-8

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of Visual and Linguistic Information in Spoken Language Comprehension. Science, 268(5217), 1632–34.

Tenenbaum, J.B., & Griffiths, T.L. (2001). The rational basis of representativeness. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society.*

Witthoft, N., & Winawer, J. (2013). Learning, Memory, and Synesthesia. Psychological Science, 24(3), 258-265. doi:10.1177/0956797612452573