

Statistically identifying basic color terms

Jason Chuang (jcchuang@CS.Stanford.Edu)

Department of Computer Science, 353 Serra Mall
Stanford, CA 94305 USA

Pat Hanrahan (hanrahan@CS.Stanford.Edu)

Department of Computer Science, 353 Serra Mall
Stanford, CA 94305 USA

Abstract

Basic color terms were originally defined by Berlin and Kay based on linguistic definitions and refer to a subset of color expressions that are universal across languages. In this paper, we investigate and report that basic color terms demonstrate much stronger statistical characteristics that differentiate them from other color words. We introduce a probabilistic interpretation of color naming data, and define the notion of term deviation which measures how much individual speakers' usages of a color term deviate from the average population. Our analysis of eight languages reveals strong correlation and evidence that basic color terms can be quantitatively characterized as terms whose usages exhibit the least amount of deviation.

Keywords: Basic Color Terms; World Color Survey.

Introduction

Berlin and Kay proposed the notation of basic color terms in 1969 and put forward the hypothesis of the universality of basic color terms across languages. A color word is said to be a basic color term if it satisfies eight linguistic criteria.¹ Subsequent research has revealed interesting results such as the non-random distribution of basic color terms (Kay & Regier, 2003), optimal partition of the color space by basic color terms (Regier, Kay, & Khetarpal, 2007) and color name-based applications in the fields of computer graphics and color imaging (Chang, Saito, & Nakajima, 2003; Motomura, 2002).

In this paper, we investigate whether meaningful statistical properties can be derived about basic color terms. Do basic color terms, collectively, exhibit any statistical characteristics that differentiate them from other color words? Is it possible to quantify such characteristics and identify basic color terms by examining a color term's usage patterns by individual speakers of a language? We answer affirmatively to both questions, and demonstrate that while basic color terms are initially defined by a set of linguistic criteria, they, in fact, infer strong statistical characteristics about the usage of the color terms.

Data and Methodology

Our analysis is based on the World Color Survey (WCS) which contains a rich set of naming data for 2300 color terms in 110 languages. Due to the setup of the survey, color terms recorded in the WCS can consist of both basic and non-basic

¹Eleven basic color terms were identified by Berlin and Kay, and correspond to white, gray, black, red, yellow, green, blue, pink, orange, brown, and purple in English.

color terms. Efforts are currently underway to analyze the WCS including classifying basic color terms in each of the 110 languages (Cook, Kay, & Regier, 2005).

A preliminary analysis of the WCS on eight languages was published by Kay, Berlin, Maffi, and Merrifield (1997). We use the basic color term classifications reported in this preliminary analysis to assess our proposed statistical measure.

We introduce a probabilistic interpretation of the color naming data, and propose the notion of *term deviation* which measures how much naming responses from individual speakers deviate from the average response given by the population.

The computation of term deviation involves three steps which are described in details in the subsections below. We first construct the *population average response* which is a multinomial probability distribution representing the aggregated response given to a color term by all speakers of a language. We then compute *speaker deviation* which measures how much a single speaker's usage of a color term deviates from the population average response. Finally, term deviation is defined as speaker deviation averaged over all speakers who use the color term in their vocabulary.

World Color Survey

The World Color Survey is a large-scale field study initiated in 1976 to collect cross-linguistic color naming data from 110 unwritten, geographically-distributed languages representing a wide range of language families.

The survey was conducted *in situ* and collected naming responses that were constrained but not restricted to a predetermined set of vocabulary. Field workers interviewed the subjects at the locations where the subjects reside. Naming information was recorded using a stimulus array consisting of 330 color chips as shown in Figure 1. The color chips were shown one at a time in random order to the speakers who were asked to name each chip. The speakers were instructed to use words which they themselves consider as basic color terms, but their responses were *not* otherwise restricted to a predetermined set of vocabulary. A total of 2300 terms were recorded from the 2616 subjects from the 110 languages.

As no explicit restrictions were placed on the responses, some of the words recorded may not necessarily qualify as basic color terms. In the preliminary analysis of the World Color Survey by Kay et al., the authors identified basic color terms for eight languages based on the linguistic criteria de-

Table 1: Basic color terms for eight languages were reported in the preliminary analysis of World Color Survey by Kay et al. (1997). The authors classified terms that qualify as basic color terms based on the linguistic definitions proposed by Berlin and Kay (1969). Of the 205 color terms recorded in the eight languages, 43 were identified as basic color terms. We use this list of words to assess our proposed statistical measure.

