Cognitive Science 31 (2007) 99–132


# Explaining Color Term Typology With an Evolutionary Model

Mike Dowman

*Department of General Systems Sciences, Graduate School of Arts and Sciences, University of Tokyo*

Received 13 June 2005; received in revised form 3 March 2006; accepted 1 May 2006

**Abstract**

An expression-induction model was used to simulate the evolution of basic color terms to test Berlin and Kay's (1969) hypothesis that the typological patterns observed in basic color term systems are produced by a process of cultural evolution under the influence of biases resulting from the special properties of universal focal colors. Ten agents were simulated, each of which could learn color term denotations by generalizing from examples using Bayesian inference, and for which universal focal red, yellow, green, and blue were especially salient, but unevenly spaced in the perceptual color space. Conversations between these agents, in which agents would learn from one another, were simulated over several generations, and the languages emerging at the end of each simulation were investigated. The proportion of color terms of each type correlated closely with the equivalent frequencies found in the World Color Survey, and most of the emergent languages could be placed on one of the evolutionary trajectories proposed by Kay and Maffi (1999). The simulation therefore demonstrates how typological patterns can emerge as a result of learning biases acting over a period of time.

*Keywords:* Basic color terms; Typology; Iterated learning model; Cultural evolution; Bayesian inference

## 1. Introduction

Berlin and Kay (1969) discovered typological patterns in the way in which languages name color. They showed that the color term systems of most languages correspond to a small set of possible systems, and that most logically possible types of color term systems are unattested. They suggested that these results were the product of a cultural evolutionary process through which color term systems develop over time. Subsequent studies have led to modifications of Berlin and Kay's typological findings, but have largely confirmed the finding that languages conform to a highly constrained range of possible color term systems (Kay, 1975; Kay, Berlin,

Correspondence should be addressed to Mike Dowman, Department of General Social Sciences, The University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan. E-mail: mike@sacral.c.u-tokyo.ac.jp

Maffi, & Merrifield, 1997; Kay, Berlin, & Merrifield, 1991; Kay & Maffi, 1999; Kay & Regier, 2003; MacLaury, 1997a; Regier, Kay, & Cook, 2005a, 2005b). Suggestions have been made that color term typology might be explainable in terms of universal aspects of human neurophysiology (especially properties of the human color vision system), environmental factors, or cultural practices. This article reports experiments using a computer model of color term evolution that shows that many of the typological patterns seen in the empirical data are explainable simply in terms of the shape of the perceptual color space, and the increased salience and uneven spacing of universal focal colors.

Berlin and Kay's (1969) typological patterns concerned a special subset of color terms, named *basic color terms*. These are the most salient and commonly used color words, and in English they are *red, orange, yellow, green, blue, purple, pink, gray, brown, black,* and *white*. Berlin and Kay noted that basic color terms have prototype properties (Rosch, 1978; Taylor, 1989); that is, there is usually a single color, the *prototype,* that is the best example of the color term, and colors become less good examples of the color term the more dissimilar they are to that color. Berlin and Kay were able to show that there was a remarkable degree of consistency in where the prototypes of basic color terms occur, both among speakers of the same language and across languages. They also found that almost all languages seem to have between 2 and 11 basic color terms, a finding that has been generally confirmed by later studies, although a few languages, such as Russian and Hungarian, appear to have 12 basic color terms[1] (MacLaury, 1997a). Berlin and Kay (1969) found clear typological regularities concerning which color terms occurred together, and they described these regularities using the implicational hierarchy shown in Fig. 1. They proposed that if a language had a color term with a prototype at any point in the hierarchy, then it would also have color terms with prototypes at all the colors to the left of that color. It is important to note, however, that 6 languages out of the 98 in Berlin and Kay's study appeared to be exceptional, and did not conform to the hierarchy.

Since Berlin and Kay's (1969) original study, a large amount of data concerning the basic color terms of a very wide range of languages from throughout the world have been collected. This has led to several revisions of Berlin and Kay's theory concerning possible color term systems (Kay, 1975; Kay et al., 1997; Kay et al., 1991; Kay & Maffi, 1999; MacLaury, 1997a). Whereas Berlin and Kay (1969) based their theory on the locations of the prototypes of color terms, more recent work has looked at the full range of hues that color terms denote. Languages with only two basic color terms, although extremely rare, either divide up the color space into two parts, one containing all light colors and the other all dark colors (MacLaury, 1997a), or, as in languages such as Dugum Dani (West Papua, Indonesia), one term denotes all light colors together with red and yellow, and the other denotes all dark colors, together with green and blue (Heider & Olivier, 1972). Three-way composite color terms of this latter type can have their prototypes at black, white, red, yellow, green, or blue (Kay et al., 1997). The

$$\begin{bmatrix}\text{white}\\ \text{black}\end{bmatrix} < \begin{bmatrix}\text{red}\end{bmatrix} < \begin{bmatrix}\text{green}\\ \text{yellow}\end{bmatrix} < \begin{bmatrix}\text{blue}\end{bmatrix} < \begin{bmatrix}\text{brown}\end{bmatrix} < \begin{bmatrix}\text{purple}\\ \text{pink}\\ \text{orange}\\ \text{grey}\end{bmatrix}$$

Fig. 1. Berlin and Kay's (1969) implicational hierarchy.

general tendency for yellow and red to be grouped with each other, and sometimes also with white, has been confirmed in later studies, as has the tendency for blue and green to be grouped with each other and with black.

Kay and Maffi (1999) presented a new theory concerning color term evolution that attempted to incorporate as many of the findings as possible of the World Color Survey, which collected data from speakers of 110 unwritten languages (Kay et al., 1997). Their theory concerned only those basic color terms that contained black, white, or one of the four universal focal colors (the red, yellow, green, and blue colors that form the prototypes of many basic color terms). *Derived* color terms, such as *purple* or *gray,* were not considered. Kay and Maffi (1999) proposed that all languages have evolved from a state in which they have just two basic color terms. The two color term systems normally contain one white-red-yellow term, and one black-green-blue one. As the language adds more color terms, these colors will gradually be separated, until each of what Kay et al. (1997) termed the six fundamental colors (i.e., black, white, red, yellow, green, and blue) are all represented by separate color terms. Fig. 2 shows the limited set of trajectories on which almost all the languages in the World Color Survey can be placed. Kay and Maffi (1999) were able to place 83% of the languages somewhere on the main

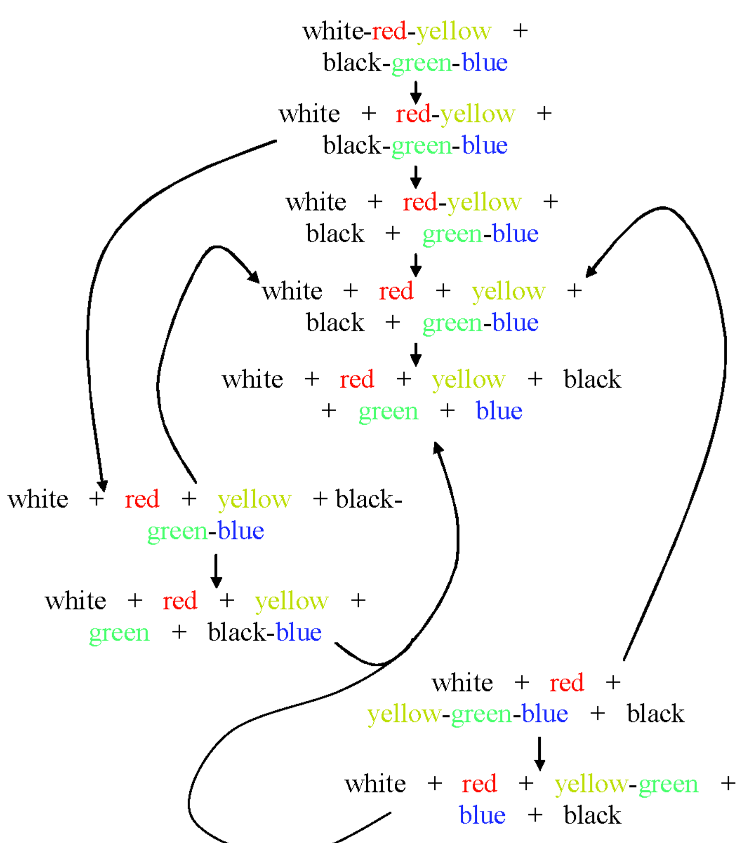



Fig. 2. The evolutionary trajectories proposed by Kay and Maffi (1999).

line of the trajectory, which starts with a two color term system at the top of Fig. 2, and proceeds straight downward, until a six color term system is reached, in the center of the diagram.

Seven languages appeared to be following the side branch appearing below and to the left of the main trajectory, as they contained separate red and yellow terms and a black-green-blue, or black-blue composite term. A further three languages contained yellow-green-blue terms, or yellow-green terms, and therefore appeared to be evolving along the trajectory shown in the lower right of Fig. 2. The arrows show possible routes between languages, and indicate that languages evolving along either of these trajectories can join the main trajectory. However, the origin of color term systems containing yellow-green-blue composites is unclear, because they could not be formed by subdividing color terms in any attested two color term system. Kay and Maffi (1999) and MacLaury (1997a) both proposed quite different explanations for the origin such color term systems, suggesting that they evolve either from languages that have no color terms for parts of the color space (Kay & Maffi, 1999), or from languages that divide the color space based simply on a light–dark split (MacLaury, 1997a). However, there is no clear evidence supporting one of these proposals over the other. Another possible explanation, and one that is compatible with the model of color term evolution proposed here, is that languages do not necessarily always evolve by subdividing color terms. If this is the case, then such composite terms could arise simply as the result of random drift in the meanings of color terms, and no specific explanation for their emergence would be required.

It should be noted that some languages appeared to be in transition between stages in the evolutionary trajectories, because some speakers of the language (usually younger speakers) used more basic color terms than others. Therefore, the language could appear to be at different stages, depending on which informant's color term system was considered. Kay and Maffi (1999) found that 25 of the 110 languages in the World Color Survey were in transition. In addition, six languages were exceptional, and could not be placed on the evolutionary trajectories. In some cases this was because there was no basic color term that consistently named some parts of the color space. For example, Kuku-Yalanji (Australia) has no consistent term for green. Some speakers name either just green or both green and blue with *kayal,* but most speakers of the language did not use this term at all. Most speakers who did not use *kayal* did however use *burrkul* or *burkul* to name yellow-green-blue, but this term was not used by those speakers with well-established green or blue terms. Such a degree of inconsistency between speakers of the same language might seem surprising, but MacLaury (1997a) reported that it is not uncommon.

Kay and Maffi (1999) found that two languages did not exhibit such inconsistencies between speakers, but still did not fit onto the evolutionary trajectories. Waorani (Ecuador) has a yellow-white term that does not include red, but no such term appears in any part of the evolutionary trajectories, and Gunu (Cameroon) contains a separate blue term as well as a black-green-blue composite. Such data should probably lead us to the conclusion that the evolution of basic color terms is a partly predictable process, but that some languages are exceptional because they diverge from the major attested types. To dispel any doubt that universal patterns do exist, Kay and Regier (2003) made a statistical analysis of the World Color Survey data, in terms of the locations in the color space of the centroids of each color term, and they showed that the tendency for the centers of color term denotations to be clustered in certain parts of the color space was highly statistically significant, thus confirming the more impressionistic results of earlier studies.

Kay and Maffi (1999) did not include the derived basic color terms in their trajectories, because the order of appearance of these terms appears to be less predictable (Kay et al., 1997). Kay (1975) noted that the appearance of gray terms is largely unpredictable, although the more other basic color terms a language has, the more likely it is that it will also have a gray term. Kay et al. (1991) reported that whereas purple and brown may be seen in languages in which there is a green-blue composite, orange and pink will not normally be seen unless a language also has separate green and blue terms. However, MacLaury (1997a) reported that orange occasionally appears in languages that do not have a purple term, but this is not usually the case. Therefore the empirical data show that purple terms are considerably more common than orange ones.

