

# Applications of a Color-Naming Database

*Nathan Moroney and Ingeborg Tastl  
Hewlett-Packard Laboratories  
Palo Alto, CA, USA*

## Abstract

An ongoing web-based color naming experiment<sup>†</sup> has collected a small number of unconstrained color names from a large number of observers. This has resulted in a large database of color names for a coarse sampling of red, green and blue display values. This paper builds on a previous paper,<sup>1</sup> that demonstrated the close agreement for this technique to earlier results for the basic colors, and presents several applications of this database of color names. First, the basic hue names can be further subdivided based on a number of modifiers. Pairs of modifiers are compared based on actual language usage patterns, rather than on a fixed hierarchical scheme. Second, given a sufficient sampling of color names using memory color modifiers such as sky or grass, comparisons can be made to other studies on memory colors and preference for color reproduction. Finally, a dissimilarity coefficient can be computed for a set of 27 color names. Multidimensional Scaling can be applied to the matrix leading to a spatial configuration, which is solely based on patterns of color naming.

## Introduction

Color naming or the linguistic encoding of color perception is a rich cross-disciplinary research area. For color imaging applications, color naming has been used in the user interface design,<sup>2</sup> in image segmentation,<sup>3</sup> and as a gamut mapping constraint.<sup>4</sup> Furthermore, it relates to some of the high level categories of preferred color reproduction. This paper extends these applications raising naming models to the next level. It supplements previous results of preferred color reproduction and serves as a new tool for the assessment of hue linearity of color spaces.

A previous paper<sup>1</sup> presented the methodology of deriving the color naming database. In addition, the agreement of the derived basic color centroids was compared to previous laboratory<sup>5-7</sup> results and the correlations were found to be quite high. The correlations for CIELAB hue and lightness were as good for the web-based experiment as the other experiments agreed with each other. The basic task was unconstrained naming of seven color patches on a white background. The seven colors were randomly generated from a 6 by 6 by 6

sampling of RGB values. Currently over 1000 participants have provided color names. One raw data element consists of an RGB triplet or node and a corresponding string of color names. A nominal sRGB<sup>8</sup> display was used. The performed analysis suggests that the exact characteristics of the assumed nominal display will minimally impact the analysis as it relates to the relatively large color name categories.

Following terminology proposed by Berlin and Kay,<sup>5</sup> the basic colors are those that are hypothesized to be shared by all fully developed languages. These colors are red, green, blue, yellow, white, black, gray, orange, pink, brown and purple. There is a further hypothesis that these names tend to enter into languages in a somewhat fixed sequence. A modifier is any term used in combination with a color name to refine the meaning of the color name, such as “light” or “dark”.

## Naming Models

Monolexical or one-word basic color names were used extensively by the observers in the web-based visual experiment. However additional modifiers and non-basic color names were also frequently used. In fact the terms “light” and “dark” were among the top ten most frequently occurring terms. These lightness modifiers were used with all of the basic color names and many of the non-basic color names. In addition, a number of other modifiers related to tangible objects were used together with multiple basic or non-basic color names.

There is current research<sup>9,12</sup> in the area of color naming models. These models vary in complexity and computational techniques. Some of these models are actually color vocabularies or systematic syntaxes for color. The use of modifiers has either not been addressed in detail or assumptions have been made about their use. Given a substantial color-naming database, it is of interest to infer patterns of modifier usage.

For example, it is possible to search the database for all instances of a modifier and compute summary statistics for the lightness, chroma and hue values in which these modifiers were used. Further it is possible to perform means testing of the data to better understand how the modifiers are used. A partial list of frequently used terms includes: “light”, “dark”, “bright”, “neon”, “deep”, “pale”, “medium”, “fluorescent”, “electric”, “true”, “dull”, “burnt”, “pastel” and “hot”. In the case of “light” and

<sup>†</sup> [http://www.hpl.hp.com/personal/Nathan\\_Moroney/color-name-hpl.html](http://www.hpl.hp.com/personal/Nathan_Moroney/color-name-hpl.html)

“dark” there is consistent widespread use of the term. In other cases, such as “burnt” or “hot” the modifier is very specific. Finally there are terms with considerable degrees of overlap, such as “pastel” and “pale” or “neon” and “fluorescent”. A table comparing the mean lightness, chroma and hues for a subset of the above modifiers is shown below..

