A Bayesian Approach To Colour Term Semantics

Mike Dowman

Mike@cs.usyd.edu.au

Basser Department of Computer Science, F09,
University of Sydney
NSW2006
Australia

1 Abstract

A Bayesian computational model is described, which is able to learn the meanings of basic colour terms from positive examples. Examples of colours named by particular colour terms are stored in a conceptual colour space, and Bayesian inference is used to determine the probability that other colours come within the range of each colour term. A fuzzy set based denotation for each colour term can be created by calculating the probability that each point in the colour space comes within the extension of each colour term. The learned categories show the prototype structure characteristic of colour terms, with there being a single best example of the category, marginal members of the category, and with intermediate colours being members of the category to a greater or lesser extent. This approach has the advantage over previous approaches of being both flexible, and so being able to account for the full range of observed colour term systems, but also of providing a precise and explicit account of colour term semantics. The model is able to account not only for the meanings of colour terms, but also for the acquisition of those meanings. Further work aims to incorporate innate biases into the model, so that it is able to account for observed typological patterns, and the learning biases of human subjects.

Key words: basic colour terms, prototype categories, semantic acquisition, Bayesian inference, fuzzy sets.

2 Introduction

This paper describes a theoretical proposal about the nature of basic colour terms. The proposal concerns not only the nature of a speaker's knowledge of the semantics of such words, but also how a person may learn the meanings of such words. There has been much work on the subject of such words in both linguistics and psychology, as well as in related disciplines, but the present work has the advantage of both being able to account for a wide range of colour term systems, while providing an explicit account of the acquisition of basic colour terms which has been implemented as a computational model.

Every language has words for colours, but in different languages words denote different ranges of colour, so that for a colour term in one language there will not always be a word in another language with exactly the same meaning. Berlin and Kay (1969) identified a subset of colour words which they called *basic colour terms*. These words are colour terms which are monomorphemic, and whose extension is not included in that of any other colour term. They must not be restricted to apply only to a narrow class of objects, and must be psychologically salient (that is, they must be prominent within the range of colour terms known by an individual, and be known by all speakers of a language). Berlin and Kay determined from a sample of 98 languages that all languages have between two and eleven such words (although it is possible that some languages have twelve, as there are some marginal examples).

2.1 The Nature of Colour

It is clear that the extension of a colour term does not consist of objects with just one specific colour, but in fact extends to objects which have any one of a range of similar colours. So in order to understand the denotations of colour terms, we must consider the nature of colour, so that we may determine how such a range of colours could be specified. Firstly, we may consider the physical properties of light, and how these properties vary between light of different colours. Light waves of different colours have different wavelengths. Red light has the longest wavelength, with orange, yellow green, blue and purple having increasingly shorter wavelengths. However, light can be made up of a mixture of these colours, and can occur at varying intensities. Hence the perceived phenomenological colour of light depends on which wavelengths are present, and on the intensity of each wavelength. Colours form a continuous range with red at one extreme and blue at the other (Thompson, 1995).

However, the fact that light has such a physical structure, does not mean that this is the form in which the nervous system processes colour. In fact, only three types of receptor cell have been identified in the retina of the eye, and each type responds differentially to light of a given wavelength. (I am referring here to the cells known as *cones*; the retina also contains cells called *rods* which are believed not to play an important part in colour vision.) Each type of cell is maximally active when presented with light of a particular wavelength, and its activity decreases gradually as light a of greater or lesser wavelength is presented. So, while physical light has the property of having a continuous range of wavelengths, the eye only senses colour in terms of how close the frequency of light reaching the eye is to the focal frequency of each of these three types of cell, and how intense it is (Thompson, 1995).

Further research has traced the neurophysiological pathway of information about the colour of light as it is processed by the nervous system, and has identified a particular class of cells known as *opponent cells*. These cells process the output of the cones to produce a signal based on a combination of the output of these cells. Some cells oppose green and red light, and so respond most in the presence of green and absence of red wavelengths. Other cells also oppose green and red light, but have opposite responses, responding most in the presence of red wavelengths and absence of green ones. Similarly, there are two types of cells which oppose blue and yellow light, each type responding most in the presence of wavelengths of one colour and the absence of wavelegnths of the other colour. These cells hence map the input light onto a two dimensional colour space, as is shown in Figure 1. The color of light will be determined by its degree of blueness or yellowness, corresponding to the horizontal axis, and its degree of redness or greenness, corresponding to the vertical axis. A third kind of opponent cell has also been postulated, which would oppose light and dark light, and so create a third dimension of lightness (Kay and McDaniel, 1978).

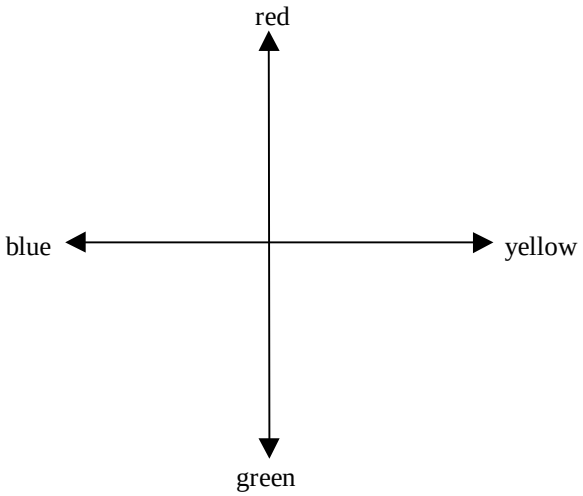

Figure 1. The Two Dimensional Colour Space

So far I have discussed the nature of colour from the perspective of its physical structure, and neurophysiological evidence as to how people process it, but we may also consider psychological and phenomenological evidence as to the nature of colour. While physically red has the longest wavelength of any colour of visible light, and purple the shortest, phenomenologically red and purple are similar colours, and there are some colours which could be considered as marginal examples of both red and purple. There has been much research on colour from psychological and phenomenological perspectives, and general consensus has been reached that, phenomenologically, colour has a three dimensional structure, which more closely relates to the properties of opponent cells than to both the physical structure of light, and to earlier stages of neurophysiological processing (Thompson, 1995).

