Explaining Color Term Typology as the Product of Cultural Evolution using a Bayesian Multi-agent Model

Mike Dowman (Mike@it.usyd.edu.au)
School of Information Technologies, F09, University of Sydney,
NSW2006 AUSTRALIA

Abstract

An expression-induction model was used to simulate the evolution of basic color terms in order 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 universal aspects of human neurophysiology. Ten agents were simulated, each of which could learn color term denotations by generalizing from examples using Bayesian inference. 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.

Introduction

This paper describes computational modeling experiments performed to explain the typological patterns observed in the color term systems of human languages. Color terms are simply words which are used to denote the property of color, and in most languages, a special subset of such words can be identified, which Berlin and Kay (1969) named *basic color terms*. Berlin and Kay listed a number of criteria which they used to distinguish basic color terms from other words used to denote color. They considered color terms to be basic only if they were known by all speakers of the language and were highly salient psychologically, and if they did not just name a subset of the colors denoted by another color word and their meanings were not predictable from the meanings of their component parts. Application of these criteria seems to distinguish clearly between basic and non-basic color terms in most languages, although there can still remain some questionable cases. English has 11 basic color words, *red*, *yellow*, *green*, *blue*, *orange*, *purple*, *pink*, *brown*, *grey*, *black* and *white*. Terms such as *crimson*, *blonde* and *royal blue* are not considered to be basic.

Basic color terms have prototype properties (Taylor, 1989) as there is usually a single color, the *prototype*, which speakers of the language consider to be the best example of the color term. Colors become increasingly less good examples of the color category as they become more dissimilar to the prototype, and the category boundaries are fuzzy, as speakers are unsure about the exact range of colors denoted by each color term.

There has been a considerable amount of research into the properties of basic color terms cross-linguistically. Perhaps the most important study was that of Berlin and Kay (1969). They examined a sample of 98 languages, and found that there was very wide variation between the color terms of different languages, in that the actual ranges of color denoted by each term differed between languages. However, they found that this variation was certainly not completely random. While the number of color terms varied between languages, which combination of color terms existed in any given language seemed to be at least partly predictable.

Berlin and Kay found that all languages have between 2 and 11 basic color terms. For 20 of the languages in their study, they asked informants to map both the outer boundary of each of the basic color terms on an array of Munsell color chips, and to identify the best or most typical example (the prototype) of each term. They discovered that the boundaries of the areas of color denoted by color terms varied greatly between languages, but that the locations of the prototypes of most basic color terms were clustered in a few parts of the color space.

A further finding emerged when Berlin and Kay investigated the combination of color terms existing in any particular language. They produced the implicational hierarchy shown in Figure 1 to explain the regularities which they found. All languages appeared to have terms with their prototypes at black and white, but some languages had no other basic color terms. However, if a language had a term for any of the colors further right in the hierarchy, it always had terms for all the colors appearing to the left of that point.

Berlin and Kay proposed that this hierarchy described the general patterns seen in color term systems cross-linguistically, but they did acknowledge the existence of

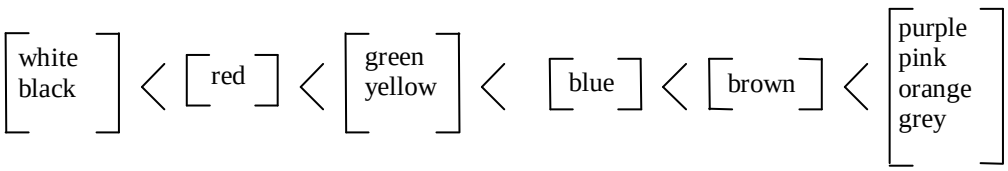

Figure 1. Berlin and Kay's Implicational Hierarchy.

some exceptional languages which could not be placed on the hierarchy. Since Berlin and Kay published their original study, there has been a great deal of interest in basic color terms, and much more data has been collected. These studies have in large part confirmed Berlin and Kay's original findings, though several modifications have been made to the hierarchy to accommodate languages of types which were not attested in the original study.

A very large survey of the color term systems of 110 minor languages, the World Color Survey (Kay, Berlin, and Merrifield, 1991), has now produced a wealth of high quality data allowing a much more complete picture of color term systems worldwide to be obtained. Using this new data, Kay and Maffi (1999) produced a new classification of color term systems, which has modified the original hierarchy of Berlin and Kay (1969) considerably, but which still shows that the attested color term systems are only a small subset of those which are logically possible.