**Table 1.** Result of means testing for modifier pairs at a 95% confidence level.

Modifier pair	Lightness	Chroma
Pale-Pastel	Equal	Equal
Light-Bright	Equal	Not Equal
Dark-Deep	Equal	Not Equal
Deep-Dull	Not Equal	Equal
True-Medium	Equal	Not Equal

Additional graphical analysis of these terms supports the results shown in Table 1. “Pale” and “pastel” appear to be used interchangeably. “Bright” is clearly preferred for use with higher chroma light colors while “deep” is used more frequently for higher chroma dark colors. In contrast, “dull” and “deep” are used for similar chroma colors but “dull” is used for lighter colors. Finally, the term “true” is on the average used with higher chroma colors than the term “medium”. The highest chroma colors also tend to have modifiers that are partially a function of the hue angle, such as “hot”, “neon”, “electric” and “fluorescent”.

These results suggest that color naming for multiple word color names is less systematic than many of the proposed vocabularies or syntaxes. For instance the terms “very”, “vivid”, or “brilliant” occur very infrequently. Furthermore, terms with statistically significant differences for lightness or chroma still have a high degree of overlap.

## Naming and Preference

A number of researchers<sup>13-19</sup> have investigated the topics of memory color, preference, and in some cases how these concepts relate to color reproduction. The large number of possible object colors has generally been limited to sky blue, grass green and flesh tones. Understanding how these critical colors should be reproduced was treated as a possible general guide for color reproduction. It has been noted that there are deviations between preferred colors and actual object colors. One trend is to prefer more chromatic or saturated reproductions relative to the original. Furthermore, some studies have suggested small hue shifts for sky blue and grass green. Psychologists being interested in the memorization process itself have observed color shifts in the direction of the most impressive chromatic attribute of an object. In the area of color reproduction color preferences of human observers have been investigated for many years. Interestingly, although different methods have been applied, the results once transformed into one

common form reveal similar trends. In specific, researchers have used monochromators asking subjects to adjust the wavelength of the emitted light in accordance with their long-term memory of colors of familiar objects.

Figures 1 through 3 show comparisons of the web-based naming results relative to previous results. These figures are in the  $u'v'$  space with  $u'$  on the x-axis and  $v'$  on the y-axis. The data for Hunt's original colors are shown with open diamonds while the Hunt memory color is shown with open squares. The range from the Hunt data is shown with a solid line. The de Ridder data is shown with a thick solid line and the Seliger data is shown with a thick dotted line. The Rider data for Figure 3 is for a single wavelength and is therefore a line instead of an area. Finally, the web-based results are shown with a solid line with filled circles.

The average wavelength plus/minus variance of a set of observers for green grass and blue sky has been converted to  $u'v'$  values and is displayed in Figure 1 and 2 together with connecting lines towards D50. Other researchers, like Hunt, Bartelson and Topfer have used actual images, varied the reproduction of skin tones, blue sky and green grass and solicited preference judgments from observers. Data from Hunt have been converted from illuminant C to illuminant D50 and are approximately reproduced in figure 1 to 3. Yet another approach is to investigate the relationship between color reproduction and naturalness of an image. Yendrikhovskih and de Ridder et al. identified certain groups of colors being related to natural objects. According to their study water, sky and distant objects fall into a hue region of 460-485nm, green plants fall into the yellowish green region of 550-575nm, where as earth and dried vegetation fall into an orange reddish bin of 575 to 590nm, and skin tones are gathered around a dominant wavelength of 590nm.