The three dimensions used in describing the phenomenological structure of colour are *hue*, *saturation*, and *lightness*. Firstly, lightness corresponds to how bright or dark a colour is, and corresponds closely to the postulated light and dark opponent cells. The dimension of hue identifies the property of light which changes continuously from red to orange, yellow, green, blue and then purple, and finally back to red again. This dimension is orthogonal to that of lightness, and is also circular, as if one progresses along it in one direction, no matter where the start point is, the original hue will eventually be reached. Hue corresponds partly to the two dimensional neurophysiological colour space represented in Figure 1, though that colour space also encompasses the phenomenological dimension of saturation. The dimension of saturation corresponds to how vivid, or undiluted a particular colour is, irrespective of its hue. At a level of zero saturation, all colours are black, white or grey, varying only on the lightness dimension, not on the dimension of hue. At all other levels of saturation, a colour may have any hue, with saturation reaching a maximum when colours are at their most vivid, for a given hue and lightness. These two dimensions are shown in Figure 2, which illustrates how saturation may be considered a radial dimension, and hue a circular one. The dimension of lightness is orthogonal to both these dimensions.

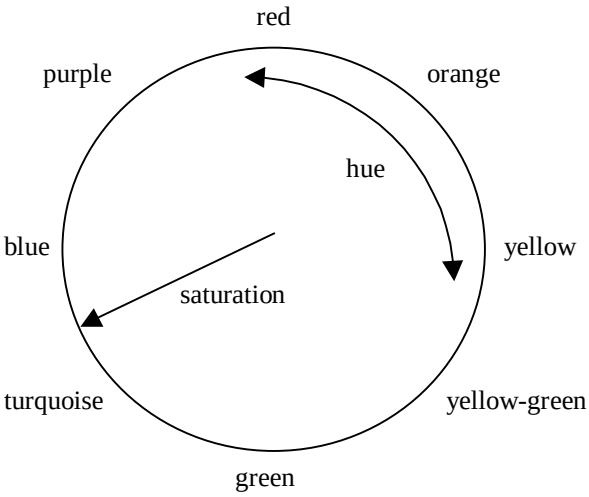

Figure 2. The Phenomenological Dimensions of Saturation and Hue

The three dimensional colour structure based on the dimensions of hue, saturation and lightness corresponds to how people describe the experience of colour, and how they categorise colours. It also appears to be the colour structure which allows the best and most concise description of the extensions of colour terms, and which has hence has been adopted as the system used in almost all investigations into basic colour terms. I now move on to consider a number of theoretical approaches to the semantics of such words, many of which share much in common with the approach used in this paper, but all of which have at least some important differences.

2.2 Approaches to Colour Term Semantics

Often language semantics are explained using a feature-based approach. The meaning of a word will be explained in terms of necessary and sufficient features. So the meaning of a word such as *bachelor* would be decomposed into the features, *man* and *unmarried*, which together are both necessary and sufficient for determining the property of being a bachelor. These features may themselves be decomposed into more primitive elements, for example *man* could be decomposed into *two-legged* and *mammal*. This approach is fairly successful at capturing the meaning of words such as *bachelor*, as there is a clear distinction between things in the world which are bachelors, and those which are not. However, it is difficult to see how such an approach could be useful in understanding the meaning of words such as colour terms.

Taylor (1989) notes that colours terms are an example of prototype categories. For any colour term, say English *blue*, speakers can readily identify a single best example colour for this term, what is called a prototype. Colours similar to this best exemplar will be considered good or bad examples of the colour term to a varying degree, depending on how similar they are to the prototype. Colours which are more dissimilar to the prototype are considered worse examples of the colour term than those which are more similar. Clearly, a point will be reached where a colour is so dissimilar to the prototype that it is no longer an example of the colour term at all, and instead comes within the denotation of a neighbouring colour. However, colour terms also exhibit another property typical of prototype categories, that there are some examples about which it is difficult to determine whether they come within the denotation of the term at all. These will be colours towards the periphery of the category, and there is often disagreement between speakers about which colour term is the correct one to name such colours.

Central to prototype theory (Taylor, 1989), is the idea that the prototype plays a role in determining which objects (or colours) come within a category, and which do not. However, exactly how the

prototype defines the extent of the category is usually not made explicit. Probably the most explicit prototype approach to colour term semantics is that of Lammens (1994). Lammens defines a colour category by specifying the category's prototype, and the size of the category. The degree of membership of a colour in the category depends on its distance from the focus, and the size of the category is determined by a numeric parameter specifying how rapidly the degree of membership in the category decreases as the distance from the focus increases. Lammens also proposes that his model can form the basis of an account of how the meanings of colour terms can be learned. However, I think in this respect there are some problems with his model, which also have implications for whether the prototype approach is the correct way to define colour categories.

Lammens' model does not specify how a language learner can establish the foci of colour categories. Instead it is assumed that these must be fixed, presumably innately (p. 143), and learning will then consist of determining the extent of the colour category. Hence, the parameter adjusted during learning simply adjusts the size of the category, and can neither affect the location of its focus, nor its shape. This seems somewhat problematic, as typological evidence seems to suggest that not all categories have universally determined foci, and the shape of categories, and hence exactly where their boundaries are, certainly varies between languages. For these reasons I believe that Lammens' work illuminates some of the most problematic aspects of the prototype approach to linguistic categories.

I am aware of only one other computational model of colour term semantics, and that is the work of Honkela (1997). Honkela uses self-organising maps, which can be seen as a form of neural network, to learn the meanings of colour terms from examples. Honkela's model is motivated partly by neuropsychological evidence, but it takes little account of the specific literature on colour perception and colour terms. Instead of basing the input to the model on the phenomenological dimensions of hue, saturation and brightness, the input is based on red, green and blue components of colours, such as are used by computers in generating colours for display on a monitor. Hence it seems unlikely that Honkela's model will account particularly well for psychological and linguistic data on colour terms[1]. However, it seems that self-organising maps may well prove useful in the computational modelling of colour term semantics, and provide an interesting alternative both to prototype approaches, and to the approach taken in this paper.