Although the World Color Survey reveals a complex pattern of cross-linguistic variation in color term naming, it should be acknowledged that there are many properties of basic color terms that are not apparent if only the range of colors that they denote is considered. Saunders (1992) noted that many languages lack words that simply denote color. Words that are typically classified as basic color terms may have cultural or religious significance, and Saunders argued that they therefore cannot be understood in isolation from other words in the language, or from the belief system of which they form a part. Levinson (2000) noted that words that denote color do not always form a coherent set that is distinct from words denoting other kinds of property, and that color is often conflated with properties such as texture or variegation. He also suggested that whole expressions sometimes fulfill Berlin and Kay's (1969) criteria for distinguishing basic color terms from other words better than individual words do. However, these findings do not seem to be incompatible with those of Kay and Maffi (1999). If we focus simply on words' meanings with respect to color, then typological patterns become apparent. Clearly this will neglect many other aspects of color term meanings, but that in no way invalidates the typological findings.

## 2. Explanations of color term universals

Most previous attempts to explain color term typology have made reference to the special status of the four universal focal colors that, as noted earlier, are very frequently chosen as the best examples of basic color terms. Ratliff (1976) proposed a psychophysiological basis for color term universals, suggesting that the order of emergence of black, white, red, yellow, green, and blue terms was determined by the relative strengths of the universal foci and the greater discriminability of red over blue. Ratliff's theory has many similarities to the model presented here, but it did not explain why we see only some types of derived color terms and not others that are logically possible, and it left unspecified some intermediate steps necessary to explain how psychophysical effects come to affect color vocabulary. MacLaury (1997a) proposed that color term typology is explainable in terms of the varying perceptual distances between the universal foci, based on the hypothesis that people form color categories by making analogies with space and motion. However, it is not clear that his theory really distinguishes correctly between which kinds of color term system can and cannot evolve. Regier et al. (2005b) showed that if the foci of the color terms in a language are known, it is possible to predict the approximate denotational range of each term. However, Regier et al.'s model did not explain which universal foci are used in any particular language, so it did not fully explain the typological patterns.

Kay and McDaniel (1978) suggested that the universal nature of color categories could be due to them being derived directly from the neural responses of cells involved in color perception (*opponent cells*). Composite categories were derived using fuzzy set unions (Zadeh, 1965) of the response functions of more than one type of opponent cell, and derived terms using fuzzy set intersections. However, Kay and McDaniel's (1978) theory appears to be too powerful, in that it predicts the presence of many types of basic color terms that do not exist (e.g., yellow-green derived terms, or red-blue composites). It also appears to be too inflexible, because it does not allow for variation in the shape or size of color terms, even though this is attested empirically. Furthermore, the neurophysiological basis for the theory is now regarded as being unsubstantiated (see section 3).

Yendrikhovskij (2001) investigated whether typological patterns could be explained in terms of the shape of the perceptual color space, and the distribution of colors in the environment. He sampled 10,000 pixels from 630 images of the natural environment, plotted their positions in color space, and grouped them into clusters based on their similarity. The cluster locations tended to mirror the centroids of color categories in human languages with the same number of terms when 3, 7, or 11 clusters were made. (Yendrikhovskij did not give details in other cases.) This suggests that typological patterns can be explained in terms of a tendency to develop color terms centered on colors that are both frequently encountered, and that are perceptually distinct. We should note however, that Yendrikhovskij did not investigate whether the same results would have been produced if colors had simply been chosen at random, and so his results did not demonstrate that consideration of the uneven distribution of colors in the environment was necessary to his explanation of typological patterns. Furthermore, Steels and Belpaeme (2005) failed to replicate these results in their application of the same techniques, so they may have been, at least in part, the result of chance correlations. Belpaeme and Bleys (2005) reported a computer model that suggests that color universals are best explained in terms of the properties of the perceptual color space and the process of linguistic transmission, rather than environmental factors (see section 5).

Finally, Foley (1997) and Saunders (1992) suggested that consideration of cultural practices is necessary to gain a full understanding of color terms. Foley (1997) suggested that "culture must be the crucial autonomous intermediary between any innate and hence universal neurological perception of color stimuli and the cognitive understanding of these" (p. 160). Although there seems to be a connection between the universal foci and the typological patterns seen in color term systems, none of the approaches mentioned earlier gives a really satisfactory account of how the foci give rise to the typological patterns. The model described in this article makes a causal link between special properties of the universal foci and color typology, incorporating some of the proposals of Ratliff (1976), Kay and McDaniel (1978), and MacLaury (1997a). It explains color term typology as the product of cultural evolution under the influence of universal, and hence presumably innate[2], biases.

## 3. Unique hues and universal foci

Although the universal foci stand out clearly from the other colors in the World Color Survey data, a variety of other evidence also highlights their special status. The universal foci vary both in hue and in lightness (focal yellow in particular is much lighter than any of the other universal foci), but there is a considerable amount of nonlinguistic evidence concerning the hues of these universal foci (the *unique hues*). Hering (1964) noted that red and green, and yellow and blue, appear to be opposite colors perceptually, as no color can be seen as containing elements of both members of either of these opponent pairs. Also red and green (or yellow and blue) lights appear gray when shone on the same spot (K. K. De Valois & De Valois, 2001).

The special status of the universal foci has also been established through psychological experiments. Rosch[3] conducted experiments with young U.S. children, and with speakers of Dugum Dani, which, as noted earlier, has only two basic color terms[4] (Heider, 1971, 1972; Heider & Olivier, 1972; Rosch, 1973). Dugum Dani speakers were used in an attempt to avoid any interference from English color categories. These studies appeared to show that the universal foci were better remembered and more salient than the other colors, both for U.S. children and for speakers of Dugum Dani. However, some of these results have since been called into question (Lucy & Shweder, 1979; Roberson, Davies, & Davidoff, 2000). In many of the experiments, participants were shown a target color chip, and then asked to pick out the matching chip in an array of Munsell[5] color chips. In several such experiments, participants more often selected the correct chip when the target was focal than when it was not. However, Roberson et al. (2000), who conducted their experiments with speakers of British English and of Berinmo[6] (Papua New Guinea), found that in some experiments participants tended to select focal colors from the array, regardless of whether the target chip was focal or not, which would in itself account for better performance when matching focal chips. When they compensated for this effect, they found no benefit for focal colors. Lucy and Shweder (1979) also found that the focal colors were perceptually farther from their neighbors than the other colors, making some tasks easier when foci were the target. This would have given the impression that the focal colors had a special psychological status, but when it was compensated for, most speakers performed no better for focal than for nonfocal colors.

However, even if we discount all of the experiments that relied on the Munsell chips in the array being evenly spaced perceptually, Rosch's work still provides clear evidence of the special status of the focal colors. First, when asked to choose color chips in free choice experiments, U.S. children picked focal color chips more often than nonfocal ones (Heider, 1971). Dugum Dani speakers also found it easier to learn to associate words with particular individual color chips when those color chips were focal ones as opposed to nonfocal ones (Heider, 1972). In a replication of this task, Roberson et al. (2000) found an advantage only for focal red, although this could be explained by the presence of a red term in Berinmo, and so might not be due to any prelinguistic bias.

There have also been suggestions that the special status of the universal foci has a neurophysiological basis. R. L. De Valois, Abramov, and Jacobs (1966) found opponent cells in the lateral geniculate nuclei of Macaque monkeys' brains that responded maximally in the presence of particular hues of red, yellow, green, or blue light, and minimally in the presence of the opposite color. These were the cells on which Kay and McDaniel (1978) based their explanation of color term universals (see section 2), on the assumption that the hues triggering the maximal firing rates of these cells correspond to the unique hues, and hence that the universal foci have a neurophysiological basis. However, it has become clear that the outputs of these cells cannot correspond directly to phenomenal color (Abramov, 1997; Derrington, Krauskopf, & Lennie, 1984; Jameson & D'Andrade, 1997). First, the null point of the cells opposing red and green is at a greenish-yellow, not at a color that is neither green nor red, as we would predict from phenomenal evidence. Second, the cells responding to long-wavelength light (which is perceptually red) fail to respond to low-wavelength (perceptually violet) light that, although being at the opposite end of the spectrum, is also perceived as reddish. Opponent cells also respond to spatial information, and so do not simply signal information about color.

Webster, Miyahara, Malkoc, and Raker (2000) also demonstrated that intersubject variation in sensitivity to each of the opponent channels does not correlate with the considerable intersubject variation as to which hues are perceived to be unique, which is what would be expected if the responses of opponent cells directly corresponded to perceived hue. R. L. De Valois and De Valois (1993) proposed a third stage of color processing to reconcile the discrepancy between opponent cell responses and phenomenal unique hues, but at present any neurological basis for the unique hues and universal foci must remain hypothetical.

Overall, we can see that, although the neurophysiological evidence for the universal foci is very doubtful, and many of the early results of Rosch have been called into doubt, the universal foci clearly display special properties in some psychological experiments. In particular, children tend to choose prototype colors in free choice experiments, suggesting that they are especially salient. Psychophysical results also consistently demonstrate that universal focal red, yellow, green, and blue are special, in that they are not seen as combinations of other hues. However, perhaps the clearest evidence for the special status of the universal foci is provided by the tendency for the prototypes of color terms to be clustered on the universal foci. This is supported by a wide range of cross-linguistic evidence, especially the analyses of the World Color Survey data performed by MacLaury (1997b) and Regier et al. (2005a, 2005b).

## 4. Perceptual color spaces

Crucial to the explanation of color term typology proposed in this article is not just the special status of the universal foci, but also their locations in perceptual color space. If two universal foci are similar, and hence close together in the color space, we might expect them to be frequently named by a composite term (because even a fairly small color category would be able to include both foci). However, we would not expect to see many derived terms between these foci, because any such term would have to fit into the small perceptual space between them. A similar, but complementary argument could be made that if two neighboring universal foci are far apart, we would expect to frequently see a derived term in between, but to rarely, if ever, see a composite term including both foci. Unfortunately, the perceived similarity between colors bears only a very indirect relation to the physical properties of light, or to anything that is directly measurable, making determination of perceptual distances between specific colors extremely difficult.

There have been many attempts to construct color spaces in which the distance between each pair of colors is proportional to how dissimilar they are. Some of the best known examples of such spaces are the Munsell (Committee on Colorimetry, Optical Society of America, 1953), the CIE L*a*b* (Kuehni, 1997; Robertson, 1977) and the Optical Society of America (MacAdam, 1974) color spaces. In all of these spaces, an attempt has been made to make the distances between colors reflect their perceptual similarity. The version of the Munsell system in use today (Renotation Munsell) was standardized by asking participants to make marks on paper to show the relative degree of perceived similarity between colors varying in terms of either hue, lightness, or saturation, or alternatively simply to rate numerically the relative distance between pairs of neighboring colors (Newhall, 1939, 1950). This produced a large number of perceptual distances between pairs of colors, and the Munsell space was organized to try and reflect all these distances as accurately as possible.

In this article the universal foci are taken to be those colors chosen most often as the foci of color categories in the World Color Survey (Regier et al., 2005a, 2005b). This places universal focal red, yellow, green, and blue at hues 1, 9, 17, and 29, respectively, which, as hue is a circular dimension with length 40, makes the distances between red and yellow and yellow and green 8 hue units, and those between green and blue and blue and red 12 hue units.[7] However, yellow is lighter than the other colors, at lightness level C, whereas red is at lightness G, and green and blue are at lightness F (each letter indexes a successively darker degree of lightness). We can convert these Munsell coordinates into points in the CIE L*a*b* color space, using Munsell–CIE L*a*b* conversions given in the World Color Survey. Because the CIE L*a*b* color space is a different shape, the distances between universal foci in this space will be different to those in the Munsell color space. In fact, there are many other color systems that attempt to represent colors in relation to their perceptual similarity, but each places the universal foci in somewhat different locations.

