Validating Large-Scale Lexical Color Resources

Nathan Moroney, Giordano Beretta

HP Laboratories
HPL-2011-226

Keyword(s):
color naming; crowd-sourcing

Abstract:

The use of the Web for crowd-sourcing lexical color resources has succeeded in creating databases consisting of millions of color terms. Various researchers have demonstrated the value of this data, but questions related to the quality and reliability of the data remain, because each large survey is tainted by a small number of disruptive subjects. The challenge is to cull the resource by identifying and eliminating the data contributed by these disruptive subjects. With a million color terms, it is no longer possible to individually inspect color terms and we need an automated process. Machine evaluation through natural language processing is possible, but this introduces the added complexity of pre-defining properties and criteria for data validity, which could improperly cloud the data. Color terms are terms associated with colors. Instead of examining the terms, we can examine their colors. Our visual system can process purely visual information at a much higher bandwidth, because the language system and its complex cognitive processes can be bypassed. In this contribution we propose a graphical approach in which the associated colors of large-scale lexical resources are first machine-sorted by color appearance so that human experts can efficiently identify outliers or questionable entries by simply looking at a graphical rendering. A recent test with the R. Munroe and E. Ellis Color Survey Data has allowed us to process over a million color terms. The methodology is as follows. First, the color terms are binned categorically, where each bin corresponds to a monolexemic color term. Second, for each term the associated red, green and blue sRGB values are further quantized and then these device values are sorted in lexicographical order. Third, the sorted device values are displayed as raster images in which each term is represented by a pixel drawn in the associated color. Finally, observers identify visually the outliers for each color term by inspecting the raster image. Using this procedure, the relatively rare disruptive subjects are efficiently identified and tagged. This process can be extended to multiple experts and a weight can be derived for the entries in the lexical resource. Experiments show that even using such crude appearance attributes as the sRGB values, the methodology is very effective and it is not crucial to use more sophisticated representations, such as for example correlates of hue, lightness and chroma. Based on this methodology, we show that the Munroe and Ellis Color Survey Data correlates well with data obtained in a controlled laboratory experiment. This is a surprising result given the informal nature of this resource. It is also a testimony of the validity of crowd-sourcing for scientific experimentation.

External Posting Date: December 8, 2011 [Fulltext] Approved for External Publication
Internal Posting Date: December 8, 2011 [Fulltext]

Published in Interaction of Colour & Light in the Arts and Sciences AIC Midterm Meeting 2011

© Copyright 2011 Hewlett-Packard Development Company, L.P.

Validating Large-Scale Lexical Color Resources

Nathan MORONEY, Giordano BERETTA
Hewlett-Packard Company

Abstract

The use of the Web for crowd-sourcing lexical color data has succeeded in creating databases consisting of millions of color terms. Various researchers have demonstrated the value of this data, but questions related to the quality and reliability of the data remain, because each large survey is tainted by a small number of disruptive subjects. We report on a controlled experiment validating a very large database.

1. Introduction

Traditionally, color naming experiments—e.g., the World Color Survey—have used reflective samples such as chips from the Munsell Book of Color or OSA UCS color chips. There has been research in the use of additive displays for color naming (Post et al. (1986), Mojsilovic (2005), Benavente et al. (2005)) but that work is not based on large numbers of observers, especially compared to the thousands of participants in the World Color Survey.

Our motivation is the creation and use of very large lexical color resources specific to displays, and builds on a decade long effort to collect unconstrained color names on the World Wide Web by Moroney (2003). Recent results by Mylonas (2010) suggest excellent agreement for two different Web-based color naming experiments. A further result[1] published on the Web by Munroe and Ellis (2010) has increased the scale of data publically available for analysis from thousands to millions of participants. This has considerable promise for a detailed understanding of the use of color terms, but requires analysis and thought with respect to a systematic method to validate the data. Due to the data size, manual inspection of individual responses is not an option. Our objective is to effectively and efficiently perform a laboratory validation of this large scale uncontrolled Web data, following Zuffi et al. (2007). This is critical given the informal nature of the survey.

2. Experiment

The Munroe and Ellis (2010) color survey asked volunteers to name color patches on black and white backgrounds. Participants were free to name as many or as few patches as they wanted and were not constrained to which terms they could use. Full details of the survey can be found on the archived Web page. The result is a relational database consisting of over 3.5 million terms, or the size of the population of Berlin. The database also includes optional demographic data such as gender, which is not considered in this paper.