Other than computational approaches, perhaps the most explicit approach to colour term semantics is that of Kay and McDaniel (1978). Instead of using a strictly prototype approach, such as that of Lammens, they give colour categories an interpretation in terms of fuzzy sets. Fuzzy sets are sets where the elements in the set can be members to a greater or lesser degree, varying from one (full membership) to zero (not a member of the set at all.). A fractional number between one and zero indicates that an element is a member of the set to an intermediate degree.

In Kay and McDaniel's approach, the prototype of a colour category is given a membership value of one, and the degree of membership in the set decreases the further away a colour is from this prototype. The membership functions, which peak at the prototype, and decrease on either side, are based on the response functions of opponent process cells in the retina of the eye. As the category membership functions, and hence the category prototypes, are determined innately by neural response functions, there is little room for cross-linguistic variation, or in fact of learning of category structure at all. While Kay and McDaniel propose that composite categories can be derived from the fuzzy union of the sets defined by the response functions of more than one type of opponent cell, and non-primary categories (such as brown and purple) from the fuzzy intersection of the response

[1] I should note here that the primary purpose of Honkela's work was concerned with language technology, and not cognitive modeling, so it is unsurprising that it does not take account of psychological and neurophysiological data.

functions, otherwise the colour categories are universally determined by the response functions of the opponent cells.

Hence Kay and McDaniel's approach is problematic, as the location of the foci of secondary colour terms, and of the boundaries of all colour terms, appears to vary between languages, and so presumably must be determined, at least in part by learning. Also, there does not seem to be any reason why the degree of membership in a colour category should relate in any direct way to the neural response functions of cells in the retina, and so a model based on this assumption might be considered somewhat unlikely *a priori*. The model presented in this paper also makes use of the fuzzy sets approach, but the shape of the membership functions of the sets defining the denotation of the colour terms is determined by learning from examples, and not from innate neural responses.

3 Bayesian Approaches to Psychology and Linguistics

The computational model used in this paper uses a form of Bayesian inference to learn semantic representations for colour terms. In this section, I will justify why it is plausible that such a model may give an accurate account of how people learn the meanings of basic colour terms. The primary reason for supposing that people may use Bayesian inference to learn basic colour terms is that there are many other psychological phenomena which appear to be best explained using Bayesian inference.

Firstly, Brent and Cartwright (1997) have used a Minimum Description Length based model to account for how children can simultaneously learn to segment speech, and to begin the process of lexical acquisition. He has shown that such an approach could allow children to make rapid progress in this task, and so it is plausible that they use some such strategy to aid in learning to segment speech and acquire a lexicon. Minimum Description Length can be seen as a version of Bayesian inference, and so such approaches suggest that Bayesian models may well give a good account of other aspects of psychology. Huttenlocher, Hedges and Vevea (2000) show that, at least in some circumstances, people appear to use Bayesian inference to improve accuracy in judging inexactly represented stimuli. Dowman (2000) has shown how Bayesian based inference (in this case Minimum Description Length) can aid in the acquisition of syntax, and in particular can account for the acquisition of the subcategorizations of many verbs. Finally, Ellison (1992) has shown how similar learning algorithms can account for the acquisition of many aspects of phonology.

Although the above is a very brief overview of some of the work in psychology and linguistics which makes use of Bayes' theorem, the implication of such work is that many psychological processes may have a Bayesian basis, and so it seems likely that this will also be true of the acquisition of basic colour terms. Hence this suggests that it is worthwhile investigating what properties basic colour terms would have if they were learned using Bayesian inference. If a Bayesian model of colour term semantics corresponds well with observations made concerning colour terms, then it will be good evidence that people do learn colour term semantics using some form of Bayesian inference.

I now want to describe in a little more detail one paper by Tenenbaum and Xu (2000), and another by Griffiths and Tenenbaum (2000), as these two papers are closely related to approach taken in this paper. Tenenbaum and Xu, in common with the approach of this paper, used Bayesian inference to model the acquisition of word meanings. The model predicted that the meaning that people would attribute to a word would depend on the number and type of examples of its use they had observed, which corresponds well with empirical findings. For example, the model predicted that if a person observed a word being used to refer to a dog, and that was the only example of the use of that word which they observed, then they would be unsure whether the extension of this word corresponded to the set of all dogs, the set of all animals, or just the particular dog referred to. However, after seeing

several examples, all of which were dogs, the model was almost certain that the word's extension was the set of dogs. Alternatively, after seeing examples consisting of three different kinds of animal, the model would be almost certain that the word denoted the set of animals, or after seeing several different examples of the same dog, the model would become almost certain that the word denoted only the particular dog. This work demonstrates how Bayesian inference may be used to learn word meanings from only positive examples of the use of a word, and Tenenbaum and Xu were also able to demonstrate that the generalizations made by their model were very similar to those made by people when presented with the same evidence.

The Bayesian model which is most closely related to the work in this paper, is that of Griffiths and Tenenbaum (2000), although their work is not itself concerned with word meanings, or with language at all. Griffiths and Tenenbaum investigate how people can predict the frequency with which some event occurs, based on observations of how long since it last occurred. For example, if people are told that on arrival at a subway station it has been 103 seconds since the last train arrived, then they will guess that it is most likely that trains run every few minutes, consistent with the predictions of the Bayesian model. However, if they are then told that on two subsequent visits to the subway station it has been 34 seconds, and then 72 seconds since the last train arrived, then they are likely to believe that trains run with a frequency much closer to 103 seconds. Again this is consistent with the predictions of the Bayesian model. Within the model, hypotheses correspond to how often trains arrive at the station. The single most likely hypothesis will be that which is large enough to include just the arrival times of all the trains, but the model finds the time such that it is equally likely that the true interval between trains is greater than or less than this time. Hence the value inferred for the length of time between when trains arrive, will always be greater than the longest observed time elapsed since a train arrived. However, as more data is observed, the inferred interval will get closer to the highest observed time, as with more observations it is more likely that at least one of the observed times comes close to the maximum time that may elapse between arrivals of trains.

