

# Modelling the Acquisition of Colour Words

## A Bayesian Approach to Fuzzy Sets

Mike Dowman

Mike@it.usyd.edu.au

School of Information Technologies,  
Madsen Building F09,  
University of Sydney,  
NSW2006, Australia

### Abstract

How Bayesian inference might be used as the basis of a system for learning and representing the meanings of colour words in natural languages was investigated. The paper is primarily concerned with cognitive modelling, but has potential applications in natural language processing. A Bayesian cognitive model was constructed to test the hypothesis that people learn language, and in particular the meanings of colour words, using Bayesian inference. The model learned the range of colours which could be named by a particular colour word from examples of colours which could be denoted by that word, and was able to do so accurately even in the presence of large quantities of random noise in the input data. The resulting meaning representations display many of the properties of colour words in natural languages, in particular prototype properties. This suggests that the Bayesian approach closely models the mechanism which people use to learn colour words, and that this approach may be useful in constructing language technology systems which require accurate representations of the words of natural languages.

### 1. Introduction

This paper describes a computational model which aims to account for how people learn the meanings of colour words. It takes as its input data examples of colours which can be named by a colour word, and then uses Bayesian inference to generalize from those examples to determine the full range of colours which that word can identify. The task of learning colour words is far from straightforward, because they differ substantially between languages, and they have complex prototype structures, with some colours being better examples of the word than others. Despite these factors, children appear to have little difficulty in learning colour words, despite the limited and noisy data from which they must learn.

All languages have many words that may be used to identify particular colours, but there is always a special subset of these words known as *basic colour terms*. Each of these words identifies a range of colour which is not included in the range of colour identified by any other colour word, and the words themselves usually divide up the colour space so that there is one colour term which can identify every colour. Furthermore, these words are psychologically salient, in that they come readily to mind when speakers of a language are asked to name colour words, and each basic colour term in a particular language will be known to every speaker of that language. These criteria were specified by Berlin and Kay (1969), who, by examining descriptions of a wide range of languages, determined that all languages had between two and eleven such words, a finding that remains largely unchallenged today.

The concept of basic colour term can probably be made clearer with some examples. English has eleven basic colour terms, 'black', 'white', 'red', 'orange', 'yellow', 'green', 'blue', 'purple', 'pink', 'brown' and 'gray', but it also has many other colour words which are not basic, such as 'crimson', 'turquoise' and 'beige'. In contrast Danian Lani, a language from Irian Jaya in Indonesia, has only two basic terms, one of which, 'laambu', names roughly white, red, yellow and very light colours, and the other, 'mili', names black, green blue and very dark colours (MacLaury, 1997). From these examples we can see that there is great variation between the colours which colour words in different languages can be used to identify, and so children must learn the range of colours which can be referred to by each colour word in their own language. This is the task for which the computational model described in this paper aims to account, but there are other properties of colour words for which the model must account for as well.

Taylor (1989) notes that basic colour terms, in common with many other kinds of word, have prototype properties. There is usually a single colour which is the best example of the category identified by the colour term, and this is called the

*prototype*. Colours gradually become less good members of the category the further they are away from this prototype, until at categories' fuzzy boundaries it becomes unclear exactly which colours are members of the category and which are not. While there is usually good agreement between speakers as to the location of the prototype, speakers tend to disagree about exactly where the category boundaries are. Prototype theory (Taylor, 1989) suggests that colour words have these properties because the prototype plays a central role in defining their meaning, but, in the model described here, prototype properties are an emergent property of the Bayesian approach to learning.

In order to construct the model, it was first necessary to consider exactly what sources of information children would have available from which to learn. There is a large body of evidence which suggests that the principal way in which children learn their languages is through observing the speech of other people, and inferring the underlying rules of grammar and meanings of words (Bloom, 2000). Explicit teaching and formal education seem to play only a relatively marginal role in language acquisition, at least for central aspects of language. In the case of learning the meaning of a colour word, it would seem that children must observe which particular colours that word was used to identify, and from a number of such examples generalize to determine the full range of colours which the word can be used to name.

The way in which children experience colour appears to affect how colour categories are formed. Physically colour is determined by the wavelength of light, with light of long wavelength appearing red, with orange, yellow, green, blue and purple having increasingly shorter wavelengths. However, perceptually colour has a three-dimensional structure, as colour can vary on any of the three dimensions of *hue*, *saturation*, or *lightness*. Saturation concerns the degree of dilution of a colour by white or black, so colours with low degrees of saturation appear greyish, while lightness is concerned with how light or dark a colour is. The hue dimension corresponds to the colours of the spectrum, but where the ends are joined together to form a circular dimension with purple next to red. The existence of this three-dimensional colour structure is well supported by evidence from psychological experiments (Thompson, 1995), and it appears to be the colour structure in which colour categories are defined. Hence it was assumed that the example colours from which children learn colour words are represented in terms of this structure. While many alternative colour spaces have been proposed (see for example Boynton and Olson, 1987 and Hård and Sivik, 1981), there appears to be consensus that people conceptualise colour using a conceptual space based on these three dimensions. The various alternative colour spaces proposed to account for human colour conceptualisation differ primarily in terms of differences in the exact shape of the space, or in terms of the conceptual distances between, and exact locations of, particular colours, but these differences are not relevant to the cognitive model presented in this paper.