The transformed values of those numbers are displayed in figure 1 to 3. One could argue an in question the precise numbers, but the interesting and important point is that the results of several of those studies performed by different groups and with different techniques line up quite well. The common point between data from those previous studies and the data extracted from the current one are judgments based on memory colors. Yet, the difference is that instead of relying on preference judgments of reproduced images the current study is based on the verbal presentation and memory of familiar objects. That is similar to the monochromator experiments, but it is augmented with the potential of data gathering over the web and can thus result in large data sets, which have the potential to reveal interesting facts about memory colors and color preferences.

## Multi-Dimensional Scaling

The raw data can also be submitted to multi-dimensional scaling.<sup>20-22</sup> This allows the spatial configuration of a set of color names to be inferred from patterns in color naming. This can be done without assuming a nominal display. Instead the only assumptions are that a nominal display exists and that the errors in color naming are randomly distributed. This analysis makes use of the statistical power of a large number of color names.

Given that the database consists of only color names and nodes, the first step is inferring a similarity measure between two color names. Measuring similarity<sup>23</sup> is not straightforward but one possibility is to use a Jaccard coefficient.<sup>24</sup> This coefficient was developed by botanists to compare species diversity across two geographic regions. It has been used by other fields and is the number of shared members for two sets divided by the total number of members. In this case a color name is equal to a set and the members are specific RGB nodes for which that name was used. For more similar color names the number of shared RGB nodes will be greater while for dissimilar color names the number of shared nodes will be minimal. This can be expressed as :

$$S_J = \frac{a}{b + c} \quad (1)$$

where  $S_J$  is the similarity between two color names,  $a$  is the number of nodes shared by the two names,  $b$  is the number of nodes for the first color name and  $c$  is the number of nodes for the second color name. As a reminder, the RGB nodes are the 216 colors corresponding to the 6 by 6 by 6 sampling used in the visual experiment. Note that the above calculation only uses presence or absence and does not make use of the frequency of name usage at each of the nodes. More complex measures are feasible and will be used in future investigations.

Equation one results in a similarity matrix for a set of color names. A total of 27 color names were chosen and the frequency of the least frequently occurring basic color, white, was used as a lower limit. The specific color names used were: red, green, yellow, blue, orange, purple, pink, brown, black, white, gray, olive, magenta, sky blue, lime green, navy blue, violet, teal, forest green, maroon, mauve, tan, grass green, lilac, peach, fuchsia and cyan. A dissimilarity matrix was computed by subtracting the values of the similarity matrix from the maximum similarity value. The diagonals of the matrix were set to zero. The dissimilarity matrix was then input to the SPSS AIscale multi-dimensional scaling routine.

The results of the first two dimensions are shown in Figure 2 for a subset of the color names. This figure and more detailed plots of all color names show a considerable agreement with the color name spacing that results from

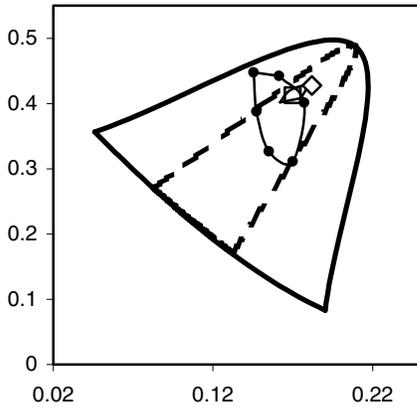


Figure 1. Comparison of web-based color naming results relative to previous results for blue sky.

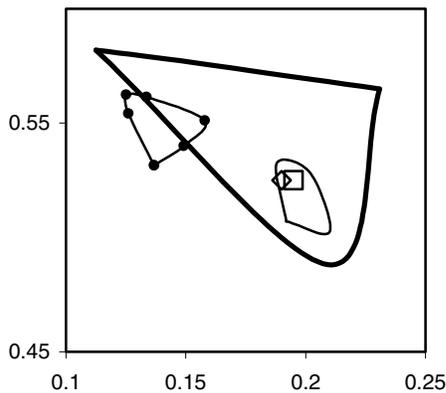


Figure 2. Comparison of web-based color naming results relative to previous results for green grass.