We processed the database using in the following steps. First, we applied minimal data cleaning, such as conversion to lower case. Next we queried the cleaned database for

[1] Munroe is the artist and writer of the Web-comic **xkcd** and Ellis implemented the software to conduct the survey. The survey was announced on Munroe's blog and data was collected for one week starting March 1, 2010. We refer to the data publically posted to the Web as Munroe & Ellis data to acknowledge their contribution in a format consistent with technical citations.

the basic color terms, which are the subject of many laboratory studies. Only exact matches were recorded, yielding a monolexemic basic color term database of 1.3 million entries. This database consists of red, green and blue 8-bit values displayed during the survey and the elicited basic color terms. The next step was a 6×6×6 quantization of the red, green and blue values, providing a simple means to sort and cluster the data, such as the novel visualization scheme in which each participant is a pixel shown in Fig. 1. The color terms are shown as square regions arranged alphabetically from left to right. Each term is frequency-sorted from top to bottom, providing a useful representation of the database.

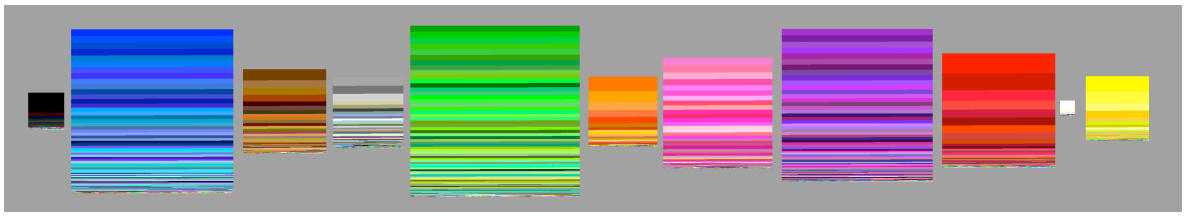


Figure 1. Sorted frequency visualization.

Our validation experiment was a multi-stimulus categorization task. Given a color term, observers were instructed to identify which color patches should be assigned to that category. A screen shot for the term green is shown in Fig. 2, which shows a screen with a summary of the instructions on the top and a 20 wide array of color patches of subtending roughly 2°. Each color patch had a check box; once observers selected the color patches corresponding to that term they progressed to the next term. The experiment was conducted on an HP Compaq LP2480zx display in sRGB conditions. A total of 16 color normal observers participated in the experiment.

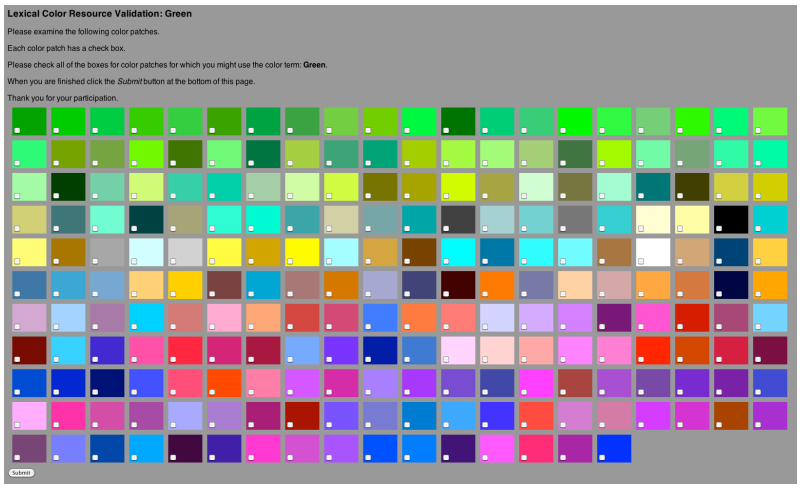


Figure 2. Screen shot for color term green.

We avoid defining a disruptive participant. The sorted frequency images in Fig. 1 suggest that there are many types of observers. There are oranges of more frequent shades of brown than color normal observers might agree upon. However, there are also greenish browns that are probably the result of color deficient observers or systematic differences in display primaries. Finally, at the bottom are the colors that are clearly not on the dichromatic confusion lines and likely to stem from adversarial participants. In our instructions to the observers, we did not define what is normal or disruptive. We simply instructed, "select the color patches you might use with the color term."