The task of determining exactly what object a word has been used to refer to through observing the speech of other people would seem to be a complex and difficult process, with much potential for error. Hence it would seem that on some occasions children might come to believe that a word had been used to name one colour, when in fact it had been used to identify a different colour. (This could happen for many reasons, such as the child misidentifying the object which was being described, or focussing on the wrong colour of a multi-coloured object.) In order to prevent occasional occurrences of this sort from having a catastrophic effect on the outcome of learning, the acquisitional model must not treat the training examples as being absolutely correct, but must instead allow some possibility that each one is erroneous.

In this paper, I propose that the mechanism which children use to learn the meanings of colour words from such examples is a form of Bayes' optimal classification (Mitchell, 1997). The model calculates the probability that the meaning of a colour word corresponds to a particular range of colours, making use of all the examples of the use of that word which have been observed. By using the standard Bayesian procedure of hypothesis averaging over all such possible sections of the colour space, it is then possible to determine the overall probability that each particular part of the colour space could be named by that word. Interestingly this produces the prototype properties characteristic of colour words, because it is most that certain colours near to the centre of the colour category can be named by the word, but it is not clear exactly where the category's boundaries are. The model is described in more detail in section 2. , but first it is necessary to provide some justification for the Bayesian approach.

The primary motivation for the Bayesian approach was simply that there already exist a number of studies which have shown that Bayesian models can accurately account for data concerning human learning in other situations. Griffiths and Tenenbaum (2000) demonstrated how Bayesian inference could be applied to predicting the frequency of periodic events, such as the frequency of trains on a subway line, based on examples of how long has elapsed between a randomly chosen time and the occurrence of the event. They showed that people's judgments as to the frequency of the trains were very similar to the predictions made by the Bayesian model, and so it seems likely that people were using Bayesian inference to make those predictions.

Tenenbaum and Xu (2000) applied Bayesian inference to a somewhat different problem – that of how to predict the full range of possible referents of a word based on examples. If a child observes a word for the first time, and it is used to name a dog, then the child might believe that the meaning of that word is 'dog'. However, the word could equally well

mean 'animal', or could even be a name referring to just that particular dog. Tenenbaum and Xu provided a Bayesian model which would predict the full range of referents of a word based on one or several examples. Psychological experiments confirmed that people tended to generalize in similar ways, to the Bayesian model, choosing a more restricted denotation when they had seen several similar examples, and a wider range of possible referents when they had seen fewer or more diverse examples. This provides evidence that people use Bayesian inference to learn concrete nouns, and if people learn concrete nouns in this way it would seem likely that they also use similar mechanisms to learn the meanings of other words, including colour words.

Before moving on to the details of the Bayesian model, I will discuss some problems with previous approaches to colour term semantics. I noted above that colour words have prototype properties, and because of this it is usually assumed that they have an underlying prototype representation. Taylor (1989) discusses this possibility at length, and suggests that the prototype plays a central role in defining the full range of colour categories. However, this seems problematic for two reasons. Firstly, as is apparent from data collected by Berlin and Kay (1969), and many subsequent researchers, the size of colour categories varies between languages. So, simply knowing the location of a category's focus is not enough to determine how far beyond this focus the category extends. However, perhaps a more fundamental problem with this approach is that categories vary not only in size but also in shape, sometimes extending further from the prototype in one direction through the colour space than in another.

Lammens (1994) implemented a computational model of colour term semantics which defined categories using prototype representations. The representation of the meaning of each colour term consisted of a specification of the location of the prototype and a parameter controlling the size of the category. Lammens made a proposal as to how a learner could determine the appropriate setting for the size parameter, but he presumed that the category prototypes must be predetermined innately. This latter assumption seems untenable, because while Berlin and Kay (1969) did find that the foci of categories in different languages were often the same, this is certainly not true for all categories in all languages. Hence it would seem that Lammens' model, at least in its present form, does not provide a plausible account of how colour categories could be learned.

MacLaury (1987) developed *vantage theory* to account for how people represent the meanings of colour words, and to explain why colour categories have the attested prototype properties. This theory is based on the concept of *vantages*, and proposes that people construct categories by contrasting the similarity and distinctiveness of points in the colour space. MacLaury attempts to account for all the available cross-linguistic data on colour words, but pays less attention to the question of how children learn colour categories. In particular it is unclear how categories which don't have a universal focus would be learned. At present there is no computational implementation of vantage theory, and I do not think that at present the theory is specified with enough precision to allow such an implementation, so no direct comparison is attempted here.

Honkela (1997) used self-organizing maps (a kind of neural network) to group colour words (including non-basic colour words) into classes with similar meanings. He also showed that self-organizing maps can model words which, like colour words, have graded membership, with some examples being members to a greater degree than others. However, Honkela did not demonstrate that the meanings of individual colour words could be learned from a series of examples, although it would appear that self-organizing maps may have the potential to do this. Furthermore, Honkela based the representation of colour on its red, green and blue components, which does not correspond to the kind of representation of colour on which linguistic colour categories are based. For these reasons, no direct comparison between the Bayesian approach of this paper and the approach of Honkela is possible, although self-organizing maps may well prove to be a fruitful avenue for further research on colour categorization.