Language	Number of color words	Number of basic color terms	Basic color terms identified
Buglere	20	6	lere, jutre, moloin, dabe, jere, leren, kwajusa
Candoshi	18	7	borshi, kavabana, ptsiyaro, kantsirpi, chobiapi, kamachpa, tarika, pozani
Kalam	66	6	tund, likan, walin, muk, minj-kimemb, mosimb
Konkomba	20	5	maman, pipin, bombon, diyun, yaankal
Kwerba	21	5	nokonum, icem, asiram, kainanesenum
Martu Wangka	44	5	yukuri-yukuri, miji-miji, maru-maru, karntawarra
Múra Pirahã	4	4	bi i'sai, ko biai, a hoa saa ga, bio pai ai
Sirionò	12	5	eirei, eshi, erondei, eruba, echo

finied by Berlin and Kay. Of the 205 color words recorded in the eight languages, only 43 were classified as basic color terms. Their results are summarized in the Table 1.

Term Maps and Population Average Response

In the same preliminary analysis on the WCS by Kay et al., the authors proposed the use of term maps to visualize the aggregated responses given to a particular color term. The authors constructed a term map by tallying the number of responses a term received for each color chip, and overlaid the tallies on top of the stimulus array. Figure 2 shows a reproduction of a term map where the areas of the chips are scaled in proportion to the number of responses received. Mathematically, we define the notation $f(\text{Color} = c, \text{Term} = t)$ as the number of responses a term t received for a color chip c .

We interpret term maps as a probability, and point out that term maps are strongly related to the notion of multinomial probability distributions. As color naming responses vary from speaker to speaker, a probabilistic interpretation provides us with a framework to model and quantify the amount of uncertainty in the color naming. A normalized term map is, in fact, equivalent to the multinomial distribu-

tion $P(\text{Color} = c | \text{Term} = t)$ which relates the use of a term to the colors referred to by the term. More specifically, $P(c|t)$ can be computed directly from a term map by normalizing the response frequencies:

$$P(c|t) = \frac{f(c,t)}{\sum_c f(c,t)}$$

Under this interpretation, a term map can be thought of as the *population average response* for color term t . When term t is used, the area of a color chip c corresponds to the *expected likelihood* that c is the color that elicited the response.

Speaker Deviations and Response Maps

We propose the notation of *speaker deviation* which measures how much a single speaker's response deviates from the population average. To compute speaker deviation, we first introduce the use of *response maps* to visualize the usage of a color term by an individual speaker. Speaker deviation then quantifies the difference between a response map and a term map.

Similar to term maps, we visualize the responses from an individual speaker by overlaying the color chips that the

Figure 1: The 330 color chips used in the World Color Survey (WCS). Naming information on 2300 color terms given by 2616 speakers in 110 languages was recorded using the stimulus array.

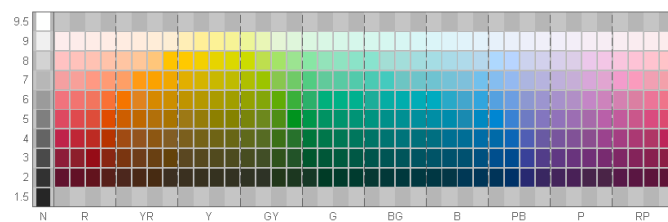
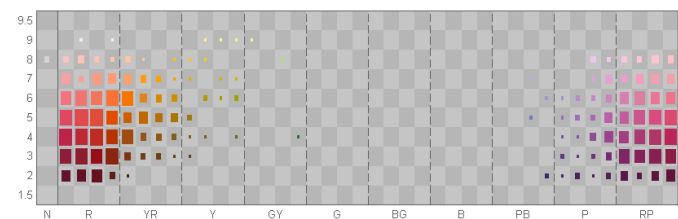


Figure 2: Example term map for 'dabe' in the language 'Buglere'. The areas of the color chips are drawn in proportion to the number of responses, and can be interpreted as the *population average response* of the likely set of colors that a color term refers to.



speaker associates with a term on top of the stimulus array. Figure 3 shows the response maps for 25 speakers (for the term ‘dabe’ in the language Bulgere, the term map for which is displayed earlier in Figure 2). We define the notation $f_i(\text{Color} = c, \text{Term} = t)$ which indicates:

$$f_i(c, t) = \begin{cases} 1, & \text{if speaker } i \text{ refers to color } c \text{ using term } t; \\ 0, & \text{otherwise.} \end{cases}$$