The reason that the approach to colour term semantics presented here is similar to Griffiths and Tenenbaum's approach to inferring the frequency of events, is that both colour and time can have continuously varying values. Hence, we can use numeric scales to represent dimensions in the colour space, in much the same way as Griffiths and Tenenbaum used such a scale to represent time. The following section describes the model of colour term semantics in detail.

4 A Bayesian Model of the Acquisition of Colour Term Semantics

The computational model proposed here is an account of how the semantics of colour terms may be learned from a finite number of examples. It assumes that, during the time when a person is learning their language, they will observe a finite number of examples of the use of a colour term, and will be able to determine the colour of the object referred to using the colour term on at least some of these occasions. It is these occasions, when a language learner matches a colour term to an example colour named by that term, which provide the input to the process of acquisition. Hence the data from which the model will learn the meanings of colour terms is a set of examples of specific colours, and each such example will be paired with a word thought to denote that colour.

4.1 Pre-Linguistic Processes and Representations

Before it is possible to propose a model of the acquisition of the meanings of colour terms, it is first necessary to determine what form the pre-linguistic input to the process of acquisition will take. Clearly people must be able to perceive colours before they are able to learn the meaning of colour

terms which denote ranges of colour[2]. Hence, in order for a person to acquire the meaning of colour terms, there must be processes which, taking as their input the physical light wave entering the eye, process this signal to form a representation of the colour which may serve as input to linguistic processes[3]. As noted above, the relationship between the physical properties of light entering the eye, and the perceived colour is not simple, and must be moderated by a number of intervening processes. Much is known about the physiology of the colour system, and there are a number of theories to account for phenomena such as perceived colour constancy despite varying illumination (see for example Land (1977)). However, this paper is not concerned with such processes; it is simply assumed that such processes exist, and that they are able to map from physical colours to a representation corresponding to phenomenological colour.

As noted above, phenomenologically colour has a three dimensional structure, varying on the dimensions of hue, saturation and lightness. At present, the acquisitional model is concerned only with the hue dimension. If only colours of maximum saturation, and the degree of lightness at which maximum saturation may be achieved are considered, then these colours will be arranged in a one dimensional colour space as shown in Figure 3. If, starting from any given hue, this hue is repeatedly changed to a neighbouring hue in a consistent direction, eventually the initial hue will be returned to after going through all the other hues for the chosen saturation and brightness.

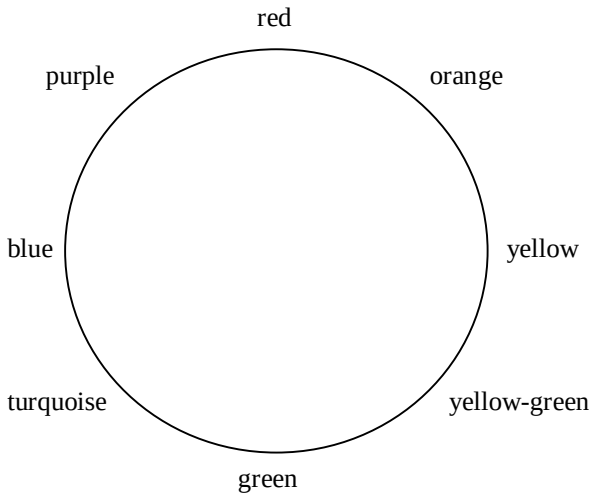

Figure 3. The Phenomenological Colour Space

The model assumes that prior to language learning such a colour space is available to a language learner, and so they will know (at least subconsciously) that, for example, red is more similar to purple than it is to blue.

Clearly, colour terms denote colours corresponding to three dimensional volumes of the colour space, and not simply to parts of a one dimensional colour space. It is hence an assumption of the

[2] It should be noted that an exception to this is the acquisition of colour terms by congenitally blind children. Landau and Gleitman (1985; cited in Bloom, 2000) found that blind children's knowledge of colour words was in many respects similar to that of sighted children of the same age. For example they were aware that colour words belong to a single domain, and that they apply only to concrete objects, as well as being aware that they denoted a property which the children themselves could not identify. While it seems that in these circumstances many aspects of the meaning of colour terms may be acquired, the most central aspect of the meaning of colour terms, that is their denotation of certain ranges of colour, cannot be acquired in such circumstances. This evidence does however suggest that contextual and morpho-syntactic cues may be an additional source of evidence used by children in determining the meaning of colour terms, although such cues are not at present utilized by the model proposed here.

[3] For present purposes it is assumed that such processes are in place before language learning begins. It is of course possible that significant development of such processes may take place during the acquisition of the meaning of colour terms, but it is assumed here that it is unlikely that such effects will have a major impact on colour term acquisition, and so may be ignored.

present model that colour term acquisition may be usefully modelled when considering only a single dimension of the colour space. This assumption is made for the purposes of simplicity only, and there is no reason to suppose the model should not be extensible to operate over the full three dimensional colour space. However, this restriction means that the model will be concerned only with the chromatic colour terms. The meanings of achromatic terms such as *black*, *white* and *gray* cannot be learned with the present model, and nor can the model distinguish between colour terms which differ principally on the dimension of lightness, for example *red* and *pink*. The model is, however, able to give an account of the acquisition of aspects of meaning determined by differences on the dimension of hue, such as those between *red*, *orange*, *yellow*, *green*, *blue*, and *purple*, and many other terms in other languages.