Kay and McDaniel (1978) attempt to derive the meanings of colour words from the neural response functions of opponent process cells in the retina of the eye. These cells transform signals coming from light sensitive cells into signals which indicate the extent to which the light striking that part of the retina is green as opposed to red, or blue as opposed to yellow. Kay and McDaniel hypothesized that the values of these response functions would correspond to the degree of membership of colours in colour categories. In order to explain how colour categories vary between languages, they proposed that fuzzy logic could be used to combine two response functions, either fuzzy intersection to produce smaller categories, or fuzzy union to produce larger ones. However, the range of colour categories that can be produced in this way is still very limited, and does not appear to include the full range of colour categories attested throughout the world's languages. Therefore it would seem that this approach is much too inflexible to account for colour term acquisition, and so an approach which involves a much greater degree of learning is required. The Bayesian model presented here is similar to Kay and McDaniel's theory in that it also uses fuzzy sets to represent colour terms, but it is much more flexible because it derives those fuzzy sets by generalizing from observations rather than by applying the operations of fuzzy logic.

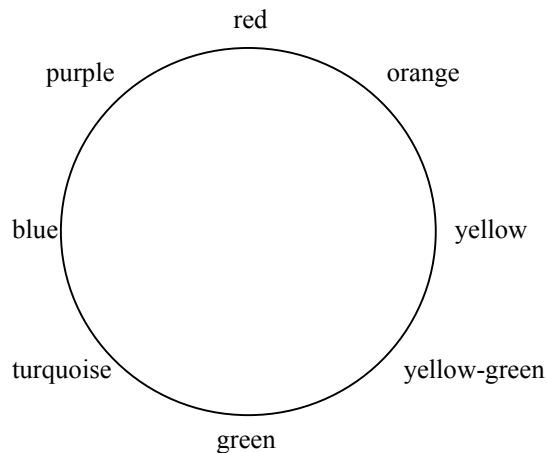
## 2. The Bayesian Model

The Bayesian model learns by making use of Bayes' rule, given in equation (1). This allows the probability of a hypothesis,  $h$ , to be determined based on some observed data,  $d$ . The hypotheses will correspond to possible meanings of colour words, and the data to examples of colours identified by them. However, before this equation can be applied, it is first necessary to determine exactly what form the colour examples take, and to specify the form of the hypotheses. This is described in the next two subsections, and then it is shown how the examples and hypotheses can be used to determine the probability that each colour comes within the meaning of each colour term, and so finally how fuzzy sets may be derived.

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

### 2.1 The Conceptual Colour Space

As described above, conceptually colour has a three dimensional structure, and it seems that, in order for people to be able to perceive and think about colours, they must have some sort of corresponding psychological conceptual space. (The idea of conceptual spaces is discussed at length by Gärdenfors (2000).) However, for reasons of simplicity, the present model is concerned only with the dimension of hue. If we consider only colours of maximum saturation and of the degree of lightness at which maximum saturation may be achieved, then the colours will form a one dimensional colour space as shown in Figure 1. It is assumed that prior to language learning such a colour space is available to children, and so they will know (at least unconsciously) that, for example, orange is more similar to yellow than it is to green. As this colour space does not include colours of zero saturation, it excludes examples of *black*, *white*, and *grey*. *pink* is excluded because it is a light version of *red*, and *brown* is excluded because it is essentially a dark and low saturation counterpart to some orange, red and yellow hues. This colour space is, however, sufficient to allow an account of the meanings of *red*, *orange*, *yellow*, *green*, *blue* and *purple*, and of the meanings of many other terms in other languages.



**Figure 1. The Phenomenological Colour Space**

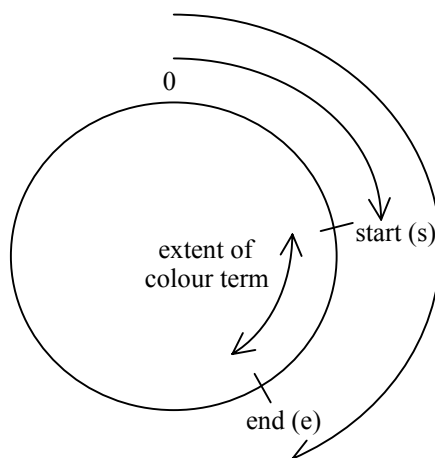
When an example of a colour which has been identified using a colour term is observed, it will be represented as a point in this colour space. For the purpose of identifying locations in the colour space an arbitrary scale from 0 to 100 will be used. This scale has its origin (location 0) at red, and increases as we move clockwise through the colour space. Location 100 will correspond to the end of the colour space, which is also the beginning, and so will be in the same place as location 0. Hence the data from which the Bayesian model learns will always consist of a list of (possibly fractional) numbers between 0 and 100, each of which corresponds to one particular example.