A normalized response map can also be considered as a multinomial distribution $Q_i(\text{Color}=c|\text{Term}=t)$ which describes the *observed* set of colors c that elicited the use of term t from speaker i .

$$Q_i(c|t) = \frac{f_i(c, t)}{\sum_c f_i(c, t)}$$

The statistical measure, Kullback-Leibler (KL) divergence estimates how much an observed probability distribution differs from the expected probability distribution (Kullback & Leibler, 1951). Treating $P(c|t)$ as the expected term response from the average population and treating $Q_i(c|t)$ as the actual observed term usage by an individual speaker i , we define speaker deviation as the KL-divergence of the speaker response from the population average response.

$$\text{Speaker Deviation}(i) = KL(Q_i||P) = \sum_c Q_i(c|t) \log \frac{Q_i(c|t)}{P(c|t)}$$

Term Deviation

Finally, we define *term deviation* for a color term t as the speaker deviation averaged over all speakers who used the term in their vocabulary.

$$\text{Term Deviation} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \text{Speaker Deviation}(i)$$

where \mathcal{S} is the set of speakers who used the color term t in their response.

Results

Our goal is to look for statistical properties that characterize basic color terms as a whole across languages, and investigate whether it is possible to derive quantitative measures that differentiate basic from non-basic color terms.

We selected 205 color terms from the eight languages analyzed by Kay et al. (1997) for which known classifications of basic color term are available. We computed term deviations for all terms except those used by 5 or few speakers as it is unclear whether such a small sample size produces term maps that reliably reflect the population average response.

Table 2 shows the results of our analysis. Each table contains the list of color terms for one language. Color terms are ordered by increasing term deviation, with the terms at the top exhibiting the least amount of deviation and at the bottom, the greatest deviation. Terms classified as basic color terms are marked with a ‘+’ sign in the BCT column and a ‘-’ sign otherwise.

We observed a complete correspondence between basic color terms and terms that exhibit the least amount of deviation. In each of the eight languages, terms classified as basic color terms always appear at the top of the list, and non-basic color terms always appear at the bottom of the list. The pattern holds for all terms in all eight languages, without a single exception.

By ordering color terms along the dimension of term deviation, basic color terms can be identified and differentiated from non-basic color term by examining whether a term exhibits deviation below a specific threshold, which varies slightly from language to language.

Conclusion

In this paper, we proposed the measure *term deviation* that determines how much individual speakers’ usages of a color term differ from average response given by the population. Our goal is to examine whether meaningful properties about basic color terms can be derived from such a statistical measure.

We report that basic color terms correspond to the set of terms that exhibit the least amount of term deviation, based on comparison with known classification results by Kay et al. The pattern is observed in all eight languages for which we have data, and holds true for every term in each of the eight languages without any exception.

Basic color terms were first defined by Berlin and Kay based on the linguistic definitions associated with the terms. Our observation indicates that basic color terms, in fact, exhibit even stronger statistical properties, and can be quantified as terms whose usages by individual speakers deviate the least from population average.

This characteristic allows us to identify basic color terms from non-basic color terms based by examining a term’s usage. By ordering color terms by deviation, basic color terms can be separated from non-basic ones if they exhibit term deviation below a sufficiently low threshold.

The threshold that defines basic color terms, however, varies slightly from language to language and invites further investigation. Some words, such as ‘eaqui’ and ‘turuma-uti’ in the language ‘Sirionò’ exhibit less deviation than ‘diyun’ and ‘yaankal’ in the language ‘Konkomba’ but are not classified as basic color terms.

While the separation of basic and non-basic terms corresponds to a large gap in term deviation in some languages such as ‘Buglere’ and ‘Candoshi’, the gap is considerably smaller in other languages such as ‘Martu Wangka’, potentially indicating that the difference between basic and non-basic color terms may be less distinctive in certain languages.

We conclude with the hope that future statistical and quantitative analysis may continue to contribute to our understanding of color naming and reveal additional insights about basic color terms and about the languages themselves.

Figure 3: Examples of response maps, speaker deviations, and term deviation. We use response maps, as shown below, to visualize individual speakers' usages of a color term. *Speaker deviation* is defined as how much the observed term usage by an individual speaker deviates from the expected population average response. *Term deviation* is then defined as the aggregated speaker deviation average over all speakers who use the term in their vocabulary. The examples below are for the term 'dabe' given by 25 speakers of the language 'Bulgere'. The response maps are ordered in increasing speaker deviation. The response map given by speaker 9 at the top of the left column exhibits the least deviation from population average response (shown earlier in Figure 2). The response given by speaker 14 at the bottom of the right column exhibits the most deviation. Term deviation, averaged over 25 speakers, is 0.883541 bits.