Even if we know the location of the universal foci in a color space, this does not necessarily allow us to determine the perceptual distance between each pair of universal foci. For example, in the Munsell color space, there are 12 hue units and 1 lightness unit (*value* in Munsell terminology) between focal blue and focal red, but 8 hue units and 3 lightness units between focal yellow and focal green. This seems to show that focal yellow is more similar to focal green than focal red is to focal blue. However lightness and hue are measured using different units in the Munsell system (Newhall, 1939), so there is no straightforward way to compare the perceptual distance between two colors varying in hue with that between two varying in lightness. If each lightness unit corresponded to a greater perceptual distance than each hue unit, then focal yellow could be farther from focal green than focal blue is from focal red, the converse of what would be the case if one lightness unit represented the same or a smaller degree of difference than one hue unit. It is not even clear that we can say that an amount of difference in hue is equal to a particular difference in lightness; hue and lightness are two different dimensions, and so may be fundamentally incomparable (Newhall, 1939).

In the CIE L*a*b* color space, an attempt was made to standardize the lightness, hue, and saturation units, so that differences in one dimension could be compared with differences in another. Typically we assume that the CIE L*a*b* space is Euclidean, and so the distance between any two colors can straightforwardly be calculated using Pythagoras' theorem (Kuehni, 1997). If we calculate the differences between the universal foci using this methodology, we find that the distances between red and yellow, yellow and green, green and blue, and blue and red are 113.4, 110.9, 97.6, and 93.1, units respectively. This places the foci at fairly even distances, with red and yellow being most dissimilar, and red and blue being most similar. However, the CIE (CIE, 1995) provides an alternative, more sophisticated color difference formula that takes account of color differences being more significant in some parts of the color space than in others, so if this new formula was used to calculate color differences, we would obtain quite different results. Using the Optical Society of America Color Space, Boynton and Olson (1987) calculated the distances between the centers of the areas of color named by English red, yellow, green, and blue, as 6.5 between green and blue, 7.3 between yellow and green, 11.2 between blue and red, and 12.2 between red and yellow, suggesting that green and blue are the most similar, and red and yellow the most dissimilar, but contradicting the CIE L*a*b* distances reported above.[8]

Indow and Ohsumi (1972) and Indow (1988) investigated how well the Munsell color space represents perceptual distances. Indow and Ohsumi (1972) used multidimensional scaling to create a color solid based on participants' judgments of perceptual similarity. They found that the Munsell system reflects perceptual distances reasonably well, except that it places red and blue too close together. In the corrected version of the Munsell system, we find the distance between the unique red hue and unique yellow hue to be 7.56 Munsell hue units, that between unique yellow and unique green 9.78 units, that between unique green and unique blue 7.44 units, and that between unique blue and unique red 15.22 units.[9] In this system, green and blue are the closest of the unique hues, and blue and red the farthest apart. This contradicts both the Munsell and CIE L*a*b* systems, but should be a more reliable result, given the methodology used (multidimensional scaling of human judgments of perceptual similarity). We should note, however, that these are the distances between the unique hues themselves, and do not take account of, for example, focal yellow being much lighter than the other universal foci. If we did so, we would find that the distances between yellow and red and yellow and green would be expanded.

The results reviewed in this section show that the various standard color spaces in use all give quite different distances between the universal foci, and they do not even agree on which universal foci are closest together and which are furthest apart. We should also note that the distances within the perceptual color space "are stretched or distorted by the influence of linguistic categories" (Roberson et al., 2000, p. 394), which implies that slightly different perceptual distances will be obtained from speakers of different languages. There are frequent references in the literature suggesting that focal green and blue are more similar to each other than are the other universal foci, but Hardin (1999) noted that this has never been studied systematically. In the color space used in the computer model reported here, the rank of distances between universal foci, from most dissimilar to most similar, was red and blue, red and yellow, yellow and green, and green and blue. This is similar to the rank of distances derived from Indow and Ohsumi (1972), except that they found red and yellow to be more dissimilar than yellow and green. (Although if we took account also of the lightness dimension, we could find that the order of these two pairs was reversed, depending on to what extent we weighted differences in lightness compared to differences in hue.)

Although much of the work reviewed here has tried to determine the correct spacing of colors in the perceptual color space, we should consider whether colors really can be represented in three-dimensional space in such a way that the distance between each pair of colors is proportional to their dissimilarity. Most color similarity judgments have been taken over small distances, and so it is not clear that they are valid over larger distances (Indow, 1988). If points A, B, and C lie along a straight line, with B in the middle, then the distance from A to C is equal to the distances from A to B plus the distance from B to C, but color dissimilarities need not follow this same rule. Perceived dissimilarity could be either increased or decreased as we consider larger as opposed to smaller distances. Furthermore, even if large perceptual distances are equal to the sum of smaller constituent ones, there need not be an arrangement of colors in Euclidean space that accurately captures the relative similarities of all pairs of colors. Indow (1988) showed that a Euclidean color space can accommodate perceptual similarity judgments quite well, but only for small regions of the color space, not for large distances such as those between the universal foci. Therefore it seems that the color space is only locally Euclidean, and that it may not be possible to capture larger color distances accurately in a Euclidean space.[10] Therefore we should probably consider color spaces to be only generally indicative of the perceptual distances between colors. Hardin (1999) noted that "it is misleading to speak of '*the* psychological color solid' as if there were a unitary and simple psychological model that captures the entire range of color phenomena as we experience them" (p. 954). The color space we use when learning color words could be quite different from the color space we use when participating in psychophysical experiments.

I hope I have shown in this section that the spacings between the universal foci in the color space with respect to which color term denotations are defined is very much an open question. However, even though we do not have a clear measure of the distances between neighboring universal foci, a number of suggestions have been made that the typological patterns in color naming can be related to these distances (Kay, 1999; Palmer, 1999). Kay noted that green-blue was the most common composite color in the World Color Survey, but that there were no derived green-blue (turquoise) terms. In contrast, derived red-blue terms (purple) were the most common type of derived term, but there were no red-blue composite categories. He suggested that these data could be explained if blue and green were more similar than blue and red, as that would explain the tendency for blue and green, but not red and blue, to form composites, and for derived terms to appear between red and blue, but not in the narrow space between green and blue. However, he noted that this approach did not work when applied to the two other pairs of neighboring universal foci, because we see both yellow-red composite and derived terms (orange), whereas yellow-green composites are extremely rare, but neither do we see any yellow-green derived terms (lime). Furthermore, the just noticeable difference[11] data on color similarity that he considered, which placed green and yellow farthest apart, followed by red and yellow, red and blue, and finally blue and green, was not compatible with an explanation of color term typology in terms of the similarity and distinctiveness of the universal foci. However, we see later that the computer model reported here is able to explain typology in terms of the distances between universal foci, because the evolutionary aspect of the model allows it to account correctly for the relative frequencies of yellow-green and yellow-red composite and derived terms.

## 5. Expression-induction modeling

The computer model used to simulate color term evolution is a kind of *expression-induction model*[12] (Hurford, 2002). These models aim to simulate the process of language change, usually over several generations. They contain a number of agents, each of which is capable both of learning some aspect of language and of using the language that they have learned to express themselves, hence creating some example utterances from which other agents can learn. If a model is run several times, the general properties of the languages that emerge can be observed. If all the emergent languages have a particular property that is also a universal in real languages, or if the emergent languages show a limited range of variation, reflecting typological patterns, then the model suggests an explanation of why these universals or typological restrictions exist.

Probably the first computer model that could be classified as an expression-induction model is that of Hurford (1987). This model was used to explain universals in the way in which languages express numbers. The model showed how a universal rule could evolve as a result of a diachronic process, even though individual speakers had no obligation to follow the rule. The induction part of Hurford's model was extremely simple, but more recent expression-induction models, including the model of color term evolution reported here, have begun to use much more sophisticated learning techniques. Kirby (2002) and K. Smith, Brighton, and Kirby (2003) created expression-induction models of the evolution of syntax, and suggested that they could explain why compositionality is ubiquitous in human language. Other expression-induction models have been used to explain phenomena as diverse as vowel typology (de Boer, 1999), vowel harmony (Harrison, Dras, & Kapicioglu, 2002), the development of lexical semantics (A. D. M. Smith, 2003; Vogt & Coumans, 2003), and the necessary preconditions for a coherent language to evolve in a community (Barr, 2004).

In common with the work reported here, Belpaeme (2002), Steels and Belpaeme (2005), and Belpaeme and Bleys (2005) applied the expression-induction methodology to color term evolution, although the details of each of these models were quite different from those of the model presented here. Belpaeme's (2002) simulations typically contained 10 agents, each of which was able to represent color categories using adaptive networks, a kind of neural network, allowing color categories of almost any size or shape to be represented. Color in the model was represented in terms of the CIE L*a*b* space, which was chosen because Lammens (1994) showed that his computer model of color naming worked best in that space. Each agent could also remember a number of word forms, each of which could be paired with a particular color category.

An agent would be presented with a topic color, and some other colors to form a context. It would then try to find a color term, chosen out of those it had already learned, that could be used to name the topic, but not the context colors. If an agent was not able to do this, it would modify its existing color categories to increase their discriminative potential, or add a new category. In some simulations, no communication between agents took place, so each agent would simply try to develop a color category system that was effective at discriminating target from context colors, through repeated iterations of this step. In simulations where communication was simulated, the speaker would then pass the word to the hearer, who would try to identify the topic color based on a category that it had already associated with that word. When communication was successful, the association between the topic color and the color word would be strengthened, and when it was not successful, the hearer would be shown the correct topic, and it would adapt its knowledge of color categories and color words to reflect this new information.

The most important result of Belpaeme (2002) was that, so long as communication was simulated, coherent color lexicons emerged that were shared by all the agents. The color lexicons would divide the color space into a number of color regions, each of which would be associated with a particular color word. The agents never agreed completely about the exact meaning of each word, but their languages were consistent enough for them to achieve rates of communicative success in excess of 85%. However, the color categories emerging in Belpaeme's model did not resemble the color terms of real languages, as they did not conform to the typological restrictions observed in color term systems cross-linguistically.[13]

Belpaeme and Bleys (2005) reported simulations conducted using a modified version of Belpaeme's (2002) model. In the new model, each color term was represented by locating its center at a point in the color space, and using the Euclidean distance from that point as a membership function. This representation is much simpler than that used by Belpaeme, but the model no longer has the ability to represent color categories of any arbitrary shape, as they are now defined by a single point. Belpaeme and Bleys (2005) used this new model to investigate whether typological patterns could be explained in terms of the constraints embodied in the CIE L*a*b* color space, the colors present in the environment, and the linguistic transmission process through which color categories are passed from one person to another. Following Yendrikhovskij (2001), they investigated the effect of colors present in the environment, by selecting example colors from those contained within digital photographs of natural scenes, instead of choosing them at random. This should bias the range of colors toward those occurring in the environments in which most languages included in the World Color Survey evolved. In particular, there would be few highly saturated colors, as most natural objects do not have such vivid colors. The effect of the linguistic transmission process was investigated by running simulations in which no communication between agents took place, and comparing them to simulations in which whole communities of agents developed shared color category systems through communicative interactions.

Belpaeme and Bleys (2005) found similarities between the World Color Survey and the languages emerging in their simulations, in terms of the locations of the centers of the emergent color categories. In the World Color Survey, most of the centers of basic color term denotations are clustered in a few parts of the color space, and the same result was found in Belpaeme and Bleys's simulations. However, the similarity was greatest when communication was simulated, and example colors were selected at random. This shows that while considering the distribution of colors in the environment did not help to explain universal patterns in color naming, the communicative aspect of Belpaeme and Bleys's model did. Belpaeme and Bleys' simulation therefore demonstrates that universals in color term naming can be explained partly in terms of the shape of the color space, but that a better explanation is obtained if linguistic transmission is also modeled.