### 2.2 Hypotheses about Colour Word Meanings

The next step in the construction of the Bayesian model is to specify the possible hypotheses, and to assign a probability to each. Each hypothesis will define a continuous section of the colour space, and will correspond to the belief that all and only those colours within that section can be named by the colour word. This restricts the range of possible

hypotheses to only those which correspond to words denoting continuous ranges of colour. This is consistent with typological evidence which appears to show that all colour words do in fact have this property, although that does not necessarily entail that learners will know that before they start learning. However, Gärdenfors (2000) has suggested that it is a general property of concepts used by humans that they do not denote disjoint sections of conceptual spaces, so a word which denoted, for example, green and red hues, but not yellow or blue, ones would be impossible, or at least highly unlikely. Hence this restriction on possible hypotheses seems reasonable, because even if it is not entirely accurate, it should at least lead to a good approximation to implicit assumptions made by people when they begin to learn colour word denotations.

As is shown in Figure 2, each hypothesis will have a start point,  $s$ , and an end point,  $e$ , and will include all those colours which come after the start point but before the end point. The size of the section of colour space corresponding to a hypothesis is thus given by  $(e-s)$ . Where a hypothesis encompasses the origin, then 100 (the size of the colour space) must be added to  $e$ , as otherwise the value of  $e$  would be less than  $s$ , resulting in a negative value for the size of the hypothesis. A hypothesis may begin and end anywhere within the colour space, and all possible start and end points are considered equally likely *a priori*. This entails that there will be a continuous space of hypotheses, with the sizes of those hypotheses ranging from not including any of the colour space at all to including the whole of the colour space. It also follows that all sizes of hypothesis will be equally likely *a priori*.



**Figure 2. A Hypothesis as to the Denotation of a Colour Term**

### 2.3 Calculating the Probability of Observed Examples

Given a hypotheses and its *a priori*, it is possible to determine how likely it is that we would have observed the colour examples if that hypothesis was correct. Firstly it is assumed that children make no *a priori* assumptions that some colours are named by colour words more often than other colours are, and that a colour word is equally likely to be used to identify colours anywhere with the range of colours corresponding to its meaning.

Given these assumptions, we can calculate how likely we would be to have observed each individual example. As the number of examples does not vary between hypotheses we need not be concerned with considering how likely it is that we would have observed the actual number of examples which we did, but instead we can concern ourselves simply with calculating the probability that an example was observed at a particular point in the colour space. In order to calculate such values, we must divide the colour space into a finite number of sections, so that there is a non-zero probability of an example being observed in each section. We will use  $q$  to represent the number of such sections into which the colour space is divided.

If we can be sure that all the examples are accurate, then there is an equally likely probability of observing an example in any of the sections of the colour space within the range of the hypothesis<sup>1</sup>. This probability,  $P$ , is given by equation (2).

<sup>1</sup> We should note here that this assumes that each hypothesis begins and ends between sections, rather than somewhere within a section, so that each section is either wholly within or wholly outside of the hypothesis. It will be shown below that the number of such sections can be made to tend to infinity, thus making the color space continuous, and so this assumption is unproblematic.

$$(2) P = \frac{100}{(e-s)q}$$

However, it would seem likely that some of the examples which a child observes might not be accurate, and so if learning is to proceed successfully in the presence of such examples children must have some degree of expectation that any particular example might not be accurate. If the example is not accurate, then a child would have no way of knowing whereabouts in the colour space it would be observed, and so it is assumed that such examples would be equally likely to occur anywhere in the colour space, and this probability, also represented by  $P$ , is given by equation (3). (We should note firstly that even if an example is erroneous, it could nevertheless come within the hypothesis simply by chance, and secondly that equation (3) is independent of the hypothesis under consideration.)

$$(3) P = \frac{1}{q}$$

A child will not know which examples are accurate examples of the colour word, and which are simply random, so it is necessary to introduce a parameter,  $p$ , which corresponds to the certainty with which a child believes an example to be accurate. This parameter can vary from 1, when the child will be completely certain that all examples are accurate, to 0, when the child believe that all examples are random. In the first of these situations a single erroneous example could have a catastrophic effect on learning, while in the second no learning would occur at all, as the child would not believe that the examples gave any indication of the meaning of the colour word. This paper assumes that this parameter is always set to a value between these extremes, so that the model will believe that the examples are indicative of the meaning of the colour word, but will still be able to learn even if some of them are misleading.

Examples which fall outside of the hypothesis must be inaccurate, and so their probability,  $P(e)$ , is given by multiplying together the probability that an example is not accurate,  $(1-p)$ , by the probability of observing an inaccurate example given in equation (3). The equation resulting from this operation is given in (4).

$$(4) P(e) = \frac{(1-p)}{q}$$

If a colour example comes within the range of the hypothesis, then it may be accurate, in which case the equation for its probability could be derived from equation (2), but it could also be inaccurate, in which case its probability could be derived from equation (3). However, when a child is learning they will not be able to be sure which of these two situations applies, and so must consider each possibility according to its probability as defined by the parameter  $p$ . We must derive an equation for the overall probability of an example, based on the possibilities of it being either accurate or inaccurate, with both of these possibilities being weighted in accordance with their probabilities, which are  $p$  and  $(1-p)$  respectively. The total probability of such an example, also written  $P(e)$ , will be found by adding its probability under each of these possibilities, which produces the equation given in (5).

$$(5) P(e) = \left( \frac{100p}{(e-s)} + (1-p) \right) \frac{1}{q}$$

The equations so far are all concerned with only a single example. The probability of all the observed examples given a particular hypothesis ( $P(d|h)$ ) can be found by multiplying together the probabilities of each individual example. Where there are  $n$  examples which come within the scope of the hypothesis, and  $m$  examples which come outside of the hypothesis, this probability can be calculated using equation (6).