3. Results and Discussion

The results for the experiment will be considered in three sets of graphs. Fig. 3 shows the

CIELAB hue, lightness and chroma correlates. The CIELAB coordinates were computed averaging in RGB space all corresponding chromatic basic color terms. The graphs show the Berlin and Kay averaged centroids on the abscissa and the Munroe and Ellis averages on the ordinate. They also show a linear fit and the corresponding r^2 value. The hue values have the highest correlation, with $r^2 = 0.97$, and the chroma values have the lowest correlation. The hue results are comparable to previously published correlations of the Berlin et al. (1999) data with the studies by Sturges et al. (1995) and Boynton et al. (1987).

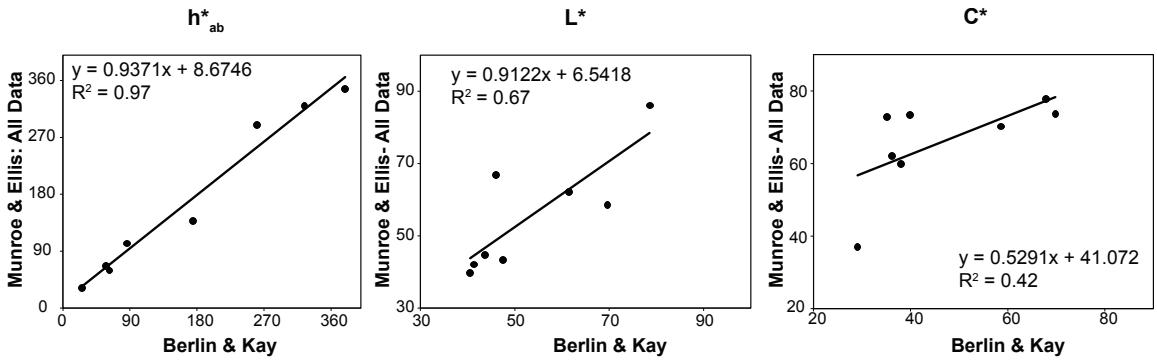


Figure 3. CIELAB hue (left), lightness (center) and chroma (right) correlations for the complete Munroe and Ellis data versus the Berlin and Kay data.

The results for the data validated in our experiment are shown in Fig. 4. The data was computed taking only the data elements selected for a given color term, averaged over all observers. The amount of data retained after validation is shown in Tab. 1. This a significant reduction in the amount of data used for calculation of the corresponding centroids.

Table 1. Percent color data retained after validation per color term.

Brown	*Purple*	*Pink*	*Orange*	*Blue*	*Yellow*	*Red*	*Green*
33%	*44%*	*49%*	*49%*	*53%*	*70%*	*73%*	*74%*

In Fig. 4 the ordinate is the validated Munore and Ellis data. The r^2 correlation values are comparable to those for the complete data set shown in Fig. 3. This raises the question: how comparable are the original and validated data sets?

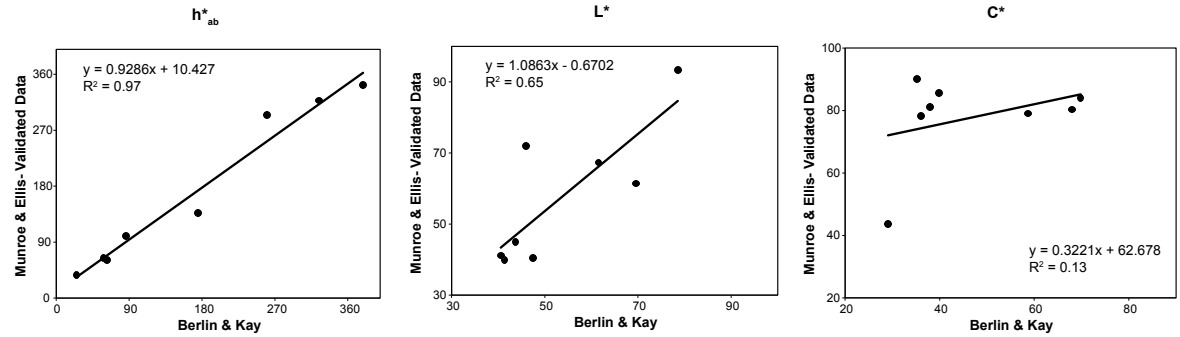


Figure 4. CIELAB hue (left), lightness (center) and chroma (right) correlations for the validated Munroe and Ellis data versus the Berlin and Kay data.