The model of color term evolution presented in this article uses a similar methodology to Belpaeme (2002), Steels and Belpaeme (2005), and Belpaeme and Bleys (2005), in that it is also a kind of expression-induction model, and it also simulates linguistic transmission, and chooses example colors at random. However, it learns and represents color categories in a completely different way, and there is no feedback mechanism that informs the agents whether communication has been successful. It also does not make the distinction between color words and color categories that is made in those papers, as each word simply names a range of colors directly, without reference to a prelinguistic category. The model took account of the special status of the universal focal colors and their uneven spacing in the perceptual color space, and it was thus able to provide an explanation concerning which types of color terms tend to emerge, and why languages appear to evolve along a small number of fixed trajectories.

## 6. A Bayesian model of color term acquisition

The acquisitional part of the expression-induction model of color term evolution learns the denotations of color words using Bayesian inference, but with an even prior over hypotheses so that no type of color term is favored a priori. It uses a version of the *size principle* of Tenenbaum (1999) and Tenenbaum and Griffiths (2001), which has the effect of favoring more restrictive hypotheses over less restrictive ones when both explain the same data. Tenenbaum (1999) showed how the size principle can be used to generalize from examples to a more general underlying concept. Griffiths and Tenenbaum (2000) showed that when this technique was used to predict the frequency of recurrent events based on a limited number of examples, its predictions closely paralleled those made by humans. Tenenbaum and Xu (2000) showed how the technique could be adapted to learn word meanings, demonstrating how the full denotational range of a word could be estimated based on some example denotation. Again it was shown that the Bayesian model made very similar generalizations to those made by humans.

To create the model of color term acquisition, it was necessary to make a number of assumptions about how children learn color words. The model is at present concerned only with a simplified color space, containing a single circular dimension that includes all the universal foci, as shown in Fig. 3. This simplification, which is the primary limitation of the model, means that it is not possible to account for the meanings of some color terms, such as *black, white, pink, brown,* and *gray,* but the acquisition of *red, orange, yellow, green, blue,* and *purple* terms can be modeled.

Another assumption was that the universal foci are not evenly spaced in the color space. The color space was divided into 40 discrete colors, so each individual color could be indexed with a number between 1 and 40.[14] The positions of the universal foci were adjusted to give the best fit to the typological data, which was obtained when focal red was placed at hue 7, yellow at 19, green at 26, and blue at 30, so that the largest distance between adjacent universal foci was 17

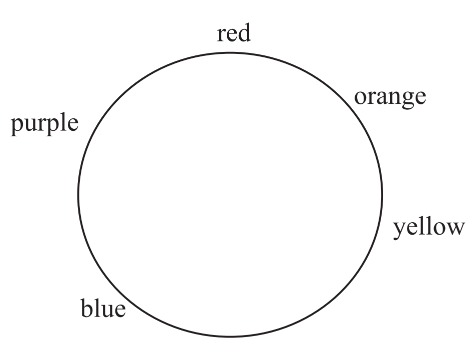



Fig. 3. The perceptual color space.

units, between blue and red, and the smallest was just 4 units, between green and blue.[15] In section 4, the distances between universal foci in several standard color spaces were discussed, but these distances do not correspond to the distances found in any of those systems. However, as the distances between the universal foci in the various perceptually standard color spaces vary radically, using universal foci locations derived from one of those color spaces would seem to be just as arbitrary as using the locations chosen here.

It might well be argued that these universal foci locations exaggerate the relative dissimilarity of blue and red compared to the similarity of green and blue beyond the greatest difference seen in any standard color space. However, changing any one part of the model (e.g., extending the color space to cover all three color dimensions, or changing the learning mechanism used) could be expected to change the exact parameter values required to replicate the typological patterns. Therefore rather than considering these universal foci locations to be the correct locations, it might instead be better to consider them only to be indicative of the relative distances between universal foci.

The next issue to be considered is what data are available from which people can learn color words. Children are typically not taught the full range of the denotations of each word they know explicitly, in terms of exactly what it can and cannot be used to denote, and so they must learn the meanings of color words primarily by observing what colors other people use those words to refer to. Hence the data from which the model learned consisted of examples of colors that a color term had been used to name. Learning then consisted of generalizing from such examples to the full range of colors that a word could denote. For example, the model might see three different shades of red, all named by the word *blah.* Learning would then consist of generalizing from those examples to estimate the full range of colors that could be denoted by the term *blah.* Earlier, evidence suggesting that the universal foci are especially salient was discussed. This suggests that agents are more likely to remember them as example denotations of color terms than they are to remember other colors. The model replicates this finding by remembering all examples of universal foci, but only 1 in 20 of the nonfocal examples.

Each color term was learned completely independently of any other color terms, so what had already been learned about the denotation of one term was not used to help learn other terms. When children learn their language they might potentially make use of the principle of contrast, and so assume that colors denoted by one term are not denoted by another, or they might infer from phrases such as *greenish-yellow* which color terms denote adjacent parts of the color space. However, it was decided to keep the acquisitional model as simple as possible, and therefore not to incorporate any such secondary cues.

The model has no a priori preference favoring color terms of any particular size, or terms that occur in any part of the color space. The model does, however, assume that color terms denote a continuous range of colors, a restriction that Roberson et al. (2000) suggested is a universal constraint on color categories (so, e.g., any color word that denotes both red and yellow must also denote the intermediate orange colors). In Bayesian terms this translates into the model considering hypotheses that vary in size from taking up one unit of the color space to including the whole of the color space. These hypotheses can occur anywhere in the color space, but must correspond to a contiguous range of colors. Because there is no built-in preference about the size or location of color terms' denotations, the model assumes that each hypothesis is equally likely a priori, so the effect of the prior is simply to determine which hypotheses are possible, rather than to distinguish between hypotheses that are more or less likely a priori.

Finally, we need to address the issue of how the model will assess the probability of each possible hypothesis, based on the example colors that have been observed. We would normally expect that if a hypothesis about the denotation of a color word were correct, then all the examples of colors that it has been used to name would come within the scope of the hypothesis. The size principle (Tenenbaum, 1999; Tenenbaum & Griffiths, 2001) tells us that the hypothesis that most tightly constrains the examples will have the highest probability. In other words, the hypotheses that cover the smallest ranges of colors will have the highest probabilities (so long as they contain all the examples), as they allow the fewest potential locations at which color examples can occur. The model also assumes that colors are equally likely to be observed anywhere in the color space, but it has to compensate for only some of the nonfocal examples being remembered.

As described so far, the model assumes that all the data it receives are completely accurate, and so that if a hypothesis is correct, all the examples will have been observed within the corresponding range of colors. However, under this assumption, the model would fail as soon as any incorrect examples were shown to it, because it would have to assign a probability of zero to any hypothesis that excluded even one example from a term's denotation. Clearly, when children learn color words they must be able to cope with misleading examples. (Perhaps they misunderstand which object is referred to using the term *red,* and so associate red with the color yellow.) Therefore, a parameter, $p$, was added to the model, which corresponds to a learner's belief concerning how likely it is that each individual example is correct. (An example would be incorrect, or erroneous, if it consisted of a word paired with a color outside of the range of colors that could correctly be named by the word.) With the addition of this parameter, the model now prefers hypotheses that predict the location of most of the examples, even if there are a few examples outside of the range of a hypothesis.

Finally, the model bases its final estimate of the correct denotation of a color word not on a single hypothesis, but on all hypotheses, using the standard Bayesian procedure of hypothesis averaging. This entails considering each color in turn, and summing the probabilities that have been calculated for each hypothesis that includes that color. The resultant probabilities concern how likely it is that each color can be denoted by the color word. If these probabilities are equated with degree of membership in a fuzzy set (Zadeh, 1965), each color will be given a degree of membership between 0 (*not a member at all*) and 1 (*full membership*) in the color category corresponding to the color term. Because the example colors are not considered to be totally reliable, the model will usually be most certain about the membership of colors toward the center of a category, and so these will have a higher degree of membership than those toward the edges, resulting in the prototype properties that are characteristic of color terms.

The preceding has described the Bayesian model, but in an informal way. Now it is described more formally, using equations, which should clarify any aspects of the model that are unclear. Bayesian inference depends on calculating the posterior probability of a hypothesis ($P(h \mid d)$), based on the prior probability of that hypothesis ($P(h)$), using Bayes's rule, given in Equation 1. (If we know the full range of possible hypotheses, and the prior probability of each, as we do in this case, then these terms can be used to derive the final term in Bayes's rule, $P(d)$.) In the case of this model, the data, $d$, consist of all the observed example colors for the word whose denotation is being learned, and each hypothesis, *h,* corresponds to a range of color that the word might denote. All hypotheses have an equal probability a priori, so *P(h)* is equal for all hypotheses.

$$P(h \mid d) = \frac{P(d \mid h)P(h)}{P(d)} \tag{1}$$

The probability of the data with respect to a hypothesis, $P(d \mid h)$, will depend on how accurately the hypothesis predicts the observed examples. If an example is accurate, then it must appear within the range of the hypothesis. If that is all we know about an example, then it is equally likely for that example to have been observed on any of the colors with the range of the hypothesis, assuming that the hypothesis is correct (this makes the assumption that accurate examples are sampled randomly and independently from the range of colors denoted by the color word). However, because only 1 in 20 nonfocal examples will be remembered, we would expect to have remembered more examples for some colors than for others. The probability of remembering an example for color $C$ will be written $R_c$, and was set at .05 for colors that did not correspond to universal foci, and at 1 for universal foci. As examples that are not remembered have no effect on the model, the better memory for universal foci was in practice always implemented by simply generating more examples of universal foci than of other colors. This has exactly the same effect on the model as if all the nonfocal examples were generated, and some of them were then forgotten. For each example, the probability that we would have remembered it at any particular color would be equal to the probability of remembering examples of that color, divided by the sum of the probabilities of remembering examples of all the colors within the range of the hypothesis, which will be written as $R_h$. This ratio, which is given in Equation 2, would correspond to the probability of an example when that example was within the range of the hypothesis, and when we knew both that the hypothesis was correct and that the example was accurate.

$$\frac{R_c}{R_h} \tag{2}$$

Erroneous examples are assumed to be equally likely to be observed anywhere in the color space, and so the probability of an erroneous example being remembered at any particular color is equal to the probability of remembering an example if it occurs at that color, divided by the sum of the probabilities of remembering examples of all colors throughout the color space ($R_t$). This ratio is expressed in Equation 3.

$$\frac{R_c}{R_t} \tag{3}$$

Equations 2 and 3 apply when we know whether an example is accurate or not, but in reality, when a person has remembered an example he or she will not be sure whether it is accurate. If we see an example outside of the hypothesis space, we know that it must be inaccurate. Because the probability that an example is accurate is *p,* the probability that it is not accurate is 1 –

$p$. Hence the overall probability of an example, $e$, that comes outside of the range of a hypothesis, is given by multiplying Equation 3 by this value, as shown in Equation 4.

$$P(e \mid h) = \frac{(1-p)R_c}{R_t} \tag{4}$$

If an example is within the scope of the hypothesis then we cannot be sure whether it is accurate or not (because it could have come within the range of the hypothesis simply by chance). So, in the case of such an example, we have to add its probability assuming that it is accurate, to what its probability would be if it was erroneous, each of which must be weighted by the probability of examples being accurate ($p$), or inaccurate ($1 - p$). The resulting overall probability of such examples is given by Equation 5.

$$P(e \mid h) = \frac{pR_c}{R_h} + \frac{(1-p)R_c}{R_t} \tag{5}$$

Equations 4 and 5 allow us to calculate the probabilities of individual examples with respect to a hypothesis, but usually we will have several examples for a particular color word, so we need to combine these individual probabilities to obtain an overall probability for all the data. This can be done simply by multiplying together the probabilities of each individual example, $e$, from the set of all examples, $E$, as shown in Equation 6. For every example we must use either Equation 4 or 5 to calculate $P(e \mid h)$, depending on whether or not the example is within the scope of the hypothesis.