When a learner observes a colour term example, it will be represented as a point in the phenomenological colour space, and this point will be labelled with the colour term itself. For the purposes of the model, hues will be numbered using an arbitrary numbering scheme, with its origin (zero) being in the red space, and increasing through orange, yellow, green, blue and purple, up to hue 99, which will again be in the red space, adjacent to hue 0. Using this scheme, it is possible to represent instances of observations of colour terms simply by a pairing of the colour term and the corresponding hue number.

4.2 Generalization from Examples to Other Colours

So far, the form of the model's conceptual colour space, and the representation of observed colours has been described. However, this does not address the issue of how people will assign names to previously unseen colours, when those colours are observed in the absence of a linguistic label. The model uses Bayesian inference to determine how likely it is that a given colour term is a correct label for a particular colour, generalising from the examples of the use of that colour term already remembered by the model.

Within the context of Bayesian inference, a hypothesis will be a specification of which colours can be correctly labelled with a given colour term, and the observed data will be the set of example colour terms which the model has remembered having been labelled with the colour term under consideration. Bayes' theorem, as given in (1) will be used to determine how likely each given hypothesis is given the observed data, but to do this it is first necessary to specify what the possible hypotheses are, and how likely each one is *a priori* (*P(h)*), and how to determine how likely it is that any given set of colours will be observed to have been labelled with the colour term in question, when assuming that a specific hypothesis is correct (*P(d | h)*).

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

4.2.1 Possible Hypotheses and their *a Priori* Probabilities

Hypotheses as to the denotation of a colour term all correspond to a continuous section of the phenomenological colour space, as shown in Figure 4 below. Each hypothesis will have a start point, *s*, and an end point, *e*. A hypothesis states that a colour is a correct label for all and only those colours which fall after the start point, and before the end point, when the colour space is followed clockwise. Each of these points can be referenced using the numbering scheme for the phenomenological colour space described above. Hence, the size of the colour space denoted by a colour term corresponding to a particular hypothesis will be given by *(e-s)*. In the case where the range of the colour term encompasses the origin, then 100 (the size of the phenomenological colour

space) must be added to *e*, as in this case the value of *e* would otherwise be less than *s*, resulting in a negative value for the size of the colour space denoted by that term.

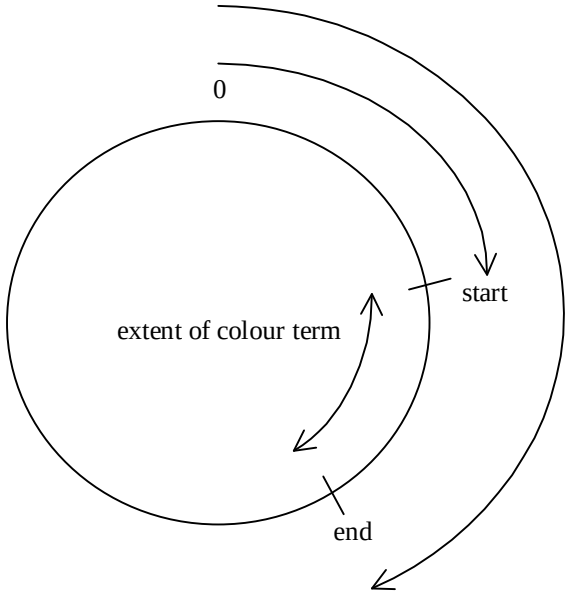

Figure 4. A Hypothesis as to the Denotation of a Colour Term

A hypothesis may begin anywhere within the phenomenological colour space, and may end anywhere within this space. All possible start points are considered equally likely *a priori*, as are all possible end points given any particular start point. Hence, there will be a continuous space of hypotheses, and hypotheses will range in size from not including any of the colour space at all, to including the whole of the colour space, with all sizes of hypothesis being considered equally likely *a priori*.

4.2.2 The Probability of Data given a Hypothesis

Given the above specification of hypotheses, it is now necessary to specify how probable it is that the remembered colour examples labelled by the current colour term would have been observed given a particular hypothesis. This will specify how to calculate the term $P(d \mid h)$ from equation (1) above.

It is assumed first that language learners make no *a priori* assumptions that certain colours are more likely to be named by colour terms than are other colours. This assumption is probably not entirely accurate, as certain colours are almost certainly more frequent in a person's environment than others, and linguistic reference to some colours is likely to fulfil greater purpose than reference to others, and so for these reasons we would not expect all colours to be named with equal likelihood. However, as in the present simulations all colours are equally likely to be named, this seems a reasonable assumption to make. In an environment where this was not the case, it seems that learners would have to take account of this factor, and adjust their learning strategy appropriately. Certainly there seems to be no reason why the model could not be extended to take account of such factors, but for present purposes this added complexity seems unnecessary..

Further issues concern how a given colour example comes to be paired with a given colour term. Such examples would be derived primarily by a person observing the use of a colour term, and inferring the entity, the colour of which the colour term was being used to identify. The learner would then pair the colour term with the perceived colour, and remember this pair for later reference.

4.2.2.1 Learning with Completely Reliable Input

The simplest assumption which we may make about how a learner views this source of evidence as to the denotation of a colour term, is that they may always consider the pairing of the colour term and the colour to be completely accurate. Given this assumption, the probability of any colour example for the colour term under consideration is given by (2)[4] for any hypothesis which includes that colour in its range. The hypothesis simply states that an example of the colour term is equally likely to occur anywhere within the range of the hypothesis. Hence the probability of a given colour example occurring at any particular point within the range of the hypothesis is equal to one divided by the number of such points (e - s).

(2) $P = \frac{1}{(e-s)}$

In most cases there will be more than a single example of each colour term, so the probability of all the examples will be equal to the product of the probability of each, so long as each is within the scope of the hypothesis. As the probability of each individual colour example given the hypothesis is given by (2), the probability of all the data given a hypothesis is given by (3)[5], where *n* is the number of examples of the colour term.