Fig. 5 shows the result of comparing the complete vs. validated Munroe and Ellis data sets. The original complete data is shown on the abscissa and the validated data is shown on the ordinate. The results for the hue and lightness are $r^2 = 0.99$. Interestingly the results for the chroma are less correlated with $r^2 = 0.82$. We lack an explanation for this consistent shift to more chromatic centroids. The basic result however may ironically be that we have

experimentally validated that validation may not be necessary for this data.

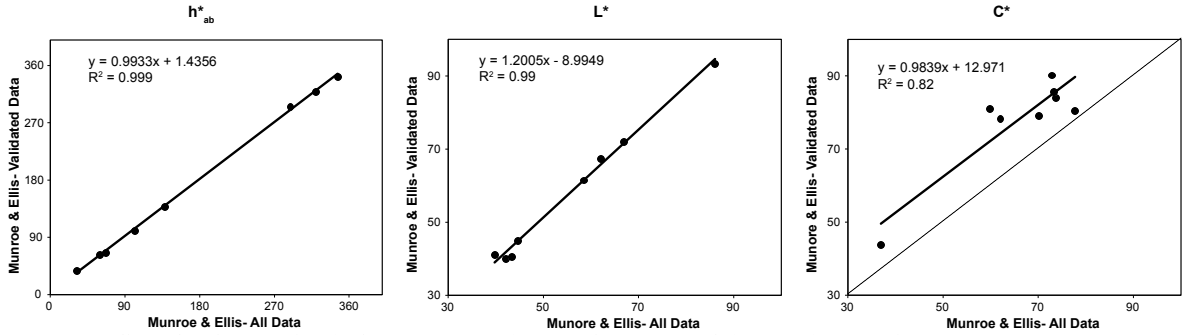


Figure 5. CIELAB hue (left), lightness (center) and chroma (right) correlations for complete Munroe and Ellis data versus the validated Munroe and Ellis data.

4. Conclusions

We have shown a means to use laboratory studies to clean large-scale uncontrolled data from the Web. The experimental subjects rejected from ⅔ of the uncontrolled data for brown to ¼ of the data for the color term green. The CIELAB hue correlation with the Web data (complete and validated) to Berlin and Kay is shown to be over 0.96. The correlations for lightness are less but still on the order 0.65. Finally the hue and lightness correlations for the complete and validated data sets is shown to be 0.99 suggesting that the scale and limited noise in this case may mean that validation is in fact not necessary.

References

Benavente R., F. Tous, R. Baldrich, and M. Vanrell. 2002. Statistical model of a color naming space. *Proc. CGIV 2002: The First European Conference on Colour in Graphics, Image and Vision*: 406–411.

Berlin, B., and P. Kay. 1999. *Basic Color Terms: Their Universality and Evolution*, CSLI Publications, Stanford, California.

Boynton, R.M., and C.X. Olson. 1987. Locating basic colors in the OSA Space. *Color Research and Application* 12(2): 94–105.

Mojsilovic, A. 2005. A computational model for color naming and describing color composition of images. IEEE Transactions on Image Processing 14(5): 690–700.

Moroney, N. 2003. Unconstrained web-based color naming experiment. In *Color Imaging VIII: Processing, Hardcopy, and Applications, Proceedings*, ed. by R. Eschbach, and G.G. Marcu. San Jose, 36–46.

Munroe, R., and E. Ellis. 2010. http://blag.xkcd.com/. Accessed: March 11, 2011.

Mylonas, D., L. MacDonald, and S. Wuerger. 2010. Towards an Online Color Naming Model. In *Eighteenth Color Imaging Conference: Color Science and Engineering Systems, Technologies, and Applications, Proceedings*, ed. by F. Imai and E. Langendijk. San Antonio, Texas, 140–144.

Post, D.L., and F.A. Greene. 1986. Color-name boundaries for equally bright stimuli on a CRT: Phase I. *SID Digest 86*: 70–73.

Sturges, J., and T.W.A. Whitfield. 1995. Locating basic colours in the Munsell Space. *Color Research and Application* 20(6): 364–376.

Zuffi, S., P. Scala, C. Brambilla, and G. Beretta. 2007. Web-based vs. controlled environment psychophysics experiment. In *Image Quality and System Performance IV, Proceedings*, ed. by L.C. Cui and Y. Miyake. 6494: 49407–49407.