$$P(d \mid h) = \prod_{e \in E} P(e \mid h) \tag{6}$$

To determine hypotheses' a posteriori probabilities, we also need to be able to calculate the probability of the data irrespective of any particular hypothesis, $P(d)$. We can calculate this probability by multiplying the probability of the data given each individual hypothesis by the a priori probability of the hypothesis, and then totaling the resulting probabilities for each hypothesis in the set of all possible hypotheses, $H$. This is expressed mathematically in Equation 7.

$$P(d) = \sum_{h \in H} [P(h)P(d \mid h)] \tag{7}$$

If we substitute Equation 7 into Bayes's rule, we obtain Equation 8, which we can simplify by canceling out the constant terms $P(h)$ and $P(h_i)$. (The $h$s of Equation 7 now have a subscript $i$ to distinguish them from the specific hypothesis under consideration, $h$. However, as the a priori probability of all hypotheses is equal, each $P(h_i)$ will be equal to $P(h)$.)

$$P(h \mid d) = \frac{P(h)P(d \mid h)}{\sum_{h_i \in H} [P(h_i)P(d \mid h_i)]} = \frac{P(d \mid h)}{\sum_{h_i \in H} P(d \mid h_i)} \tag{8}$$

Equation 8 lets us calculate the posterior probability of individual hypotheses, each of which corresponds to a possible denotation of the color word. Finally, we use Equation 9 to determine a degree of membership for each color *x* in the color category *C* named by the color word, by summing the posterior probabilities of all the hypotheses in *S,* the set of hypotheses that include color *x*.

$$P(x \in C \mid d) = \sum_{h \in S} P(h \mid d) \tag{9}$$

The ability of the model to learn color term systems was investigated by presenting it with examples corresponding to the color term system of Urdu. The denotations of Urdu color terms shown on a chart in Berlin and Kay (1969) were mapped onto the model's color space, and example colors were generated based on these denotations. Both the boundaries within which examples were generated for each color term and the learned denotations are shown in Fig. 4. As with all the simulations reported in this article, the parameter *p* was set to .5. This corresponds to a belief by a child learning the language that he or she is only able to identify the color named by a color word correctly on half of all occasions. However, the exact value of this parameter was chosen fairly arbitrarily, and minor variations in this parameter do not appear to greatly alter the model's predictions.

Random hues from within the range of colors denoted by each color term were selected, and these examples were passed to the model until it had remembered 40 of them. (As described earlier, examples of each focal color would be generated, on average, 20 times as often as each nonfocal color, to simulate the increased tendency to remember focal examples over nonfocal ones.) In the representations learned by the model, each of the color terms had prototype properties, as each had a single best example and the degree of membership declined gradually moving away from that color. Each term that contained a universal focal color had that color as its prototype, even when this was near the boundary of the color category rather than at the center. Both of these phenomena are consistent with empirical findings, and show that the model is

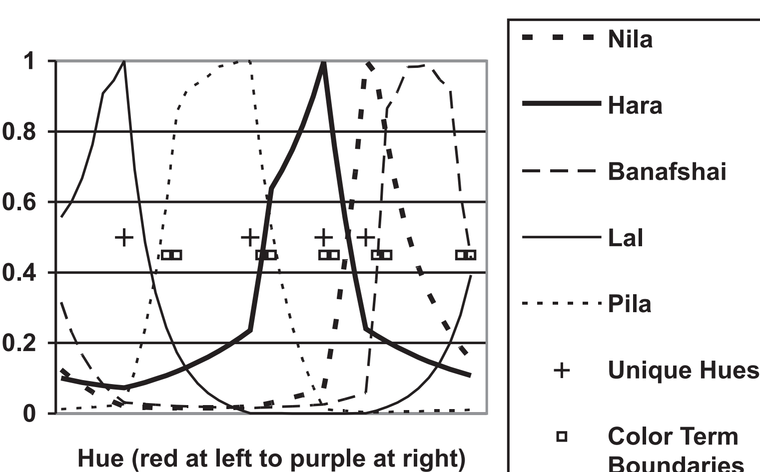



Fig. 4. Learned denotations for Urdu color terms (*y* axis shows degree of membership in each color category).

able to learn the color term system of a real language. However, the acquisitional model does not explain the typological data, because it is also able to learn color term systems of unattested types.

## 7. The evolutionary model

The acquisitional model described in the previous section was used as part of an expression-induction model. To simulate a whole community of people, 10 copies of the acquisitional model were created, each of which acted as an agent. These agents could induce color word meanings based on examples that had been expressed by other agents, hence allowing the evolution of color words to be simulated.

In the initial state of the model, each agent had observed a different color term together with one completely random example of it. These color terms would be used in the first interactions between agents, and some of them would become established in the language, whereas others would likely be lost within the first few generations. Each agent was assigned a random age between zero and the maximum age to which agents could live, so that the first generation of agents would not all die at the same point in the simulation. The simulations then proceeded using the algorithm summarized in Fig. 5. First one agent was selected at random to speak, and another to hear. A color for the speaker to name would then be chosen. This color would be randomly selected, but each universal focus was chosen 20 times more often than each of the other colors. (This ensured that the model would remember 20 times as many examples of universal foci as of other colors, as described earlier.) The speaker would then find the word that they thought most likely to be the correct label for the color, based on all the color examples it had observed up to that point. This word, together with the corresponding color, would then be observed by the hearer, and remembered by it as an example.

However, 1 time in every 1,000, instead of the speaker choosing the best word based on the observations it had made, it would be creative instead, and make up a new word. This made it possible for new color words to enter the language, so there was no upper bound on the number of color terms that could emerge in the simulations. If agents were not able to introduce new words in this way, the number of words present would have gradually decreased throughout each simulation, as words will occasionally be lost if by chance they are never used productively.

A parameter in the model controlled how long each agent lived for, measured in terms of how many color examples agents remembered during their lifetimes. (As agents are hearers, and hence remember color terms, as often as they are speakers, this is equivalent to measuring life span in terms of one half of the average total number of interactions in which an agent partakes.) The actual life span of each agent was varied randomly by an amount of up to 20% either above or below the chosen average life span. Once an agent reached the end of its life span it would be replaced by a new agent that had not observed any color term examples. (If an agent was chosen as the speaker before it had observed any color terms, then the program would just go back and choose another agent instead.) This evolutionary model makes explicit the process through which it is proposed that color language is passed from person to person over several generations. Because the language is transmitted through a bottleneck (because it is passed

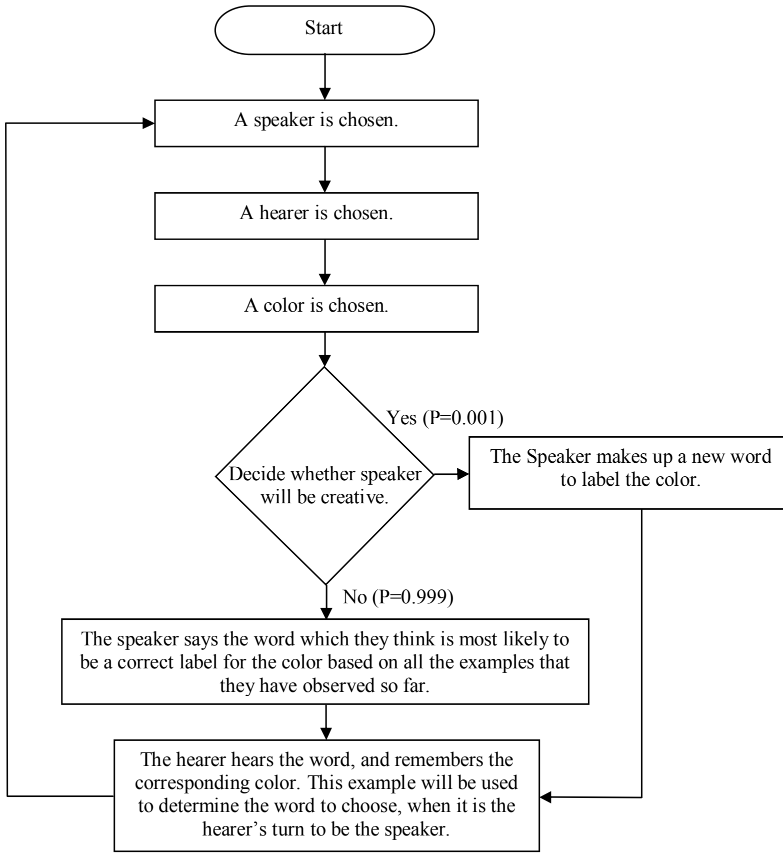


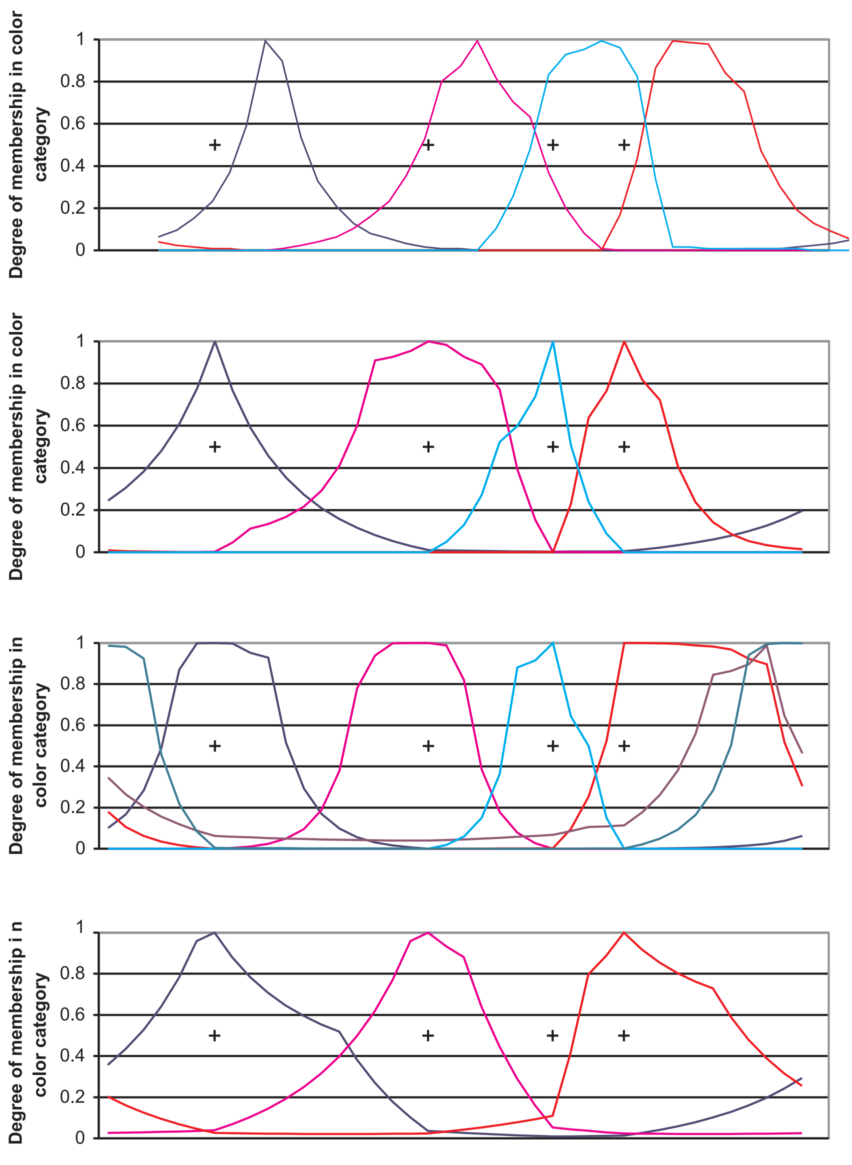

Fig. 5. Outline of the evolutionary algorithm ($p$ stands for the probability of making each choice).

from agent to agent only through a limited number of examples), we would expect agents to reproduce the color term system imperfectly, resulting in gradual semantic drift.