(3) $P(d \mid h) = \frac{1}{(e-s)^n}$

However, the above equation only applies in the case that all the observed colour examples are within the range of the hypothesis. As the assumption has been made that all the observed colours are correct examples of the colour term, the probability of any individual colour example being outside of the range of the hypothesis is zero. Hence the probability of the data given a hypothesis in which one or more colour term example is given an incorrect label is given by (4).

(4) $P(d \mid h) = 0$

4.2.2.2 Learning with Unreliable Input

It seems worth considering whether the above assumption, that a person will always assume that a given colour term and colour example will always have been correctly paired, is correct. There are good reasons to believe that a language learner would not be completely certain that they had always correctly identified the colour being referred to by a particular colour term. The pairing of a colour term and a colour relies on correctly identifying the colour term being used, and establishing what colour the speaker intended to identify using this term. There are a number of factors which could lead to a person incorrectly pairing a colour term and a colour, creating considerable scope for error in deriving colour term and colour pairs.

Firstly, a speaker may use the wrong colour term to describe a colour which they wish to refer to. Such an error could arise for a number of reasons. On some occasions this kind of error would arise

[4] Note that this equation assumes that colours are specified with an accuracy corresponding to the unit size of the scale of hues used. However, if colours are specified to a different accuracy, this will not affect the specification of the model, as below where a formula for the final semantic interpretation of a colour term is specified, the component *P(d | h)* is divided by the integral of this term over all hypotheses, and so any constant term scaling for the precision with which colours are recorded would cancel out.

[5] This formula assumes that the number of colour examples is known *a priori*. As this value does not change between hypotheses, this assumption will not have any effect on the model.

simply through a speech error, or the speaker may have incorrectly perceived the colour. This possibility would be especially likely under poor or unusual illumination, as these conditions would make the task of accurately perceiving the colour more difficult, though even small errors in perceiving colour for colours near the margins of the denotations of colour terms could produce this kind of error.

However, even if a speaker perceives a colour accurately, and does not make a performance error, they may still label a colour term inaccurately. This would happen in one of two cases, either because a speaker was unsure of the correct colour term with which to label the colour in question, or because they had incorrectly learned the denotation of the colour term. If a speaker has observed only a limited number of examples of a colour term, then they are unlikely to be certain about exactly which colours this term can be correctly used to refer to, and this is especially likely to affect colours near the margins of its denotation. It is also possible that a speaker has simply learned the denotation of a colour term inaccurately. This could happen for a number of reasons, including the speaker having learned from inaccurate input. Either of these two factors could lead to a learner incorrectly pairing colour terms with colours.

The above discussion may seem to presuppose a single correct denotation for all colour terms. However, this may seem problematic, as exactly what a 'correct' denotation could be, other than the denotation believed to be correct by a speaker of the language, is unclear. It does seem reasonable, however, to consider there to be a 'correct' denotation, based not on the semantic representation of an individual speaker of a language, but on some kind of norm of the speech community as a whole. I will not discuss exactly how such a speech community norm may arise from the knowledge of individual speakers. Instead I will simply assume that in practice individual uses of colour terms to name colours may be classified as correct or incorrect, irrespective of the language use of any individual speaker.

Having considered the cases in which a speaker may come to use an incorrect colour term to describe a particular example colour, it now remains to consider errors which may arise even in cases in which a speaker has used a colour term correctly. Errors could arise through the learner believing that the speaker had used a colour term to identify a colour, when in fact no word had been used to denote a colour, when a hearer incorrectly inferred the colour which the colour term was used to describe, or the hearer may simply have misperceived the colour.

Firstly, a hearer could come to believe a colour term had been used to describe a colour, when in fact it had not been used to describe a colour at all. Colour words are usually polysemous, and may also have homonyms. Consider, for example, the sentence *Mary has green fingers*. A language learner might then infer that the term *green* may be used to describe the colour of Mary's fingers, when in fact the speaker meant to say that Mary is very good at gardening. Errors of this sort could also happen if the learner misperceives a word which is not a colour term, in such a way that they believe that it is. They are then likely to associate a colour with the word they believe to have been uttered, when in fact that word had not been used by the speaker. Such errors are especially likely given that acquisition of other areas of language is taking place simultaneously to the acquisition of the semantics of colour terms.

Secondly, a hearer may misperceive a colour in exactly the same way as a speaker might, with the same consequences for learning. However, possibly the most important factor leading to a learner coming to pair a colour term with a colour example incorrectly, is that they may misidentify the object, or the particular part of that object, the colour of which was being referred to. Determining what colour a colour term was used to indicate on a particular occasion of use is a complex process, in which a language learner must make inferences about exactly what object it is that has the colour which is denoted by the term the speaker is using. Much has been written about this process (for

example Bloom (2000) discusses it at length), but the computational model discussed here will not focus on this aspect of word learning. For present purposes it is sufficient to note that this is a complex process, and there are many opportunities for error. Here it will simply be assumed that colour examples which do not come within the denotation of a particular colour term may nonetheless on some occasions come to be paired with that term. Such erroneous pairing may arise for any of the reasons mentioned above, and possibly also for other reasons not considered here.

Given all of the above considerations, it seems clear that a learner will not be completely certain that in all cases the colour which they believe to have been named with a particular colour term, is actually a colour that correctly falls within that term's denotation. Hence formulas (3) and (4) above do not correctly specify how $P(d \mid h)$ should be calculated, as both these formulas assume that the language learner believes the input data to the process of generalisation from examples to be completely accurate.

It seems necessary to incorporate into these formulas a factor accounting for the possibility that some of the example colours for the colour term under consideration are likely to be incorrect. For this reason a constant p is introduced, which corresponds to a person's belief that a colour has been correctly associated with a colour term, as opposed to the colour arbitrarily being anywhere within the phenomenological colour space. More specifically, p is the probability with which the learner assumes that each colour example has been correctly paired with its colour term label, and $(1 - p)$ is the probability with which the learner assumes that the applied label is completely arbitrary, and hence and may or may not correspond to a correct colour term for the colour in question. In the case where p is equal to one, the model will correspond to the case outlined above for complete certainty of correctness in colour and colour term pairing.