Table 2: Comparison of term deviation and basic color term classification in eight languages. Each table below lists the color terms in one language and their corresponding term deviation, ordered by increasing deviation. Terms classified as a basic color term by Kay et al. (1997) are denoted with a '+' sign in the BCT column and a '-' sign otherwise. We observed a complete correspondence between terms that exhibited the least amount of deviation and terms that were classified as basic color terms. Basic color terms were separated from non-basic color terms by a threshold in term deviation in each of the eight languages.

Language: Buglere

Term	BCT	Deviation
lere/lerere/lejre	+	0.574989
jutre/jusa	+	0.685247
moloin/moloinre	+	0.851454
dabe/dabere	+	0.883541
jere/jerere	+	1.07413
leren	+	1.18094
kwajusa	-	2.20122
...
(13 more terms)	-	< 5 speakers

Language: Candoshi

Term	BCT	Deviation
borshi	+	0.343470
kavabana	+	0.414831
ptsiyaro	+	0.451685
kantsirpi	+	0.461904
chobiapi	+	0.527526
kamachpa	+	0.893788
tarika	+	1.44197
ponzai	-	2.62913
...
(10 more terms)	-	< 5 speakers

Language: Kalam

Term	BCT	Deviation
tund	+	0.529957
likan	+	0.657256
walin	+	0.723300
muk	+	1.05129
minj-kimemb	+	1.25444
mosimb	+	1.28497
anjerj-ay	-	1.69209
spay	-	2.11845
wilik	-	2.18710
sambiy-arjgin	-	2.24722
maym	-	2.65602
...
(55 more terms)	-	< 5 speakers

Language: Konkomba

Term	BCT	Deviation
maman	+	0.457822
pipin	+	0.524728
bombon	+	0.584065
diyun	+	1.37567
yaankal	+	1.39744
siin	-	1.98829
kunii	-	3.36576
...
(13 more terms)	-	< 5 speakers

Language: Kwerba

Term	BCT	Deviation
nokonum	+	0.395546
icem	+	0.676523
asiram/aherem	+	0.689575
kainanesenum	+	1.36948
masibucinom	-	1.65985
kacenum	-	2.09075
...
(15 more terms)	-	< 5 speakers

Language: Múra Pirahã

Term	BCT	Deviation
bi i'sai	+	0.262299
ko bai	+	0.403820
a hoa saa ga	+	0.448935
bio pai ai	+	0.589687

Language: Martu Wangka

Term	BCT	Deviation
yukuri-yukuri	+	0.602560
miji-miji	+	0.729724
maru-maru	+	0.784300
karntawarra	+	1.22315
piila-piila	+	1.44980
ngarnka	-	1.51277
parnaly-parnaly	-	1.60269
martaly-martaly	-	2.11928
munga-turrkatingu	-	2.32329
...
(35 more terms)	-	< 5 speakers

Language: Sirionò

Term	BCT	Deviation
eirei	+	0.439152
eshi	+	0.573232
erondei	+	0.853320
eruba	+	0.879355
echo	+	0.951372
eaqui	-	1.32206
turuma-uti	-	1.32360
enumbi	-	2.46146
...
(4 more terms)	-	< 5 speakers

References

- Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley and Los Angeles: University of California Press.
- Chang, Y., Saito, S., & Nakajima, M. (2003). A framework for transfer colors based on the basic color categories. In *Computer graphics international* (pp. 176–181).
- Cook, R. S., Kay, P., & Regier, T. (2005). The world color survey database: History and use. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorisation in the cognitive sciences*. Elsevier.
- Kay, P., Berlin, B., Maffi, L., & Merrifield, W. (1997). Color naming across languages. In C. L. Hardin & L. Maffi (Eds.), *Color categories in thought and language*. Cambridge.
- Kay, P., & Regier, T. (2003). Resolving the question of color naming universals. *Proceedings of the National Academy of Sciences*, 100(15), 9085-9089.
- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1), 79-86.
- Motomura, H. (2002). Analysis of gamut mapping algorithms from the viewpoint of color name matching. *Journal of the Society for Information Display*, 10(3), 247–254.
- Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, 104(4), 1436-1441.