## 8. Simulations and results

The expression-induction model described previously was used to simulate the evolution of color term systems, and at the conclusion of these simulations the emergent languages were analyzed to determine whether they reproduced the typological patterns identified by Kay and Maffi (1999). Simulations were conducted with the life span parameter set variously at 18, 20, 22, 24, 25, 27, 30, 35, 40, 50, 60, 70, 80, 90, 100, 110, or 120. The model was run 25 times in each of these conditions, resulting in a total of 425 separate simulations. Each time the model was run for a time equal to 20 average life spans, and the results reported here are based on the languages spoken by the agents at the end of the simulations.

A general picture of the kind of color term system typically emerging in the simulations can be obtained by examining Fig. 6. This shows the color term systems of four agents from the end of one evolutionary simulation, in which the average life span was set at 100. As is the case with all the results reported in this chapter, only color terms for which the agent had observed at least four examples were included. This was because one of the necessary criteria for a color term to be considered basic is that it must be salient for a speaker (Berlin & Kay, 1969), and it seems reasonable to propose that if a person has observed only one or two examples of a color




(a) (b) (c) (d)

Fig. 6. The basic color term systems of four agents from the same simulation (+ marks universal foci).

word, then that word would be less salient for that speaker than one for which he or she had observed more examples. We can see that, with respect to this criterion, Speakers a and b both knew four basic color terms, with their prototypes at the red, yellow, green, and blue universal foci. Speaker c, who was older than Speakers a and b, and hence had seen more color term examples, knew these same four terms, plus two extra terms, which both denoted purple hues, and Speaker d, who was younger, and had seen fewer examples, did not know the green term, and so had only three basic color terms. Whereas Speakers a and b had both observed examples of each of the purple terms, they had not seen enough examples to consider either of these terms to be basic. Speakers a, b, and c all also knew an orange term, but none of them had seen sufficient examples of its use for it to achieve basic color term status.

Fig. 6 shows that a color term system emerged, most aspects of which were shared by most members of the community, but in which there was considerable variation between individual agents. The variation concerned which color terms each agent knew, which they considered to be basic, and the exact denotation that they had learned for each. Sources such as MacLaury (1997a) reported that these are all phenomena that are prevalent in real language communities, and so these results, which are representative of the simulations as a whole, are consistent with the empirical evidence. More terms tended to emerge in simulations in which the agents had longer life spans, varying from a mean of 2.13 terms with an average life span of 18, up to 5.18 terms when the average life span was 120. In this, and all further analyses, only agents whose age was greater than or equal to half the average life span were included, as younger agents could not have been expected to have reliably learned the color term system spoken by the other agents.

To investigate whether the typological patterns identified by Kay and Maffi (1999) were replicated in the simulations, it was necessary to classify each emergent language in terms of which kinds of basic color terms it contained. These classifications were based on the language emerging in each simulation as a whole, rather than on the language of individual agents. This was so the analysis of the results could be based on how many times each type of language arose independently. If we had simply included each agent in a simulation individually, we would have gotten the impression that the type of color term system that was predominant in that simulation had arisen on several occasions, when in fact it had occurred only once. An alternative approach would have been to base the analysis on the language spoken by a single agent from each simulation, but it was decided that a better picture of the simulation results would be obtained by taking a consensus from several agents, and that this approach was also more consistent with that taken in the analysis of the World Color Survey data reported by Kay and Maffi (1999).

The typological analysis was begun by first finding the color term each agent would use to name each hue. Fig. 7 shows the result of this process for one community that had six agents old enough to be included in the analysis, with the results for each agent on a separate row. Each column corresponds to a color, with hue 1 at the left and hue 40 at the right. Four different color terms are represented by the letters *A, B, C,* and *D.* The denotation of each color term was considered to be the smallest range of colors that included all those hues that the term would name. (This could potentially include some hues that would be named by a different color term, because there might be another term that had greater confidence values for some of the hues within the range of hues named by the first term.) Color terms were classified as red, yellow, green, or blue if their denotations included the corresponding universal focus. Terms whose denotations did not include any universal foci were classified as orange, lime, turquoise, or purple, depending on which universal foci their denotations came between. If a color term included more than one universal focus it would be classified as a composite of those universal foci; for example, red-yellow or yellow-green-blue.

The next stage of the analysis consisted of determining which color terms the language spoken by each community as whole could be said to contain. This was not straightforward, because the agents usually did not agree exactly on the denotation of each word, nor did they necessarily use identical sets of color words. In the community shown in Fig. 7 the classification of all of the terms would be the same for the first five speakers, as each contains just one universal focus, and this is the same universal focus for each of those speakers. However, the sixth speaker names red hues with term *C,* which was therefore classified as a red-yellow composite term for this speaker.

A number of rules were devised to arrive at a consistent classification for each emergent language, even when the agents who spoke it were not consistent in the way in which they named colors. First, a color term was included in the analysis only if it was considered basic by at least half the agents who were sufficiently old to be included in the analysis. If all the agents did not agree on the classification of a term, then the classification that was supported by the greatest number of agents would be chosen. If two or more possible classifications were supported by equal numbers of agents, then if one of the terms contained fewer universal foci it would be chosen, but otherwise the whole language would be excluded from the analysis. Agents who had not observed at least four examples of two or more color terms were also not considered. After the application of all these criteria, a unique classification was obtained for the languages emerging in 420 of the 425 runs of the simulation.

The number of terms that were classified as being of each type in all the emergent languages is listed in Table 1. For the terms that contain a universal focus, these data were converted to percentages, shown in Fig. 8. Both Table 1 and Fig. 8 also contain equivalent data from the World Color Survey, as reported in Kay and Maffi (1999). Kay and Maffi did not take account of derived terms in making their classifications, hence no figures for those terms appear here. The counts of each type of color term for the languages in the World Color Survey do not include terms from languages that were in transition between evolutionary stages, or that did not fit on the evolutionary trajectories at all. Color terms in the World Color Survey that were either achromatic or distinguished from other colors on the basis of some dimension other than hue

| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | A | A | A | A | A | $\boxed{A}$ | A | A | A | A | A | C | C | C | C | C | C | $\boxed{C}$ | C | C | D | D | D | D | $\boxed{D}$ | D | B | B | $\boxed{B}$ | B | B | B | B | B | B | B | B | B | A |
| B | A | A | A | A | A | $\boxed{A}$ | A | A | A | A | A | A | A | C | C | C | C | $\boxed{C}$ | C | C | C | D | D | D | $\boxed{D}$ | D | D | B | $\boxed{B}$ | B | B | B | B | B | B | B | B | B | B |
| B | B | A | A | A | A | $\boxed{A}$ | A | A | A | A | A | A | C | C | C | C | C | $\boxed{C}$ | C | C | D | D | D | D | $\boxed{D}$ | D | D | B | $\boxed{B}$ | B | B | B | B | B | B | B | B | B | B |
| A | A | A | A | A | A | $\boxed{A}$ | A | A | A | A | A | A | A | C | C | C | C | $\boxed{C}$ | C | C | C | D | D | D | $\boxed{D}$ | D | D | B | $\boxed{B}$ | B | B | B | B | B | B | B | B | B | A |
| A | A | A | A | A | A | $\boxed{A}$ | A | A | A | A | A | A | A | C | C | C | C | $\boxed{C}$ | C | C | C | D | D | D | $\boxed{D}$ | D | D | B | $\boxed{B}$ | B | B | B | B | B | B | B | A | A | A |
| B | C | C | C | C | C | $\boxed{C}$ | C | C | C | C | C | C | C | C | C | C | C | $\boxed{C}$ | C | C | C | C | D | D | $\boxed{D}$ | D | B | B | $\boxed{B}$ | B | B | B | B | B | B | B | B | B | B |

← **Hue** →

Fig. 7. Denotations of basic color terms for all adults in a community (each row is a separate agent, and the boxed columns are universal foci).

Table 1
Frequencies of color terms of each type in the simulations and the World Color Survey

| Type of Color Term | World Color Survey | Simulations |
|---|---|---|
| Orange | n/a | 20 |
| Lime | n/a | 4 |
| Turquoise | n/a | 0 |
| Purple | n/a | 80 |
| Red | 70 | 382 |
| Yellow | 67 | 334 |
| Green | 26 | 191 |
| Blue | 27 | 214 |
| Red-yellow | 9 | 31 |
| Yellow-green | 1 | 12 |
| Green-blue | 50 | 170 |
| Blue-red | 0 | 1 |
| Red-yellow-green | 0 | 1 |
| Yellow-green-blue | 2 | 38 |
| Green-blue-red | 0 | 3 |
| Blue-red-yellow | 0 | 0 |

have simply been excluded from the analysis, and composite terms that included white or black and one or more universal foci were treated as if they only contained the universal foci. The total counts for each type of color term were derived simply by counting them once for every language in the World Color Survey that contained that type of basic color term.

Fig. 8 shows that there is a close relation between the frequency of each term in the World Color Survey and in the simulations. The key differences are that the simulations produce somewhat too many yellow-green-blue composites, and too few green-blue ones. We can also see that there are more purple terms than orange terms, which is consistent with the finding that purple terms are more common than orange ones (MacLaury, 1997a). There were no turquoise terms, which is also in line with expectations, as no language has a turquoise basic color term.

The simulations did produce a few terms of types that have not been attested empirically. There was one blue-red composite, one red-yellow-green composite, three green-blue-red composites, and four lime terms. The presence of a small number of previously unattested color terms should not be surprising. The evolutionary model does not place absolute restrictions on the types of color terms that can evolve, but simply introduces biases, so that some kinds of color terms emerge much more frequently than others. If the typological patterns seen in real languages are produced in the same kind of way, we would expect to occasionally discover new types of color terms as we looked at greater numbers of languages. As linguists have examined the color terms of more and more languages, color terms of types that were not found in Berlin and Kay's (1969) original survey have been discovered, but it is possible that some very rare types of color terms remain undiscovered.

Looking at the languages overall, 340 could be placed on Kay and Maffi's (1999) evolutionary trajectories. Nine languages deviated from the trajectories because they contained unattested color terms, 35 because no term consistently named one or more of the universal foci,

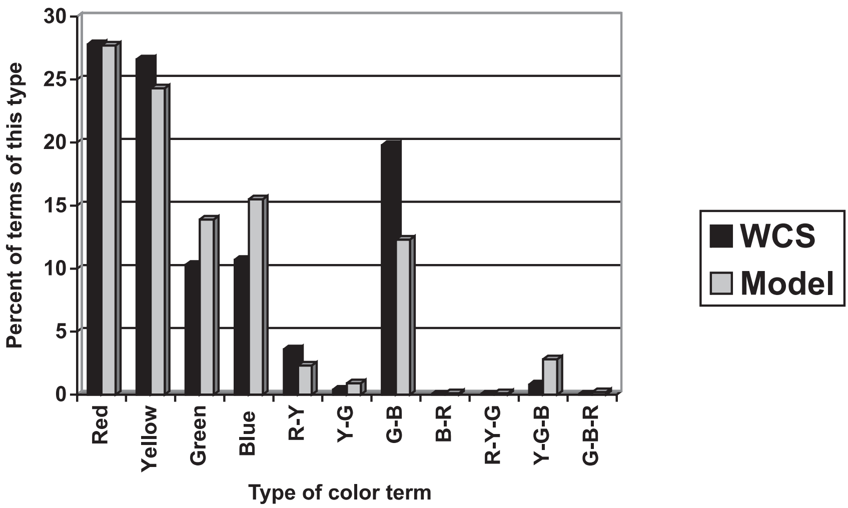



Fig. 8. Percentage of color terms of each type in the simulations and the World Color Survey.

and 37 because there was more than one term that could name one of the universal foci, or there was more than one purple term.[16] The 353 most common types of color term system were all 1 of the 11 types listed in Table 2, which includes equivalent data from the World Color Survey. (The World Color Survey data were collated using the same principles as those used to produce Table 1. We should note in particular that systems containing derived color terms were counted together with equivalent systems that did not have such terms, and hence there are no separate frequencies for such systems.) Table 2 contains examples of all of the types of systems that Kay and Maffi (1999) placed in their evolutionary sequences, plus three unattested types. (As Kay and Maffi did not report which languages contained derived terms, the presence of a single purple or orange term was ignored when determining whether a system could be placed on an

Table 2
The most common color term systems emerging in the simulations

| | Number of Languages With This Type of Color Term System | |
|---|---|---|
| Basic Color Terms in Language | World Color Survey | Simulations |
| Red, yellow, green, blue | 26 | 112 |
| Red, yellow, green-blue | 41 | 110 |
| Purple, red, yellow, green, blue | — | 44 |
| Red, yellow-green-blue | 2 | 30 |
| Red-yellow, green-blue | 9 | 22 |
| Orange, purple, red, yellow, green, blue | — | 7 |
| Red, yellow, blue | 0 | 7 |
| Red, green-blue | 0 | 6 |
| Orange, red, yellow, green, blue | — | 5 |
| Red, blue, yellow-green | 1 | 5 |
| Red, red-yellow, green-blue | 0 | 5 |

evolutionary trajectory.) Two kinds of system—red, yellow, blue, and red, green-blue—do not have a term that consistently names one of the universal foci. The third unexpected type of system is the red, red-yellow, green-blue system, which has two different terms that can be used to name focal red.