There appear to be six fundamental colors, corresponding to the colors which are usually the prototypes of red, yellow, green, blue, black and white color terms. The order of emergence of terms containing these fundamental colors is fairly predictable, but the order of emergence of other basic color terms, such as purple and orange terms, is less predictable. Kay and Maffi's (1999) classification of color term systems was made by considering only terms whose denotation included at least one of the fundamental colors, but Kay, Berlin and Merrifield (1991) noted that purple and brown terms tend to emerge before orange or pink ones.

Kay and Maffi (1999) found that 83% of the languages in the World Color Survey lie somewhere along the trajectory shown in Figure 2, which represents a progression in which languages evolve from a state in which they contain only two basic color terms, to states in which each of the fundamental colors is represented by a different basic color term. Kay and Maffi also proposed side branches to the main trajectory in order to accommodate some less common language types, such as those containing yellow-green-blue, yellow-green or black-blue composites. There were four languages which Kay and Maffi were unable to place anywhere on their trajectories, and which were simply classed as exceptions. This paper attempts to address the issue of what causes these typological patterns, by relating them to fundamental properties of the human visual system.

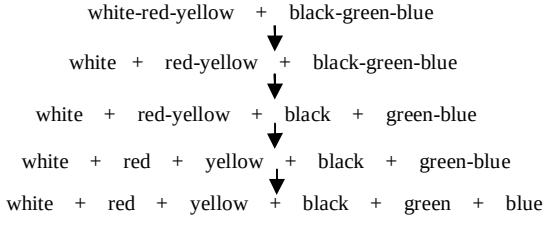

Figure 2. The Main Line of the Evolutionary Trajectory.

De Valois and Jacobs (1968) showed that Macaque monkeys had four kinds of opponent cells, each of which responded maximally in the presence of light of a particular red, yellow, green or blue hue. Kay and McDaniel (1978) proposed that the output of the opponent cells would map directly to fuzzy set membership in color term categories. However, this proposal appears to be too restrictive, because it implies a limited set of universal color categories, while the empirical evidence shows that the boundaries of color term denotations are quite variable, and that it is only the prototypes of color terms that are consistent across languages.

Four special red, yellow, green and blue hues have also been identified by psychological evidence, and they are termed unique hues, because they can't be described as blends of other colors. Heider (1971) showed that children are more likely to pick unique hues rather than other colors out of selections of color chips, and so she proposed that unique hues are especially salient. In Heider (1972) she showed that people are best able to pick out a previously seen color from an array of color chips when that color is a unique hue, which suggested that unique hues can be remembered more easily than other colors. We should note that Heider did not distinguish between the four unique hues and the colors which formed the prototypes of color terms such as *purple* and *orange*, but a large amount of other evidence points to the special status of unique hues. Also Kay and Maffi (1999) have questioned whether the unique hues apparent from the linguistic and psychological studies really correspond to those found using neurophysiological techniques. However, the consensus of opinion seems to be that the four unique hues have a special psychological status, regardless of its cause.

The computer model described in this paper is a kind of expression-induction model. These models, the first of which was described in Hurford (1987), aim to simulate cultural evolution of language, usually over a number of generations. They contain several agents, each of which is capable both of learning some aspect of language, and also using the language which they have learned to express themselves, hence creating example utterances from which other agents can learn. Usually expression-induction models are run several times, so that the general properties of the languages which emerge in them can be observed. If all the emergent languages have a particular property which is also a universal in real languages, or if the emergent languages show a limited range of variation, reflecting typological patterns, then the model can be said to explain why these universals or typological patterns exist. Expression-induction models have been developed to account for a wide range of linguistic phenomenon. Belpaeme (2002) used such a model to simulate color term evolution, but his model did not account for typological patterns.

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, but it learns in a similar way to a concept learning program developed by Tenenbaum (1999). In order to create the acquisitional model it was necessary to make a number of assumptions about how children learn color words. It was assumed that, before people begin to learn the meanings of color terms, they must have some sort of conceptual color space. We experience color as having a three-dimensional structure, where the three dimensions are *hue*, *saturation* and *lightness*. The Bayesian model is at present concerned

only with the dimension of *hue*, which is a circular dimension, as shown in Figure 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* or *white*, but the acquisition of red, orange, yellow, green, blue and purple terms can still be modeled.

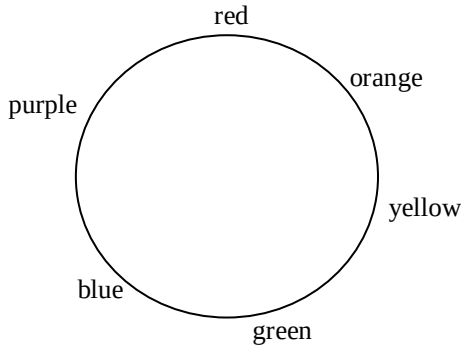

Figure 3. The Conceptual Color Space.