$$(6) P(d|h) = \left( \left( \frac{100p}{(e-s)} + (1-p) \right) \frac{1}{q} \right)^n \left( \frac{(1-p)}{q} \right)^m$$

We can extract  $q$  from the terms put to the power of  $n$  and the power of  $m$ , to create a new term of  $q$  to the power of  $n$  plus  $m$ . However,  $n$  plus  $m$  is always the total number of examples observed, and so this value will be constant across all hypotheses. Equation (6) is rewritten as (7) below, where  $r$  is used to represent the total number of examples.

$$(7) P(d|h) = \left( \frac{100p}{(e-s)} + (1-p) \right)^n (1-p)^m \frac{1}{q^r}$$

## 2.4 Calculating Hypothesis Probabilities

So far we have specified how probabilities will be assigned to two of the three terms on the right hand side of Bayes' rule (equation (1)), but in order to calculate the probability of a hypothesis we must also be able to calculate the *a priori* probability of the data,  $P(d)$ . We can do this by taking the product of the probability of the data given a hypothesis and the *a priori* probability of the hypothesis, and finding the total of all these products for all the hypotheses.

However, ideally we do not want to divide the colour space into a number of arbitrarily sized sections, as there does not appear to be any empirical motivation to do so. Hence it would seem desirable that we increase the number of sections of the colour space,  $q$ , so that  $q$  tends to infinity, and the colour space effectively becomes continuous.  $P(d)$  can then be calculated using calculus, as in (8), where  $H$  is the set of all possible hypotheses. (We will see below that the term  $q$  will cancel out of the equations, and so its exact value is unimportant.)

$$(8) P(d) = \int_{h \in H} P(d | h)P(h)dh$$

We now have all the terms which we need in order to calculate the *a posteriori* probability of a hypothesis using Bayes' rule. Substituting equation (8) into Bayes' rule we obtain equation (9), where the hypothesis whose probability is being calculated is now labeled  $h_i$ . All hypotheses have equal *a priori* probability, so the terms  $P(h_i)$  and  $P(h)$  will cancel out. The terms  $P(d | h_i)$  and  $P(d | h)$  also both contain the constant term  $q^r$ , and so this term will also cancel, as was noted above.

$$(9) P(h_i | d) = \frac{P(d | h_i)P(h_i)}{\int_{h \in H} P(d | h)P(h)dh}$$

## 2.5 Generalizing from Examples to Other Colours

The aim of the model is not to determine the probability of any one hypothesis, but to determine how likely it is that any particular colour can be named with the colour word. We can express the probability that a particular colour,  $x$ , comes within the set of colours which can be named by the colour word,  $C$ , using the expression  $P(x \in C | h_i)$ . However this expression only applies when we are sure that  $h_i$  is correct, in which case if  $x$  comes within  $C$  this expression will evaluate to one, and otherwise it will evaluate to zero.

However, what is really needed is an expression for the probability that a colour can be named by the colour word which takes account of all the possible hypotheses. This can be achieved by using the procedure of hypothesis averaging, where the probability that a colour can be named by the colour word if a particular hypothesis is correct is multiplied by the probability of that hypothesis given all of the observed data. Equation (10) shows how the overall probability that the colour can be named by the colour word can be found by summing over these products for all the hypotheses.

$$(10) P(x \in C | d) = \int_{h_i \in H} P(x \in C | h_i)P(h_i | d)dh_i$$

The summation can be performed using calculus, because the colour space is continuous. However, the number of examples which are within the hypotheses changes discretely at the locations of the colour space where the examples occur, and the value of  $P(x \in C | h_i)$  changes discretely at the location of the colour under consideration. For these reasons, the value of the sum must be calculated separately for sections of the colour space between such points, and then these values added together. There is not sufficient space to give the full derivation of the integrals here, though this is presented in detail in (Dowman, 2001). It is worth noting however that the integration is straightforward if the binomial expansion is used to transform the first part of equation (7), although there will be special cases when there are less than three examples within the scope of a hypothesis, or when the hypotheses include either all of the colour examples or none of them.