This treatment of a learner's confidence in the correctness of the data is of course very simple. It assumes that for all colour terms, and for all example colours, there will be an equal probability of erroneously matching a colour and colour term. However, there are a number of factors which might affect the confidence which a person had about correctly identifying the usage of a colour term. Firstly the circumstances in which the colour terms were matched to particular colours might lead a learner to be more certain that they had correctly paired a colour term with the colour it was used to refer to on some occasions than on others. This would imply that different values of the constant p might be appropriate for different colour examples.

People might also adjust their confidence that they are able to accurately pair colour terms to colours, by considering how well the inferred examples of each colour term correspond to plausible denotations. Such considerations would give a person some indication of how probable it is that they had been successful in pairing each colour term to the colour it was used to denote. In the model this would correspond to adjusting the value of the parameter p to an appropriate value taking account of the observed data. The effect of these factors could be investigated in extensions of the model, but for present purposes a person's confidence in the correctness of the data is modelled simply by assuming a fixed value for the parameter p.

Another assumption of the current model is that, if a colour is not a correct example of the colour term, it is equally likely to occur anywhere in the phenomenological colour space. Even ignoring the possibility that some colours are more frequently named than others (as discussed above), this assumption may not be accurate. Whether this assumption is valid would seem to depend on how the colour term had come to be paired with an incorrect example colour. For example, if the error arose through the speaker or hearer having misperceived the colour, we would expect the colour to be near

the boundaries of the denotation of the colour term. In contrast, if the hearer had misidentified the object to which the colour term applied, or the speaker had been using a colour term to denote a property other than colour, then the assumption would seem justified. Hence the extent to which the assumption is justified is dependent on the relative importance of factors such as these.

However, during the course of the acquisition of colour terms, it would seem to be a very difficult task for children to assess how accurate the data from which they are learning is. Only when they had identified the ranges of the denotations of each colour term with a high level of accuracy, would children be able to make accurate estimates of how likely it is that their observations of colour examples for each colour term were correct. Hence, it seems plausible that they make simple assumptions such as those proposed here, about both the reliability of their input examples, and the properties of any erroneous data. Certainly there are alternative assumptions that could be made, many of which might be equally plausible. Consideration of such alternatives would be a worthwhile avenue for research, but at present it seems sufficient to consider only the one possibility.

Given this parameter for confidence in the data, it is now necessary to derive new formulas specifying the probability of the data given a particular hypothesis. First of all considering, only a single colour example, if it is known that it has been correctly paired with the hypothesis, then, assuming the hypothesis to be correct (and hence also that the colour example comes within its denotation), its probability is given by (5). It may be noted that this is the same as equation (2) above for completely reliable data.

(5) $P = \frac{1}{(e - s)}$

If it is known that a colour example was not correctly paired with a colour term, and hence we assume that it is equally likely to occur anywhere within the colour space, its probability is given by (6).

(6) $P = \frac{1}{100}$

However, a learner will not in fact know whether a colour has been correctly paired with its hypothesis, so we must incorporate into these formulas the probability with which a learner assumes that the colour is, or is not, a correct example of the use of the colour term. Firstly we can consider the case in which the colour example is outside of the range of the hypothesis. In this case the colour example must have been incorrectly paired with the hypothesis, and so its probability is given by (6), weighted by the probability that an arbitrary example colour has been identified to correspond to the colour term. This probability is *(1 – p)*, so the resulting formula is that given in (7).

(7) $P = (1 - p) \times \frac{1}{100} = \frac{(1 - p)}{100}$

If a colour example comes within the range of the hypothesis, then it may have been correctly paired with the colour term, in which case its probability would be derived from (5), or it may have been incorrectly paired with the colour term, but nonetheless come within its denotation, in which case its probability would be derived from (6). In fact, in order to determine the overall probability of the example given the hypothesis, we sum the probability of the example under both of these conditions, weighted by the probabilities with which the learner assumes either of these two possibilities. The

probabilities are *p* and *(1-p)* respectively, so the overall probability of such an example is given by (8).

(8) $$P = p \times \frac{1}{(e-s)} + (1-p) \times \frac{1}{100} = \frac{p}{(e-s)} + \frac{(1-p)}{100}$$

In order to determine the probability of all the data given a particular hypothesis (*P(d | h)*), it is necessary to multiply together the probability of each individual example. Where there are *n* examples which fall within the range of the hypothesis, and *m* examples which fall outside of the range of the hypothesis, this probability is given by (9).

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

The above formula is the one which has been implemented so as to allow the probability of the observed data given any of the permissible hypotheses to be determined. However, it is not how probable data is given a hypothesis that we are interested in, but in fact it is the probability of hypotheses given the data. In order to calculate this probability using Bayes' theorem we need to consider the probability of one further component, as described in the following section.

4.2.3 The *A Priori* Probability of the Data

The final term on the right hand side of Bayes theorem (as given in (1)), is *P(d)*, corresponding to the *a priori* probability of the data. This term specifies how likely the data was before we considered any particular hypothesis. While this term is constant across all possible hypotheses, we need to consider it in order to determine the probability of each individual hypothesis given the observed data. The value of the term can be calculated by summing the probability of the data given each individual hypothesis, multiplied by the *a priori* probability of that hypothesis. As the hypotheses are continuous (that is, there are no discreet boundaries between hypotheses) we can use calculus to calculate the value of this sum. This is expressed mathematically in (10), where *H* is the set of all possible hypotheses.

(10) $$P(d) = \int_{h \in H} P(d \mid h) P(h) dh$$

4.2.4 The Probability of a Hypothesis given Observed Data

We now have a specification of all the terms which we need to complete the right hand side of Bayes' theorem as given in (1). Substituting the expression given in (10) into this equation, we obtain (11). (It should be noted that some of the occurrences of *h* have been relabelled h_i, as otherwise there would be two different variables both given the label *h*.)