The simulations also produced 67 systems of types that occurred four times or less, most of which contained more than one term that could name a universal focus, or no term that consistently named one of the universal foci. However, nine of the less common systems contained an unattested type of color term, and five of them conformed to the evolutionary trajectories (four of these systems were purple, red, yellow, green-blue, and one was purple, red, yellow-green-blue). What is clear from these results is that there is a small set of color term systems that occur very frequently, and that the color term systems of the majority of languages can be classified as belonging to one of these types. There is also a significant subset of languages that diverge from the trajectories, in that they have extra color terms that could be classified as basic, one part of the color space is not named consistently, or the exact combination of terms does not conform to the expected pattern. However, this is consistent with the empirical findings of Kay and Maffi (1999) and MacLaury (1997a), who noted that a minority of color term systems do not appear to correspond to any widely attested type.

## 9. Discussion

The results of the simulations show that the typological patterns observable in basic color term systems can be explained if it is assumed that the universal foci are not evenly spaced in the perceptual color space, and that people remember the universal foci better than other colors. However, we should consider how confident we should be that this is the correct explanation for color term typology, and how closely we can expect the parameters of the model to correspond to their real-world equivalents.

The unequal spacing of universal foci is uncontroversial, and is exemplified by most color order systems, such as the Munsell system (Cleland, 1937; Newhall, 1939, 1950; Newhall, Nickerson, & Judd, 1943), as these systems do not place the universal foci at equidistant locations in the color space. The primary evidence for the universal foci locations used in the model is the data contained in the World Color Survey, but a question should be asked concerning just how closely those data constrain the locations. The set of parameter settings used was arrived at by a process of trial and error, but it was not possible to exhaustively test every possible set of parameters, so it is possible that there exists a quite different set of locations for universal foci that would account equally well for the typological patterns.

It was also not clear to what extent the results presented earlier were obtained simply because of the spacing of the universal foci in the color space, and to what extent the added salience of the universal foci affected the results. To investigate this possibility, the simulations were repeated without the universal foci having any special properties (i.e., the agents saw no more examples of the universal foci than of other colors). The locations of the universal foci, however, remained the same. The overall frequency of each type of color term was broadly similar to the previous simulations reported, except that there were far more derived terms (644 purple, 374 orange, 118 lime, and 16 turquoise) and more terms of unattested types (e.g., there were seven blue-red terms, compared to one in the previous simulations). Looking at the emergent languages as a whole, much greater divergence from the empirical data was apparent, as only 21% of languages conformed to the evolutionary trajectories, compared to 81% previously. This was either because of the presence of unattested types of color term, or because there were missing or extra terms. We can see from these results that the empirical data could be partly explained simply in terms of the locations of the universal foci, but that also modeling their added salience created a tendency for most color terms to have a universal focus and thus resulted in a fuller explanation of the data.

An interesting question concerning the model is whether the typological patterns could have been predicted directly from the uneven spacings of the universal foci in the color space, without recourse to evolutionary simulations. If the perceptual distance between two universal foci can be used to predict how often they appear in a composite, and how often a derived term appears between those universal foci, it could be argued that the evolutionary part of the model plays no explanatory role. However, this neglects the issue of which types of color term occur together. For example, from the spacing of the green, blue, and red universal foci, we can predict that green-blue composites and purple terms will both be common, but it is not so clear that purple will not usually occur together with a green-blue composite. However, that finding was replicated in the evolutionary simulations, in which the universal foci are usually distributed between the color terms as evenly as possible, given the number of terms (so, e.g., we usually would not see a composite color term containing two universal foci in a language that also had a derived color term containing none).

Further reasons why an evolutionary model is necessary become apparent if we pay close attention to the frequencies of composite categories. The distances between universal foci in the model do not actually correctly predict the relative frequency of each type of composite category, even though the model as a whole was successful in doing so. For example, focal yellow is further from focal red than focal green, but in both the simulation results and the World Color Survey there are more red-yellow terms than yellow-green ones. This is presumably because the strong tendency for green and blue to form a composite makes it less likely for green to be available to form a composite with yellow. Also, as focal red is so far from focal blue, there will be a tendency for it to form a composite with yellow, simply because that is the only other available focal hue. It would be impossible to predict the effect of all these interacting pressures just from the distances between the universal foci themselves, so the evolutionary part of the model is needed to predict both the frequency of each type of color term and the combinations in which color terms occur.

Another question that could be asked about the model concerns the particular learning algorithm used, a Bayesian learning algorithm based on the size principle of Tenenbaum (1999). Clearly, alternative assumptions could have been made when creating the Bayesian learning algorithm, which would have led to different generalizations being made from a particular set of examples. Furthermore, although there is a considerable amount of evidence supporting the hypothesis that people use an inference procedure that is based on, or is similar to, this form of Bayesian inference (Griffiths & Tenenbaum, 2000; Tenenbaum, 1999; Tenenbaum & Griffiths, 2001; Tenenbaum & Xu, 2000), many alternative learning algorithms have been proposed as models of human learning (e.g., consider the adaptive networks used by Belpaeme, 2002). Therefore it seems important to consider whether the choice of a Bayesian acquisitional procedure was critical to the results of the simulations, or whether the same results would have been produced if the Bayesian learner had been replaced by a different learning algorithm. Bayesian learning was chosen because it makes any prior biases completely clear and explicit, so that it was possible to be sure that the learning algorithm had no built-in preference favoring one type of color category over any other. However, I would expect that very similar results would be obtained if the Bayesian learner were replaced with another learning algorithm, so long as that algorithm also had no strong bias to prefer one type of color term over another. Therefore, although some form of acquisition procedure was necessary to complete the expression-induction model, the choice of Bayesian inference was essentially fairly arbitrary.

Overall, the major limitation of the model was that it neglected the dimensions of saturation and lightness. It would be desirable to extend the model so that it used a full three-dimensional color space, as this would allow the denotations of a greater range of color terms to be modeled, and would more accurately reflect the real-world perceptual color space. There is no reason in principle why this could not be done, but it would make the acquisitional model more complex. In this model, color term denotations are represented simply as linear sections of the color space, but if a three-dimensional color space were used, then color term denotations would correspond to three-dimensional volumes of the color space. It would therefore be necessary to specify a probability distribution over the possible shapes of the denotations, not just over their size and location, as is the case at present. This remains a possibility for future research, but it is worth noting that the well-known color model of Kay and McDaniel (1978) was also essentially one dimensional. Their model used fuzzy set operations on the outputs of opponent process cells (under the now discredited assumption that these could directly explain the denotations of color terms), but they only fully specified these in terms of single dimensions (either the hue or the lightness dimension). Although Kay and McDaniel clearly intended their model to cover the full three-dimensional color space, and hence the full range of color terms, this relied on extrapolation from, or extension of, the formal fuzzy logic operations.

When evaluating the computer model, attention should also be paid to the empirical data, as it is important to consider whether these give an accurate picture of the tendency for each type of color term system to emerge. Because many of the languages in the World Color Survey were probably closely related (either genetically, in the sense of being descended from a common ancestor, or because of language contact), some of the color terms and color term systems might not have arisen independently. This could skew the results so that some types of color term appeared to be common simply because they appeared in several related languages. Furthermore, the World Color Survey contains data on a fairly limited number of languages, so that there are very few data concerning the less common types of color term systems. The survey might therefore considerably overestimate or underestimate the frequency of some of these systems, and may have missed some types entirely.[17] Therefore we should treat the World Color Survey as being only broadly indicative of the types and frequencies of potential color term systems. Discrepancies between the World Color Survey data and the simulation results could potentially be attributable to inaccuracies in the frequencies derived from the World Color Survey, and need not necessarily indicate deficiencies of the computer model.

The limitations of the color space used, and the uncertainties about the World Color Survey data, should lead us to be cautious about what claims are made concerning the universal foci locations predicted by the model. They should probably be interpreted simply as being indicative of the relative distances between the universal foci in the perceptual color space. It seems likely that in reality the relative distances between universal foci are ordered from green and blue being closest, followed by green and yellow, then yellow and red, and with blue and red being the furthest apart. However, making a more precise claim about the locations of universal foci would be unjustified, and it is not possible to be certain even about this ordering of relative distances. A more cautious interpretation of the results would be to conclude that the computer model has shown how learning biases can influence color term evolution, and that such learning biases could result in the emergence of a range of languages that collectively mirror typological patterns, but not to make any more specific claim. Even if we do not accept that there is convincing evidence that the details of the model are correct, it has provided a plausible, detailed and rigorous explanation of data that have previously resisted a coherent explanation.

In conclusion, the model presented here has provided an account both of how color term denotations can be acquired, and of why we see typological patterns in color term systems cross-linguistically. The prototype properties of color term systems are emergent properties of the Bayesian learning mechanism used, whereas the typological patterns are emergent properties of the cultural evolution of color term systems over time. This generally supports Berlin and Kay's (1969) hypothesis of an evolutionary trajectory, but it allows for more flexibility in how languages can evolve, and can readily accommodate exceptional languages. The evolutionary model ties together observations concerning the universal foci and cross-linguistic typology, with a degree of explicitness that has not been achieved by any other theory.

## Notes

1. There has even been a recent report that one language, Pirahã (Brazil), has no basic color terms at all (Everett, 2005), although this finding remains controversial (Kay, 2005).
2. We should note that there are other possible explanations of universals besides innateness, such as universal aspects of the environment or of human culture, both of which could lead to psychological universals that are not innate. What is important for the theory presented here, however, is the special properties of universal foci. Whether they are the product of cultural, environmental, or genetic processes is less important.
3. Some of this work was published under her earlier name, Heider.
4. We should note, however, that Dugum Dani also has nonbasic color terms, so it is possible that these color terms could to some extent have influenced Rosch's results.
5. Munsell chips are pieces of cardboard, each of which is painted in a specific color. The Munsell colors were chosen to cover the color space as comprehensively as possible, and an attempt was made to ensure that there was an equal perceptual distance between all neighboring chips (Cleland, 1937; Newhall, 1939, 1950; Newhall et al., 1943).
6. Berinmo has only five basic color terms, whose denotations correspond approximately to black, white, red, yellow, and green-blue.
7. Here I use the World Color Survey coordinate system, which is a simplification of the Munsell system. For Munsell–World Color Survey conversions see the World Color Survey at http://www.icsi.berkeley.edu/wcs/.