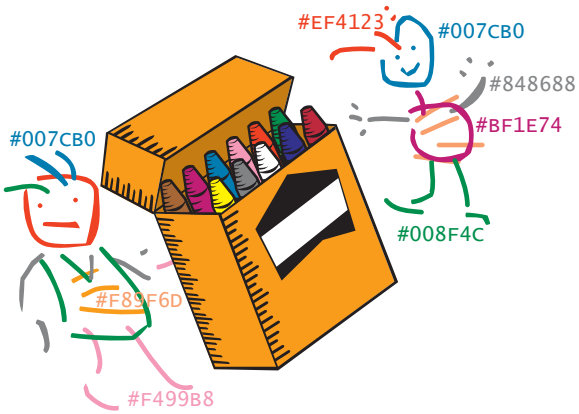
Address: Nathan Moroney, Giordano Beretta, Hewlett-Packard Company
HP Laboratories, 1501 Page Mill Road, m/s 1161, Palo Alto, CA 94304, USA
E-mails: nathan.moroney@hp.com, giordano.beretta@hp.com

Validating large-scale lexical color resources

Giordano Beretta & Nathan Moroney, Hewlett-Packard
Date: 8/June/2011

Color naming

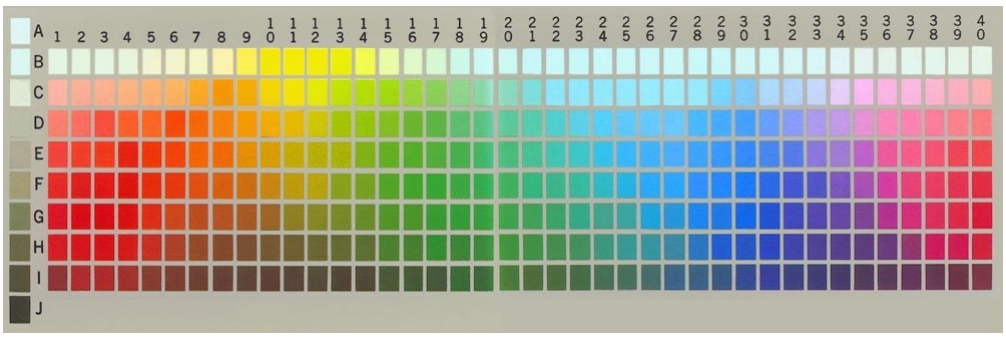
- Colorimetric values are best for communicating color via machine
- Color terms are best for communicating color among humans
- Problem: how can we find the most effective color terms?



© Copyright 2011 Hewlett-Packard Development Company, L.P.
The information contained herein is subject to change without notice.
Created 26/4/2011

World Color Survey, Berlin & Kay, 1969

- Munsell Sheets of Color are shown to respondents to elicit color terms
- A snapshot in time
- Experiment in the wild
- Several similar experiments, e.g., ISCC–NBS



© Copyright 2011 Hewlett-Packard Development Company, L.P.
The information contained herein is subject to change without notice.
Created 26/4/2011

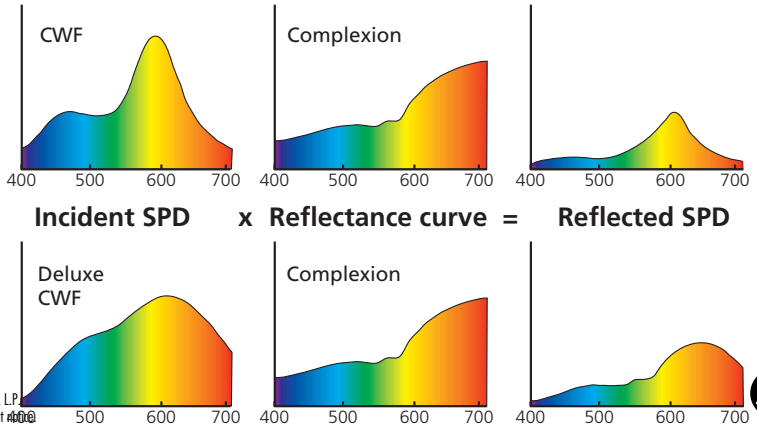
Light + object

The spectral power distribution of the light reflected to the eye by an object is the product, at each wavelength, of the object's spectral reflectance value by the spectral power distribution of the light source