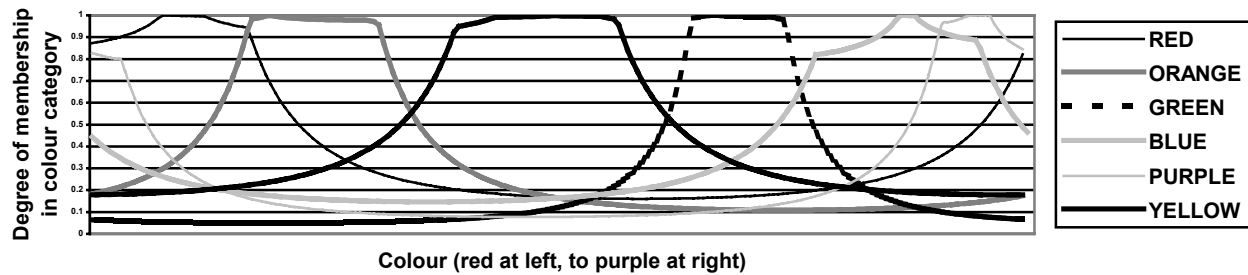


Figure 3 Fuzzy Category Memberships Learned for English Color Words

So far the model has been described simply with respect to determining whether one particular colour can be named by a colour word. However, if we consider not just a single colour, but the full range of colours, then we can assign a probability to each. These probabilities can be interpreted as corresponding to the degree of membership in a fuzzy set, and so the Bayesian model can be seen as defining a fuzzy set representation for the meanings of colour words. This inference procedure is similar to a proposal made by Tenenbaum (1999), except that Tenenbaum's procedure did not allow for parameterization so that inference could be modelled under conditions where a learner's degree of confidence in that data varied. Tenenbaum's approach also did not deal with learning in a circular concept space, although it would seem to be unproblematic to adapt it so that it could be applied in such a situation.

### 3. Learning English and Berinmo Colour Words

In order to investigate the performance of the Bayesian model empirically it was trained on the six basic chromatic colour words of English. The approximate range of each colour word was determined from data published in Berlin and Kay (1969) and ten examples of each word chosen randomly from within the range of colours which the word can name were then given to the model. (The ranges can only be approximate, as even speakers of the same language disagree about the exact range of each colour word.) The model's degree of certainty that each example was correct was set at 0.8, and the probability of membership was calculated for each colour word at 200 points evenly spaced along the colour space. These values were then plotted to produce Figure 3, in which the horizontal axis corresponds to the conceptual colour space, with red at the left, then orange, yellow, green, blue and purple. As the colour space is circular, the left and right edges of the graph represent adjacent points in the colour space.

We can see from the graph that the learned meaning representations have the prototype properties characteristic of colour words. For each colour word there is a single colour which is the best example of it (that is has the greatest degree of membership in the colour category). Moving away from this prototype colour, membership in the category gradually decreases, which corresponds to colours becoming increasingly poor members of the category the further that they are from the prototype. The categories also have fuzzy boundaries, because there is no clear point at which colours stop being members of the category, so some colours may be considered marginal members, especially where their degree of membership is around 0.5.

In order to demonstrate that the system is able to learn a wide variety of types of colour system it was also trained on the chromatic colour words of Berinmo, a language spoken in Indonesia. (The data concerning the meanings of the Berinmo colour words was taken from Roberson, Davies and Davidoff (2000)). Berinmo has five colour words, but one of these

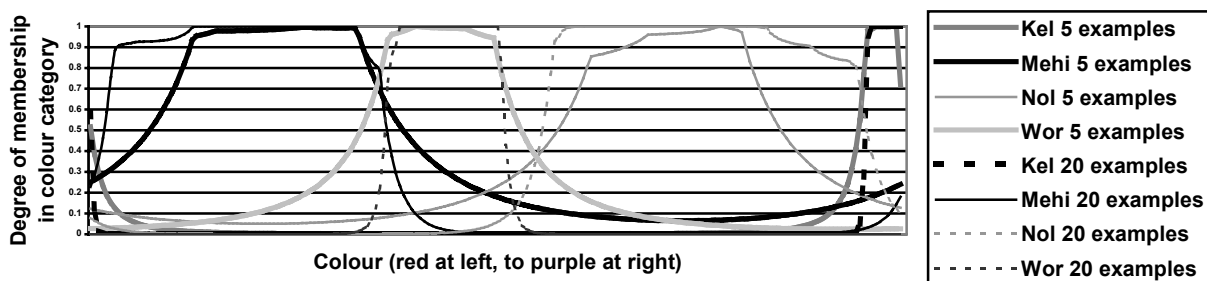


Figure 4 Fuzzy Category Memberships Learned for Berinmo Color Words

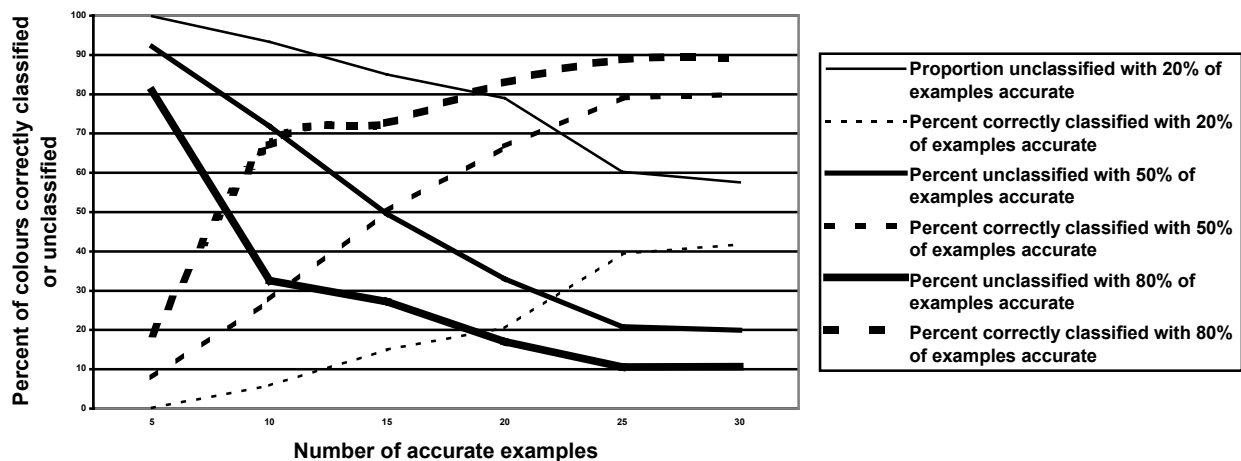


Figure 5 Accuracy of Learning with Noisy Data

only names very dark colours, so could not be learned by the colour model. The result of training the model on the Berinmo colour words is shown in Figure 4, both after the model was trained on only five examples of each word, and after it was trained of twenty examples of each word.