Another assumption is that the unique hues are not evenly spaced in the color space. The color space was divided into 40 discrete colors, and so each individual color could be indexed with a number between 1 and 40. Using this coordinate system, the red unique hue point was placed at hue 7, yellow at 19, green at 26, and blue at 30, so that the largest distance between adjacent unique hue points is 17 units between blue and red, and the smallest is just 4 units between green and blue.

The motivation for placing the unique hues at these points is primarily that it results in an explanation of the typological patterns, as will be shown below. MacLaury (1997) reported that there is some evidence to suggest that the green and blue unique hues are in some way closer than any of the other hues are to each other, but there is no clear objective way to measure distances in the conceptual color space, and so we could obtain different conceptual distances depending on what method was used to measure them.

The special salience and memorability of the unique hues was simulated by associating with each color a probability corresponding to how likely a person was to remember an example of it. These values will be written R_c, and were set at 0.05 for colors which did not correspond to unique hue points, and at 1 for unique hues, so that it was 20 times as likely that examples of unique hues would be remembered as examples of other colors.

The next issue to be considered is what data is 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 learns consists of examples of colors which a color term can denote. Learning then involves generalizing from those examples to the full range of colors which come within the word's denotation. Because it is possible that some examples could be erroneous, a parameter, *p*, was added to the model, which corresponds to a learner's belief concerning the probability that each individual example is correct.

The model learns using Bayes' rule, given in (1). Each word is learned separately from all the others, and so the data, *d*, will consist of all the observed example colors for one color word, and the hypothesis, *h*, will correspond to the range of color which the word denotes. Hypotheses can vary in size from taking up one unit of the color space, to including the whole of the color space, and can correspond to any contiguous range of colors. It is proposed that children will consider all such hypotheses to be equally likely *a priori*, so that the model has no inbuilt bias to prefer color terms corresponding to one part of the color space as opposed to any another, and so the term $P(h)$ will have the same value for all hypotheses.

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

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's equally likely for that example to have been observed on any of the colors with the range of the hypothesis[1], assuming that the hypothesis is correct. However, because some examples will be forgotten, the proportion of examples which we would expect to have remembered for each particular color would be equal to the probability of remembering examples of that color, divided by the sum of the probabilities of remembering examples on all the colors within the range of the hypothesis, which will be written as R_h. This ratio, which is given in (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.

$$(2)\quad \frac{R_c}{R_h}$$

Erroneous examples are 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 (3).

$$(3)\quad \frac{R_c}{R_t}$$

(2) and (3) apply when we know whether an example is accurate or not, but in reality, when a person has remembered an example they won't 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

[1] This assumes that people are equally likely to observe examples of colors anywhere in the color space.

that it is not accurate is *(1-p)*. Hence the overall probability of an example, *e*, which comes outside of the range of a hypothesis, is given by multiplying (3) by this value, as shown in (4).

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

If an example is within the scope of the hypothesis then we can't 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 (5).

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

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 (6). For every example we must use either equation (4) or (5) to calculate *P(e | h)*, depending on whether or not the example is within the scope of the hypothesis.

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

In order 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 (7).

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

If we substitute (7) into Bayes' 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)*.)

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

Equation (8) lets us calculate the probability of individual hypotheses corresponding to possible denotations of the color word. However, if we use the standard Bayesian procedure of hypothesis averaging, it is possible to calculate how likely it is that any particular color can be denoted by the color word. We can do this by calculating the *a posteriori* probability of each hypothesis which includes the color within its range, and adding together all these probabilities. This will determine the overall probability that the color comes within the denotation of the color word. If such values are calculated for all colors, then we can derive a fuzzy set by equating the probabilities with degree of membership in the set.

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, and the learned color term system is shown in Figure 4. Example colors were generated based on the denotations of Urdu color terms shown on a chart in Berlin and Kay (1969), and these examples were passed to the model until it had remembered 40 of them. Each of the color terms which contains a unique hue point has that point as its prototype, and the degree of membership declines gradually moving away from this point, which is consistent with empirical findings. This shows that the model is able to learn the color term system of a real language, but it does not explain the typological data, because color term systems of unattested types could also be learned by the model.