$$R = k \int E(\lambda)S(\lambda)\bar{r}(\lambda)d\lambda$$

$$G = k \int E(\lambda)S(\lambda)\bar{g}(\lambda)d\lambda$$

$$B = k \int E(\lambda)S(\lambda)\bar{b}(\lambda)d\lambda$$

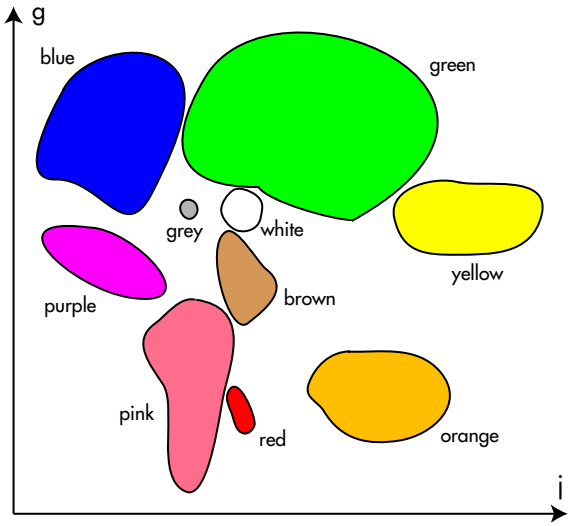


© Copyright 2011 Hewlett-Packard Development Company, L.P.
The information contained herein is subject to change without notice.
Created 26/4/2011

Is it robust?

- Experiment by Boynton & Olson proves robustness w.r.t. light source

Robert M. Boynton, *Insights gained from naming the OSA colors*, Color categories in thought and language (Clyde L. Hardin and Luisa Maffi, eds.), Cambridge University Press, 1997, pp. 135–150.

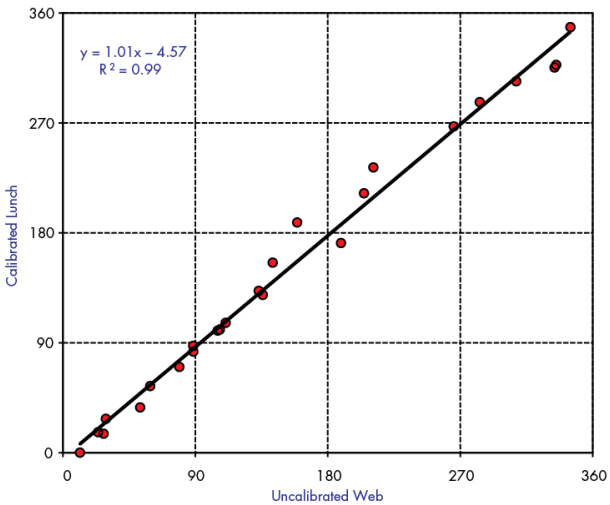


© Copyright 2011 Hewlett-Packard Development Company, L.P.
The information contained herein is subject to change without notice.
Created 26/4/2011

Emissive color?

- Excellent correlation between controlled reflection experiment and uncontrolled crowd-sourced experiment on the Web

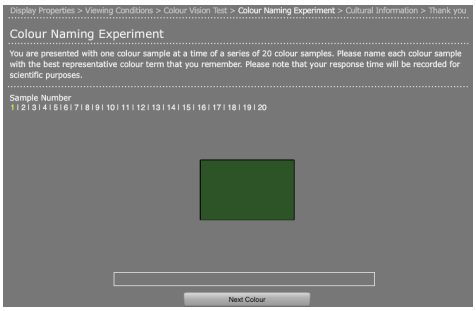
Giordano B. Beretta and Nathan M. Moroney, *Is it turquoise + fuchsia = purple or is it turquoise + fuchsia = blue?* , vol. 7866, SPIE, January 2011, p. 78660H.

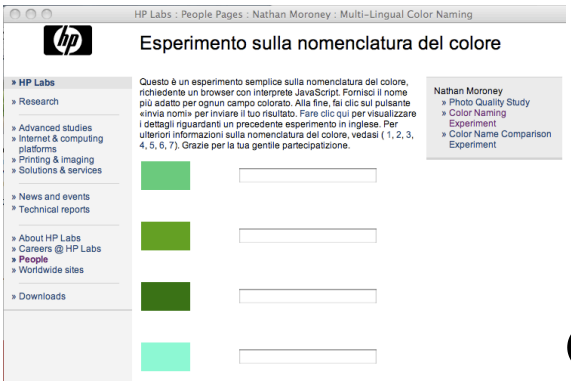


© Copyright 2011 Hewlett-Packard Development Company, L.P.
The information contained herein is subject to change without notice.
Created 26/4/2011