We can see that all the learned meaning representations have prototype properties, although when the model has observed more examples of a category it is much more sure about whether colours are members of a category or not, with the curves getting nearer to the top and bottom of the graph. (Although it may appear that at some points the curves are completely flat, this is simply a property of the accuracy with which the graph is plotted, as the degree of membership never reaches the limits of either one or zero.) The areas where category membership is marginal also become much smaller, with the curves rising and falling much more steeply between the areas of high and low degrees of membership. This is because, with more examples of a category it is possible to be much more sure about which colours are members of the category and which are not.

#### 4. Robustness of the System to Noisy Data

Having seen that the system is able to learn when presented with accurate data, it was then decided to investigate to what extent the learning process would be disrupted by the addition of random noise. A target category of size 30 was created for the model to learn (the whole colour space being 100 units wide). Examples of this category were generated in the same way as for the English and Berinmo colour categories, but varying amounts of completely random examples were added to the training data to simulate noise. The parameter,  $p$ , was adjusted, so that it always accurately reflected the proportion of examples which were accurate, so that the model had advance knowledge of how reliable the data was, though it did not know which particular examples were accurate and which were not.

The model was judged to have correctly categorized a colour if the colour came within the category and was assigned a degree of membership of greater than 0.95, or if the colour came outside of the category and it was assigned a degree of membership of less than 0.05. If a colour was assigned a degree of membership between 0.05 and 0.95 then that colour would be considered not to have been classified. Examples were wrongly classified if they were assigned a degree of membership greater than 0.95, but were in fact outside of the category, or if they were assigned a degree of membership of less than 0.05 but came within the category.

The results of these experiments are given in Figure 5, which shows the proportion of colours classified accurately, and left unclassified, when the level of noise in the data varied from 20% to 80%, and the number of accurate training examples observed varied from 5 to 30. These examples would in each case be accompanied by the number of random examples needed to simulate the appropriate level of noise. The results were derived from sampling at 100 evenly spaced points in the colour space, and in each case investigating whether that colour was classified by the system as coming within the colour category, outside of the colour category, or whether the system did not classify that colour at all. If we view these experiments from the perspective the standard machine learning test data–training data paradigm, then these 100 points would correspond to a stratified sample of 100 test data items, each of which has to be classified. The results from which Figure 5 was plotted were all averages over 20 runs of the system.

We can see that, as more training examples are observed, a higher proportion of colours are correctly classified. Also, if a higher proportion of training examples are accurate, this leads to more accurate classification. However, even when

80% of the data are random, once 30 accurate training examples have been observed (by which time 120 random training examples would also have been seen), over 40% of test colours are classified accurately. When 50% percent of the data was accurate, the model achieved very good performance with 30 accurate training examples, correctly classifying over 80% of the test colours.

Of course classifying a high proportion of examples accurately would not be an impressive result if a large proportion of examples were also classified inaccurately. However, the highest proportion of test colours which were ever classified inaccurately, in any condition, was 0.8%. (This was when 80% of examples were random, and only 10 accurate examples had been observed, and 0.8% is an average, based on all 20 runs of the system in this condition.) In most cases there were even fewer colours classified inaccurately, and in many cases none at all. So we can see that learning can proceed with a very high degree of precision even in the presence of large quantities of noise. When there are very high levels of noise in the training data, or when there is only a small number of training examples available, a large proportion of test colours are left uncategorized. However, even in such conditions, very few examples are classified incorrectly.

## 5. Discussion

The model described in this paper has shown that a Bayesian inference procedure can learn the meanings of colour words from the same kind of evidence that is available to children when they learn the same words. Moreover, the representations it learns account well for psycholinguistic evidence concerning colour words, in particular that they have prototype properties, and so this provides additional support for the proposal that colour words may be learned using a Bayesian inference procedure, or some procedure which approximates Bayesian inference. However, most common nouns, and many other types of word also have prototype properties (for example we may view blackbirds or robins as prototypical birds, but penguins are very marginal members of the bird category). Hence I would suggest that it may be the case that people use Bayesian inference to learn other aspects of language, and that there is a very big potential for Bayesian modelling in linguistics.

While the focus of this paper is on cognitive modelling, this research may also be of benefit in the construction of natural language understanding systems. If such systems are to perform well, they need to understand natural languages in the same way as humans do, because such systems aim to determine the meaning that was intended by the person with whom they are interacting. While the current approach is probably not of immediate practical use in most present day natural language systems, if such systems are ever to achieve a level of language competence approaching that of people, closer attention to some of the more subtle features of natural language, such as the prototype properties focused on in this paper, may be warranted. Furthermore, it may become possible to teach computers the meanings of words through the presentation of examples, and this might result in more accurate representations of word meanings, and could also reduce the cost of producing natural language understanding systems.