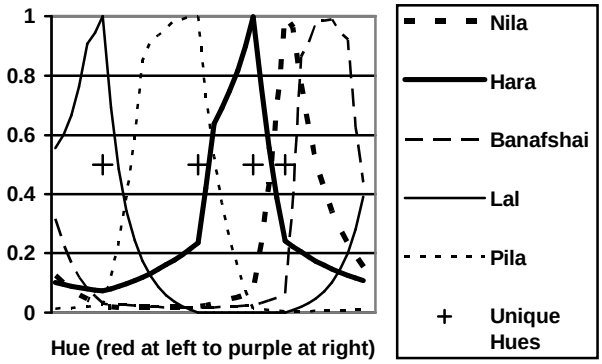

Figure 4. Learned Denotations for Urdu Color Terms. (Y-axis shows degree of membership in color category.)

Simulating Color Term Evolution

In order to simulate a whole community of people, ten copies of the acquisitional model were created, each of which acted as an agent. Conversations between the agents were simulated over several generations, 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).

In the initial state of the model, each agent had observed a different color term together with one completely random example of it, and each agent was assigned a random age between zero and the maximum age which agents could live to. The simulations proceeded by first selecting one agent at random to speak, and another to hear. A color for the speaker to name would then be chosen. The color would be randomly selected, but each unique hue was chosen 20 times more often than each of the other colors. The speaker would then find the word which they thought most likely to be the correct label for the color, based on all the color examples which 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. One time in every thousand, instead of the speaker choosing the best word based on the observations they had made, it would be creative instead, and make up a new word.

A parameter in the model controlled how long each agent lived for, measured in terms of how many times an agent would speak during its lifetime. 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 which 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.)

These simulations were repeated 425 times, with the average lifespan set variously at 18, 20, 22, 24, 25, 27, 30, 35, 40, 50, 60, 70, 80, 90, 100, 110 or 120, with 25 separate simulations being made in each condition. The simulations were all run for a time equal to twenty average life-spans, and the results reported below are based on the languages spoken by the agents at the end of the simulations.

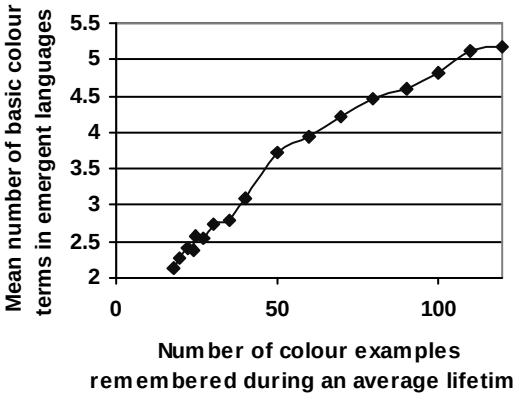

Figure 5. Number of Color Terms in Emergent Languages.

In order to analyze the results it was necessary to develop some procedures for classifying the languages emerging in the simulations. Throughout all the analyses, only agents whose age was greater than or equal to half the average lifespan were included, and only color terms for which they had observed at least 4 examples were considered. Using these criteria, the mean number of color words spoken by agents for each setting of the life expectancy parameter, was calculated, and these means are plotted in Figure 5. There is a clear positive correlation between how often agents used color terms and the number of color terms in their language. This effect is not because more color terms allowed the agents to communicate more accurately, as the agents were not rewarded for successful communication. This suggests that languages with large numbers of color terms may have so many color terms simply because the people who speak those languages frequently talk about color, rather than because of any functional benefit arising from a larger color vocabulary.

In order to investigate whether the typological patterns identified by Kay and Maffi (1999) were replicated in the simulations, it was necessary to classify each language in terms of which kinds of basic color terms it contained. For every hue, the color term which the agent would use to name that hue was found, and the denotation of each color term was then considered to be the smallest range of colors which included all those hues which that term would name. (This could potentially include some hues which would be named by a different color term, because there might be another term which 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 unique hue. Terms whose denotations didn't include any unique hue points were classified as *orange*, *lime*, *turquoise* or *purple*, depending on which unique hue points their denotations came between. If a color term included more than one unique hue point it would be classified as a composite of those unique hue points, 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, like with real color term systems (MacLaury, 1997), the agents did not agree exactly on the denotation of each word, and nor did they necessarily use identical sets of color words. A color term was included in the analysis only if at least half the agents had observed at least four examples of it. If all the agents did not agree on the classification of a term, then that classification which was supported by the greatest number of people would be chosen. If two or more possible classifications were supported by equal numbers of people, then if one of the terms contained fewer unique hue points 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 proportion of terms which were classified as being of each type in all the emergent languages is shown in Figure 6. Figure 6 also contains equivalent data from the world color survey (WCS), as reported in Kay and Maffi (1999). Kay and Maffi did not take account of color terms which did not contain a unique hue point in making their classifications, hence the relevant data on these terms does not appear here, and the counts of color terms for the world color survey do not include terms from languages which were in transition between evolutionary stages, or which did not fit on the evolutionary trajectories at all. Color terms which are either achromatic or distinguished from other colors on the basis of some dimension other than hue have simply been excluded from the analysis, while composite terms which included, white or black and one or more unique hues were treated as if they only contained the unique hue.