Advantage of Web experiments

- Crowd-sourcing uses the World Wide Web to recruit thousands of respondents
- Persistence in time can account for ephemerality of color terms
- Respondents recruited mainly from the color community
- Nathan Moroney, Dimitris Mylonas



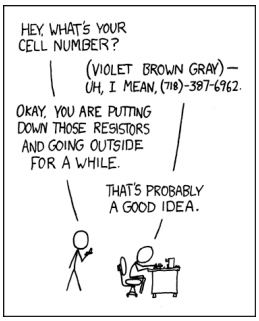


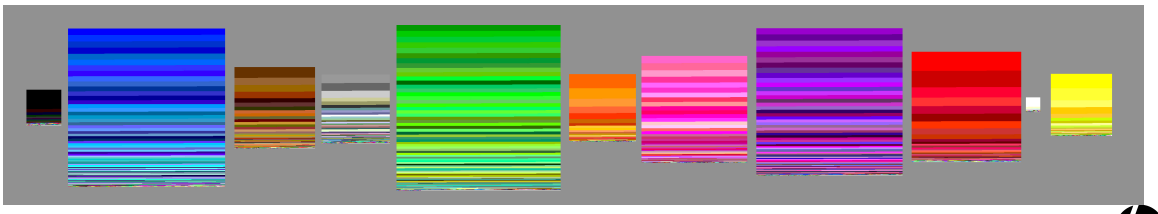
© Copyright 2011 Hewlett-Packard Development Company, L.P.
The information contained herein is subject to change without notice.
Created 26/4/2011

We do not have so many friends

Can we leverage a larger class of respondents?

- Munroe and Ellis of *xkcd* fame have performed a color naming experiment among their readers
- Is there any scientific value in such totally uncontrolled data?
- Frequency sorted color term data points contributed (TDP):





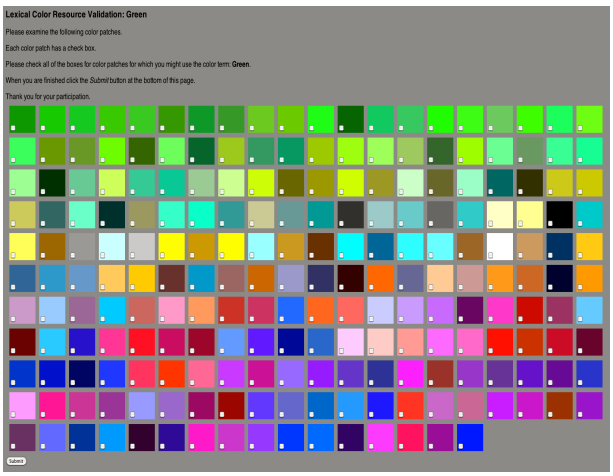
© Copyright 2011 Hewlett-Packard Development Company, L.P.
The information contained herein is subject to change without notice.
Created 26/4/2011

Validation experiment

Can we leverage a larger class of respondents?

- Multi-stimulus categorization task
- Strict sRGB conditions
- 16 color normal observers
- Instructions: "select the color patches you might use with the color term"

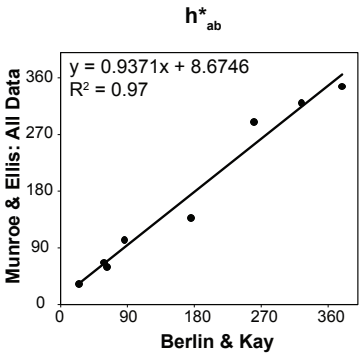
- Contributed color stimuli per term (TDP without frequency), CCS:

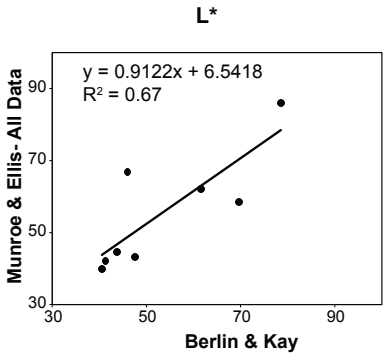


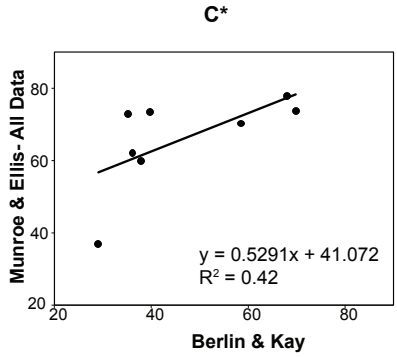
© Copyright 2011 Hewlett-Packard Development Company, L.P.
The information contained herein is subject to change without notice.
Created 26/4/2011

Complete Munroe & Ellis versus the Berlin & Kay data

Graphs show the Berlin and Kay averaged centroids on the abscissa and the Munroe and Ellis averages on the ordinate.