We should also note that the model is able to accurately determine set membership, even when presented with very noisy data. Clearly this is helpful for a cognitive model, as any such model should be robust and able to learn in realistic situations. However, this does suggest that the approach used here, or an adaptation of it, could be useful for any learning task in which the aim is to determine set membership based on examples of members of that set, especially where the data was unreliable or corrupted by high levels of noise.

From a cognitive perspective, the main limitation of the Bayesian model is that it does not account for data from the field of linguistic typology. Large surveys of the colour vocabularies of speakers of many languages have been conducted (for example Berlin and Kay, 1969 and MacLaury, 1987), and these have revealed that, while both the number of colour words and the ranges of colours which each word can name varies between languages, this variation is not completely random. In particular the location of the prototypes of the colour words is partly predictable, so, for example, if a language has three basic colour words these are always focussed on black, white and red.

Experiments have been conducted to try to account for the typological data using the Bayesian model. Firstly learning biases were added to the model so that it preferentially remembered examples of focal red, yellow, green and blue. Then several copies of the model were created to form a community of artificial people (agents) who then named colours to one another, usually based on the colour words they had learned up till that point, but occasionally being creative and making up a new colour word. At random intervals one of the agents would be replaced by a new one, and so the evolution of the language was simulated over several generations. This work builds on that of Belpaeme (2002), who also performed computational evolutionary simulations of the cultural evolution of colour term systems, although his agents learned using adaptive networks, not Bayesian inference. Belpaeme was able to show how coherent colour term systems could emerge in a population, but his model was not able to account for typological patterns. Preliminary results using the augmented Bayesian model show that the colour term systems which emerge in the simulations display many

of the properties observed in colour term systems across languages. While the evolutionary model is not the topic of this paper, the Bayesian model's ability to account for aspects of linguistic typology helps support the proposal that people learn colour words with Bayesian inference.

## Acknowledgements

I would like to thank all the people who have helped me with this paper and my work on Colour Terms and Bayesian inference in any way, including Judy Akinbolu, Brett Baker, Cassily Charles, Garry Cottrel, Michelle Ellefson, Eva Endrey-Walder, Bill Foley, Alexander Francis, Yukari Fujiwara, Paul Green-Armytage, Catherine Harris, Timo Honkela, Jim Hurford, Simon Kirby, Johan Lammens, Darren Ler, Emmanuel Letellier, Rob MacLaury, Jon Patrick, Anders Steinvall, Roslyn Frank and my Ph.D. supervisor Judy Kay. I also acknowledge financial support from the Australian Government and the University of Sydney in the form of IPRS and IPA scholarships respectively.

## References

- Belpaeme, Tony (2002) *Factors influencing the origins of colour categories*. Ph.D. Thesis, Artificial Intelligence Lab, Vrije Universiteit Brussel.
- Berlin, B. & Kay, P. (1969). *Basic Color Terms*. Berkeley: University of California Press.
- Bloom, P. (2000). *How Children Learn the Meanings of Words*. Cambridge, MA: MIT Press.
- Boynton, R. M. & Olson, C. X. (1987). Locating Basic Colors in the OSA Space. *COLOR research and application*, Volume 12, No. 2, 94-105.
- Dowman. (2001). *A Bayesian Approach to Colour Term Semantics*. (Technical Report Number 528). Sydney: Basser Department of Computer Science, University of Sydney.
- Gärdenfors, P. (2000). *The Geometry of Thought*. Cambridge, MA: MIT Press.
- 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*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Hård, Anders and Sivik, Lars (1981). NCS – Natural Colour System: A Swedish Standard for Colour Notation. *Color Research and Application*, Volume 6, No. 3, 129-138.
- Honkela, T. (1997). *Self-organizing Maps in Natural Language Processing*. Doctor of Philosophy thesis, Helsinki University of Technology.
- Kay and McDaniel (1978). The Linguistic Significance of the Meanings of Basic Color Terms. *Language*, Volume 54, Number 3.
- Lammens, J. M. G. (1994). *A Computational Model of Color Perception and Color Naming*. Doctor of Philosophy dissertation, State University of New York at Buffalo.
- MacLaury, R. E. (1997). *Color and Cognition in Mesoamerica: Construing Categories as Vantages*. Austin: University of Texas Press.
- Mitchell, T. M. (1997). *Machine Learning*. New York, NY: McGraw-Hill.
- Roberson, D., Davies, I. and Davidoff, J. (2000). Color Categories are not Universal: Replications and New Evidence from a Stone-Age Culture. *Journal of Experimental Psychology: General*, Volume 129, No. 3, 369-398.
- Taylor, J. R. (1989). *Linguistic Categorization: Prototypes in Linguistic Theory*. Oxford: Oxford University Press.
- Tenenbaum, J. B. (1999). *A Bayesian Framework for Concept Learning*. Doctor of Philosophy Thesis, Massachusetts Institute of Technology.
- 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.
- Thompson, E. (1995). *Color Vision: A Study in Cognitive Science and the Philosophy of Perception*. New York, NY: Routledge.