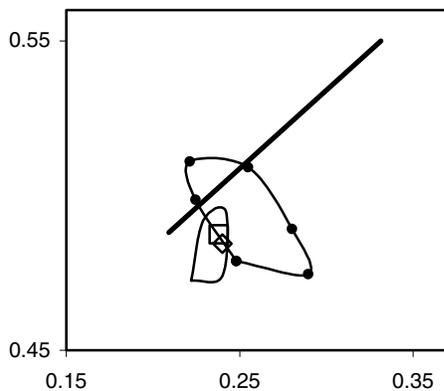


Figure 3. Comparison of web-based color naming results relative to previous result for skin tones.

assuming a nominal display and computing name centroids in different color spaces. Furthermore the hue linearity of CIELAB and CIECAM02 can be compared with each other using this data as independent data set. The significant gap in CIELAB hues of almost 180 degrees between blue and green is not evident in the CIECAM02 plot or in the spatial configuration show in Figure 2.

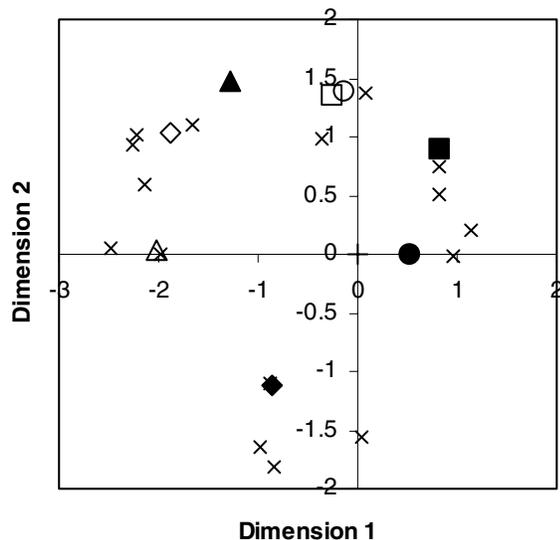


Figure 2. Spatial reconstruction of 27 color names based on multi-dimensional scaling of the dissimilarities as estimated applying a Jaccard coefficient to the color names and nodes. The non-basic colors are shown with an 'x' while the white and black points are located at the origin. The basic colors are shown with filled and open shapes

Figure two was created based on the first two dimensions of the MDS output. The non-basic colors are shown plotted with "x"s. The basic colors of red, yellow green and blue are shown with a filled triangle, square, circle and diamond, respectively. The basic colors of orange, brown, pink and purple are shown with an open circle, square, diamond and triangle, respectively. The reconstruction of the color name hue angles is quite impressive, especially for the non-basic color names. There are some definite differences corresponding to chroma but these are minimal and likely relate to the apparent lack of basic and non-basic color names relating to low chroma colors. The reconstruction of the third dimension or lightness was also approximate. For figure 2 the raw data was translated and rotated such that black and white points are both at the origin for the first two dimensions. The location for gray was roughly collinear with black and white, but not exactly and therefore the origin was plotted relative to the white and black. More investigation is required to verify the reconstruction of the

lightness axis. However, the results in figure two show various asymmetries,<sup>25,26</sup> and clustering and raise the question of the orthogonality of the red-green and yellow-blue axis.

## Conclusion

A large color-naming database has been used to consider how naming models might be extended to include modifiers. It is hypothesized that color syntax is not as detailed or hierarchical as previously published vocabularies suggest. The database has also been used to infer memory colors and these memory colors have been compared to those published in previous studies on color preference. Finally, the color naming was used as input to a multi-dimensional scaling function. The result is a spatial configuration that is based only on patterns of coincidence in color naming.