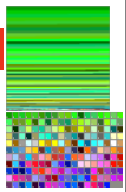


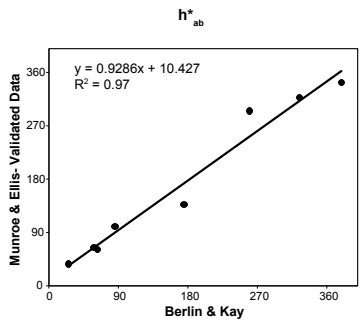
© Copyright 2011 Hewlett-Packard Development Company, L.P.
The information contained herein is subject to change without notice.
Created 26/4/2011

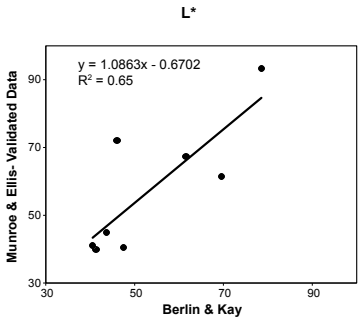
Validated Munroe & Ellis versus the Berlin & Kay

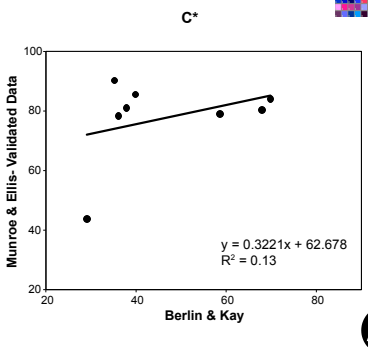
Color data retained after validation:

	Brown	Purple	Pink	Orange	Blue	Yellow	Red	Green
TDP	33%	44%	49%	49%	53%	70%	73%	74%
CCS	2%	7%	6%	3%	7%	3%	5%	18%



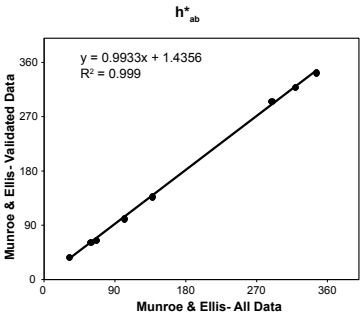


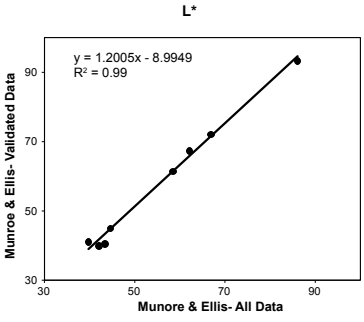


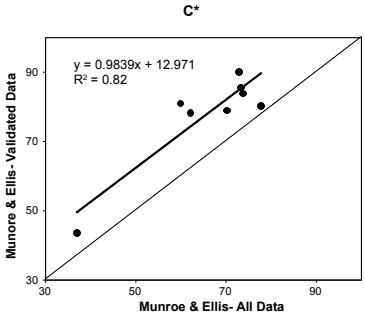


© Copyright 2011 Hewlett-Packard Development Company, L.P.
The information contained herein is subject to change without notice.
Created 26/4/2011

Complete versus validated Munroe and Ellis data







© Copyright 2011 Hewlett-Packard Development Company, L.P.
The information contained herein is subject to change without notice.
Created 26/4/2011

Conclusions

- Color terms are well suited for human communication
- Color naming experiments are robust w.r.t.
  - light sources
  - reflection vs. emissive patches
  - crowd-sourcing
  - disruptive users in large experiments
- Color terms are ephemeral and need continuous experiments
- Future work:
  - basic terms vs. long tail
  - how does color naming scale?

© Copyright 2011 Hewlett-Packard Development Company, L.P.
The information contained herein is subject to change without notice.
Created 26/4/2011