Figure 6 clearly shows 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. There were 80 purple terms in emergent languages, but only 20 orange ones, which is consistent with the finding that purple terms tend to emerge before orange ones. 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 which have not been attested empirically. There was 1 Blue-Red composite, 1 Red-Yellow-Green composite, 3 Green-Blue-Red composites and 4 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 which can evolve, but simply introduces biases, so that some kinds of color term 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 which were not found in Berlin and Kay's original survey have been discovered, but it is possible that some very rare types of color term remain undiscovered.

Looking at the languages overall, 340 could be placed on Kay and Maffi's (1999) evolutionary trajectories. 9 languages deviated from the trajectories because they contained unattested color terms, 35 because no term consistently named one or more of the unique hues, and 37 because there was more than one term which could name one of the unique hues, or there was more than one purple term. What is clear from these results is that there is a small set of color term systems which occur very frequently, and that the color term systems of the majority of languages can be classified as belonging to one of these types. A significant number of languages diverge from the trajectories in some way or another, but this is consistent with empirical findings.

Conclusion

The computer model has shown that the typological patterns observed in basic color term systems can be explained if it is assumed that the unique hues are not evenly spaced in the conceptual color space and that people remember the unique hues better than other colors. These assumptions produced learning biases which affected the way that languages evolved. The language which each agent in the simulation learned was a product both of the agent's learning mechanism, and of the language spoken by the other agents, and this suggests that human languages can be understood only as a product of both innate biases and cultural pressures interacting over a period of several generations.

Acknowledgments

I would like to thank my Ph.D. supervisor Judy Kay and everyone else who has helped me with this research which was supported by IPRS and IPA scholarships.

References

Belpaeme, Tony (2002). Factors influencing the origins of color categories. Ph.D. Thesis, Vrije Universiteit Brussel.

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

De Valois, R. L. & Jacobs, G. H. (1968). Primate color vision. *Science, 162*, 533-540.

Heider, E. R. (1971). "Focal" Color Areas and the Development of Color Names. *Developmental Psychology, 4 (3),* 447-445.

Heider, E . R. (1972). Universals of Color Naming and Memory. *Journal of Experimental Psychology, 93*, 10-20.

Hurford, J. R. (1987). *Language and Number The Emergence of a Cognitive System.* New York: Blackwell.

Kay, P., Berlin, B. & Merrifield, W. (1991). Biocultural Implications of Systems of Color Naming. *Journal of Linguistic Anthropology, 1 (1),* 12-25.

Kay, P. & McDaniel, K. (1978). The Linguistic Significance of the Meanings of Basic Color Terms. *Language, 54 (3),* 610-646.

Kay, P. & Maffi, L. (1999). Color Appearance and the Emergence and Evolution of Basic Color Lexicons. *American Anthropologist, 101*, 743-760.

MacLaury, R. E. (1997). *Color and Cognition in Mesoamerica: Construing Categories as Vantages.* Austin, Texas: University of Texas Press.

Taylor, J. R. (1989). *Linguistic Categorization: Prototypes in Linguistic Theory*. Oxford University Press.

Tenenbaum, J. B. (1999). *A Bayesian Framework for Concept Learning*. Ph.D. Thesis, MIT.

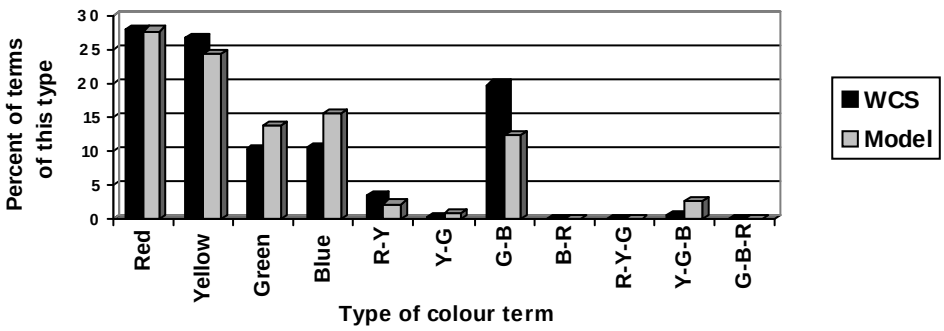

Figure 6. Percentage of Color Terms of each type in the Simulations and the World Color Survey.