## References

1. N. Moroney, "Unconstrained web-based color naming experiment", (2003).
2. Toby Berk, Lee Brownston and Arie Kaufman, "A human factors study of color notation system for computer graphics", *Communications of the ACM*, **25**(8), 547-550 (1982).
3. A. Mojsilovic, "A method for color naming and description of color composition in images", *Proc. Int. Conf. Image Processing, ICIIP 2002*, Rochester, New York, (2002).
4. H. Motomura, "Analysis of gamut mapping algorithms from the viewpoint of color name matching", *J. SID*, **10**(3), 247-254 (2002).
5. Brent Berlin and Paul Kay, *Basic Color Terms: Their Universality and Evolution*, CSLI Publications, Stanford, California (1999).
6. Robert M. Boynton and Conrad X. Olson, "Locating basic colors in the OSA Space", *Color Res. App.* **12**(2) 94-105 (1987).
7. Julia Sturges and T.W. Allan Whitfield, "Locating basic colours in the Munsell Space", *Color Res. App.* **20**(6), 364-376 (1995).
8. International Electrotechnical Commission, *Part 2-1: Colour Management - Default RGB Color Space- sRGB*, IEC 61966-2-1, (1999).
9. H. Lin, M.R. Luo, L.W. MacDonald, and A.W.S. Tarrant, "A cross-cultural colour-naming study. Part III - a colour-naming model", *Color Res. Appl.* **26**(4) 270-277 (2001).
10. Robert Benavente, Francesc Tous, Ramon Baldrich and Maria Vanrell, "Statistical model of a color naming space", *Proc. CGIV 2002: The First European Conference on Colour in Graphics, Image and Vision*, 406-411 (2002).
11. Shoji Tominaga, "A color-naming method for computer color vision", *Proc. IEEE Int. Conf. On Cybernetics and Society*, 573-577 (1985).
12. K. Okajima, A.R. Robertson, G.H. Fielder, "A quantitative network model for color categorization", *Color Res. App.* **27**(4), 225-232 (2002).

13. D.L. MacAdam, "Reproduction of colors in outdoor scenes", *Proc. IRE*, 166-174 (1954).
14. R.W.G. Hunt, I.T. Pitt and L.M. Winter, "The preferred reproduction of blue, sky, green grass and Caucasian sky in colour photography", *J. Photogr. Sci.*, **22**, 144-149 (1974).
15. M. Yamamoto, Y. Lim, N. Aoki and H. Kobayashi, "On the preferred reproduction of flesh color in Japan and South Korea, investigated for all generation", *J. Soc. Photogr. Sci. Japan*, **65**(5), 363-368 (2002).
16. K. Topfer and R. Cookingham, "The quantitative aspects of color rendering for memory colors", Proc. IS&T PICS conference, 94-98 (2000).
17. S.N.Yendrikhovskij, F.J.J. Blommaert, and H. de Ridder, "Color reproduction and the naturalness constraint", *Color Res. App.* **24**(1), 52-67 (1999).
18. P. Bodrogi and T. Tarczali, "Colour memory for various sky, skin, and plant colours: effect of the image context", *Col. Res. Appl.* **26**(4), 278-289 (2001).
19. H.H. Seliger, "Measurement of memory of color", *Color Res. App.* **27**(4), 233-242 (2002).
20. J.B. Kruskal and M. Wish, *Multidimensional Scaling*, Sage University Paper series on Quantitative Applications in Social Sciences, Series no. 07-011, Newbury Park, CA (1993).
21. T. Indow, "Predictions based on Munsell notation. II. Principle hue components", *Color Res. App.*, **24**(1), 19-32 (1999).
22. T. Indow, "Principle hue curves and color difference", *Color Res. App.* **24**(4), 266-279 (1999).
23. S. Santini and R. Jain, "Similarity measures", *IEEE Trans. Pattern Analysis and Machine Intelligence*, **21**(9), 871-883 (1999).
24. P. Jaccard, "The distribution of the flora of the alpine zone", *New Phytologist*, **11**, 37-50 (1912).
25. P. Kay, "Asymmetries in the Distribution of Composite and Derived Basic Color Categories", *Behavioral and Brain Sciences* **22**, 957-958 (1999).
26. C.S. McCamy, "The primary hue circle", *Color Res. Appl.* **18**(1), 3-10 (1993).