8. However, the discrepancy is not necessarily as great as suggested by these data, as Boynton and Olson calculated distances between the centers of areas of color named by English words, not between the universal foci.
9. These distances were obtained by measuring those given on a diagram in Indow and Ohsumi (1972). The diagram did not show any of the unique hues themselves, as none of these lie on a major axis of the Munsell system. Hence their locations were derived by interpolation from the closest hues shown on the diagram, and the distances were then scaled to World Color Survey hue units.
10. If it seems incoherent that a space can have a property locally that is not preserved over greater distances, consider the surface of the Earth. Locally it is flat, but globally it is round.
11. A *just noticeable difference* is the smallest change in a color that can be detected by a human observer. By counting the number of such changes between a pair of colors, we obtain a measure of their dissimilarity, but one that is quite different to the perceptual similarity judgments that were used in standardizing the Munsell color system.
12. These models are also known as iterated learning models (Brighton & Kirby, 2001).
13. Belpaeme (2002) did suggest that the split into light and dark colors seen in languages with only two color terms might be explainable in terms of his model, because this might be the easiest way to divide up the color space, but the model was not able to account for any other aspects of color term typology.
14. The reader should note that this in no way corresponds to the 40 hues of the World Color Survey coordinate system. Universal foci vary in both lightness and hue, so the model's color space covers a range of lightness values as well as the full range of hues.
15. In section 4 it was noted that the language a person speaks has some influence on how dissimilar he or she perceives colors to be. Therefore, it could be argued that these universal foci locations should be set differently for speakers of different languages. However, these locations are the universal foci locations before the process of acquisition has begun, and so presumably are the same for speakers of all languages. Ideally they would be adjusted slightly during the acquisition process to reflect relativistic effects of the acquired language, but such details are beyond the scope of this model.
16. One system contained both two purple terms and a lime term, and hence is counted twice, as it diverges from the trajectories in two different ways.
17. We should note that it included no languages with only two color terms, so it has missed at least one possible color term system. MacLaury (1997a) suggested that there are other types of color term systems, such as ones containing yellow-green terms, that are significantly underrepresented in the World Color Survey.

## Acknowledgments

This work was supported by an Economic and Social Research Council Postdoctoral Fellowship (award No. PTA–026–27–0760). An earlier version of this article was presented at the 25th Annual Meeting of the Cognitive Science Society.

I would like to thank Judy Kay who supervised the original research leading to this article; Tony Belpaeme for his help in clarifying some issues; and Jim Hurford, Terry Regier, and two anonymous reviewers for their helpful comments that have greatly improved this article, as well as everyone else who helped me with any stage of this research.

## References

Abramov, I. (1997). Physiological mechanisms of color vision. In C. L. Hardin & L. Maffi (Eds.), *Color categories in thought and language* (pp. 89–117). Cambridge, UK: Cambridge University Press.

Barr, D. J. (2004). Establishing conventional communication systems: Is common knowledge necessary? *Cognitive Science, 28,* 937–962.

Belpaeme, T. (2002). *Factors influencing the origins of color categories.* Unpublished doctoral thesis, Artificial Intelligence Lab, Vrije Universiteit Brussel, Brussels, Belgium.

Belpaeme, T., & Bleys, J. (2005). Explaining universal color categories through a constrained acquisition process. *Adaptive Behavior, 13,* 293–310.

Berlin, B., & Kay, P. (1969). *Basic color terms.* Berkeley: University of California Press.

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

Brighton, H., & Kirby, S. (2001). The survival of the smallest: Stability conditions for the cultural evolution of compositional language. In J. Kelemen & P. Sosík (Eds.), *Advances in artificial life* (pp. 592–601). Berlin: Springer.

CIE. (1995). *Industrial color-difference evaluation.* Vienna, Austria: Author.

Cleland, T. M. (1937). *A practical description of the Munsell color system with suggestions for its use.* Baltimore: Munsell Color.

Committee on Colorimetry, Optical Society of America. (1953). *The science of color.* New York: Crowell.

de Boer, B. (1999). Evolution and self-organization in vowel systems. *Evolution of Communication, 3,* 79–103.

Derrington, A. M., Krauskopf, J., & Lennie, P. (1984). Chromatic mechanisms in lateral geniculate nucleus of macaque. *Journal of Physiology, 357,* 241–265.

De Valois, K. K., & De Valois, R. L. (2001). Color vision. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (pp. 2252–2256). Amsterdam: Elsevier.

De Valois, R. L., Abramov, I., & Jacobs, G. H. (1966). Analysis of response patterns of LGN cells. *Journal of the Optical Society of America, 56,* 966–977.

De Valois, R. L., & De Valois, K. K. (1993). A multi-stage color model. *Vision Research, 33,* 1053–1065.

Everett, D. L. (2005). Cultural constraints on grammar and cognition in Pirahã: Another look at the design features of human language. *Current Anthropology, 46,* 621–646.

Foley, W. A. (1997). *Anthropological linguistics: An introduction.* Malden, MA: Blackwell.

Griffiths, T. L., & Tenenbaum, J. B. (2000). Teacakes, trains, taxicabs and toxins: A Bayesian account of predicting the future. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the twenty-second annual conference of the Cognitive Science Society* (pp. 202–207). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hardin, C. L. (1999). Color relations and the power of complexity: Commentary on Palmer (1999). *Behavioral and Brain Sciences, 22,* 953–954.

Harrison, K. D., Dras, M., & Kapicioglu, B. (2002). Agent-based modeling of the evolution of vowel harmony. In M. Hirotani (Ed.), *Proceedings of the Northeast Linguistic Society* (p. 32). Amherst, MA: GLSA.

Heider, E. R. (1971). "Focal" color areas and the development of color names. *Developmental Psychology, 4,* 447–455.

Heider, E. R. (1972). Universals of color naming and memory. *Journal of Experimental Psychology, 93,* 10–20.

Heider, E. R., & Olivier, D. C. (1972). The structure of the color space in naming and memory for two languages. *Cognitive Psychology, 3,* 337–354.

Hering, E. (1964). *Outlines of a theory of the light sense* (L. M. Hurvich & D. Jameson, Trans.). Cambridge, MA: Harvard University Press.

Hurford, J. R. (1987). *Language and number: The emergence of a cognitive system.* New York: Basil Blackwell.

Hurford, J. R. (2002). Expression/induction models of language evolution: Dimensions and issues. In T. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models* (pp. 301–344). Cambridge, UK: Cambridge University Press.

Indow, T. (1988). Multidimensional studies of Munsell color solid. *Psychological Review, 95,* 456–470.

Indow, T., & Ohsumi, K. (1972). Multidimensional mapping of sixty Munsell colors by non-metric procedure. In J. J. Vos, L. F. C. Friele, & P. L. Walraven (Eds.), *Color metrics: Proceedings of the Helmholts Memorial Symposium on Color Metrics*. Soesterberg, Netherlands: Institute for Perception.

Jameson, K., & D'Andrade, R. G. (1997). It's not really red, yellow, green, blue: An inquiry into perceptual color space. In C. L. Hardin & L. Maffi (Eds.), *Color categories in thought and language* (pp. 295–319). Cambridge, UK: Cambridge University Press.

Kay, P. (1975). Synchronic variability and diachronic change in basic color terms. *Journal of Language in Society, 4,* 257–270.

Kay, P. (1999). Asymmetries in the distribution of composite and derived basic color categories. Commentary on Palmer (1999). *Behavioral and Brain Sciences, 22,* 957–958.

Kay, P. (2005). Commentary on Everett (2005). *Current Anthropology*, *46,* 636–637.

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* (pp. 21–56). Cambridge, UK: Cambridge University Press.

Kay, P., Berlin B., & Merrifield, W. R. (1991). Biocultural implications of systems of color naming. *Journal of Linguistic Anthropology, 1,* 12–25.

Kay, P., & Maffi, L. (1999). Color appearance and the emergence and evolution of basic color lexicons. *American Anthropologist, 101,* 743–760.

Kay, P., & McDaniel, K. (1978). The linguistic significance of the meanings of basic color terms. *Language, 54,* 610–646.

Kay, P., & Regier, T. (2003). Resolving the question of color naming universals. *Proceedings of the National Academy of Sciences of the United States of America, 100,* 9085–9089.

Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. In E. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models* (pp. 173–204). Cambridge, UK: Cambridge University Press.

Kuehni, R. G. (1997). *Color: An introduction to practice and principles.* New York: Wiley.

Lammens, J. M. G. (1994). *A computational model of color perception and color naming.* Unpublished doctoral dissertation, State University of New York at Buffalo, Buffalo, NY.

Levinson, S. C. (2000). Yélî Dnye and the theory of basic color terms. *Journal of Linguistic Anthropology, 10,* 3–55.

Lucy, J. A., & Shweder, R. A. (1979). Whorf and his critics: Linguistic and nonlinguistic influences on color memory. *American Anthropologist, 81,* 581–615.

MacAdam, D. L. (1974). Uniform color scales. *Journal of the Optical Society of America, 64,* 1691–1702.

MacLaury, R. E. (1997a). *Color and cognition in mesoamerica: Construing categories as vantages.* Austin: University of Texas Press.

MacLaury, R. E. (1997b). Ethnographic evidence of unique hues and elemental colors. *Behavioral and Brain Sciences, 20,* 202–203.

Newhall, S. M. (1939). The ratio method in the review of the Munsell colors. *American Journal of Psychology, 52,* 394–405.

Newhall, S. M. (1950). A method of evaluating the spacing of visual scales. *American Journal of Psychology, 63,* 221–228.

Newhall, S. M., Nickerson, D., & Judd, D. B. (1943). Final report of the OSA subcommittee on the spacing of the Munsell colors. *Journal of the Optical Society of America, 33,* 385–418.

Palmer, S. E. (1999). Color, consciousness, and the isomorphism constraint. *Behavioral and Brain Sciences, 22,* 923–943.

Ratliff, F. (1976). On the psychophysiological bases of universal color terms. *Proceedings of the American Philosophical Society, 120,* 311–330.

Regier, T., Kay, P., & Cook, R. S. (2005a). Focal colors are universal after all. *Proceedings of the National Academy of Sciences, 102,* 8386–8391.

Regier, T., Kay, P., & Cook, R. S. (2005b). Universal foci and varying boundaries in linguistic color categories. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the XXVII annual conference of the Cognitive Science Society.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Roberson, D., Davies, I., & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a Stone-Age culture. *Journal of Experimental Psychology: General, 129,* 369–398.

Robertson, A. R. (1977). The CIE 1976 color-difference formulae. *Color Research and Application, 2,* 7–11.

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology, 4,* 328–350.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Saunders, B. A. C. (1992). *The invention of basic color terms*. Utrecht, Netherlands: ISOR.

Smith, A. D. M. (2003). Intelligent meaning creation in a clumpy world helps communication. *Artificial Life, 9,* 559–574.

Smith, K., Brighton, H., & Kirby, S. (2003). Complex systems in language evolution: The cultural emergence of compositional structure. *Advances in Complex Systems, 6,* 537–558.

Steels, L., & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for color. *Behavioral and Brain Sciences, 28,* 469–529.

Taylor, J. R. (1989). *Linguistic categorization: Prototypes in linguistic theory.* Oxford, UK: Oxford University Press.

Tenenbaum, J. B. (1999). *A Bayesian framework for concept learning*. Unpublished doctoral thesis, MIT, Cambridge, MA.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioural and Brain Sciences, 24,* 629–640.

Tenenbaum, J. B., & Xu, F. (2000). Word learning as Bayesian inference. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the twenty-second Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Vogt, P., & Coumans, H. (2003). Investigating social interaction strategies for bootstrapping lexicon development. *Journal of Artificial Societies and Social Simulation, 6*(1).

Webster, M. A., Miyahara, E., Malkoc, G., & Raker, V. E. (2000). Variations in normal color vision: II. Unique hues. *Journal of the Optical Society of America A, 17,* 1545–1555.

Yendrikhovskij, S. N. (2001). Computing color categories from statistics of natural images. *Journal of Imaging Science and Technology, 45,* 409–417.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8,* 338–353.