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

As all hypotheses are equally likely *a priori*, the terms $P(h_i)$ and $P(h)$ are both constants and equal to one another, and hence will cancel, allowing the equation to be rewritten as (12). This equation is used in the next section to determine how likely it is that each possible colour is correctly labelled with the colour term under consideration.

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

4.2.5 Generalising to New Colours

The model uses Bayesian inference, not to determine the probability of any individual hypotheses, but to determine how likely it is that any colour which may be of interest is correctly labelled by a particular colour term (that is how likely it is that the colour comes within the denotation of the colour term). For a location in the colour space, *x*, we can express the probability that it is within the scope of a colour term *C*, given that we assume that hypothesis h_i is correct, using the expression $P(x \in C \mid h_i)$. When the location *x* is within the range of the hypothesis, then this expression will have a value of one, as if a hypothesis is correct, then all hues with its range come within the denotation of the colour term. However, when the location *x* is outside of the range of the hypothesis, then the expression will have a value of zero, as if the hypothesis is correct, then all locations outside of its range do not come within the denotation of the corresponding colour term.

However, what remains to be done is to derive a final expression for the probability that a point in the colour space is within the denotation of a colour term given, not a particular hypothesis, h_i, but given all the observed data, *d*. In order to do this, it is necessary to use the standard Bayesian procedure of hypothesis averaging. The probability that a location in the colour space is within the denotation of the colour term is given by summing the probability of this given each individual hypothesis multiplied by the probability of each hypothesis given the data. As was noted above, hypotheses are continuous, so calculus is used to perform this summation, which is expressed mathematically in (13), where *H* is again the set of all possible hypotheses.

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

This formula completes the specification of the Bayesian model of acquisition, by specifying how a person learning a language can learn to predict which colours come within the denotation of a particular colour term, when their only information about this is a finite number of examples of the use of a colour term.

4.2.6 More than One Colour Term

So far the model has been described with respect to only a single colour term. This is because each colour term which a child tries to learn is considered independently of all the others. Every example colour will be remembered in memory in exactly the same way, but each will be paired with a corresponding colour term. In considering the denotation of each colour term, account will be taken only of those colour examples paired with it. This means that, for every possible colour, a probability can be obtained that it is within the denotation of each colour term encountered by the model. Hence, it is possible that the model will predict that a particular colour is very likely to come within the denotation of more than one colour term, or that it is very unlikely to come within the denotation of

any colour terms at all. However, the model will never conclude with absolute certainty that any colour is outside of the range of the denotation of any particular colour term.

However, it is important to note that considering colour terms one at a time is an assumption, and that alternative learning strategies are possible. The most obvious modification of the strategy proposed here in this respect, is that a learner might reason that if a particular part of the colour space comes within the denotation of one colour, then it is less likely to come within the denotation of another colour. For example, if a learner were to believe that a colour came within the denotation of a term *toki*, they might then consider it unlikely that that colour would come within the denotation of any other term. A stronger version of this would state that a learner might assume that the denotations of basic colour terms in a particular language do not overlap at all, and so if a colour is within the denotation of one colour term, then it will not come within the denotation of another colour term.

Beliefs of this kind, that the denotations of colour terms do not overlap might be supported by inferences made from the linguistic context of a colour term. A learner might make inferences as to whether colour terms overlap one another, based on pragmatic interpretation of observed dialogs. For example, if a learner hears a conversation in which the first speaker says 'Is that balloon purple?' and a second person replies 'No it's blue.', they might take this as a cue that the denotations of red and orange do not overlap, and so consider the denotation of each of these terms when inferring the denotation of the other. However, even in this case it seems that this assumption may be incorrect, as would be the case if in the above question the first speaker had asked 'Is that balloon turquoise?', but received the same reply. This is because the denotations of the colour terms *turquoise* and *blue* overlap, with some colours being both turquoise and blue. So it can be seen that input of this sort is not reliable in providing evidence that the denotations of colour terms do not overlap.

However, it is at least a very common property of languages that basic colour terms in a language not only do not overlap, but also partition[6] the colour space. (This issue is discussed in detail by Kay and Maffi (1999).) The reason that *turquoise* and *blue* overlap, is because *turquoise* is not a basic colour term, but a secondary colour term. The denotations of secondary colour terms can overlap with that of more than one basic colour term, as is the case with *turquoise*, which overlaps the denotations of both *blue* and *green*. Alternatively secondary colour terms may denote a range of colour within the denotation of a basic colour term, as is the case with *crimson*, which denotes colours within the scope of the denotation of *red*. What this implies is that, if a language learner were able to reliably identify which colour terms in a language were basic colour terms, they might be able to learn more effectively by making the assumption that the full set of these terms in the language partitioned the colour space.

Clearly, relying on cues of this sort is a plausible learning strategy, but it does add the difficulty of determining which colour terms are basic, which would seem to be difficult, at least until the acquisition of the colour term system was at a fairly advanced stage. Hence it seems unlikely that adoption of such a strategy can be of major importance in enabling the successful acquisition of colour terms. Further, such a learning strategy is then unable to account for how secondary colour terms are acquired, as secondary colour terms certainly do not partition the colour space. Even greater difficulties for such an approach are caused by a small minority of languages in which the basic colour terms do not appear to partition the colour space, some colours having no corresponding linguistic label (Kay and Maffi, 1999).

[6] By *partition* it is meant that the whole of the colour space is divided up so that each colour is denoted by exactly one colour term.

4.3 Deriving Fuzzy Sets using Bayesian Inference

As described up to the present point, the model simply determines the probability that a specific colour comes within the denotation of a particular colour term. However, so far no consideration has been given to how a semantic representation of a colour term might be derived from the Bayesian model. However, the Bayesian model implicitly defines a fuzzy set representation for the denotation of colour terms.

Instead of considering the probability that individual colours are denoted by a particular colour term, we can consider, for each colour in the phenomenological colour space, the probability that it is within the denotation of the colour term. This will result in a probability for each colour that it can be denoted by the colour term of interest, and these values can be interpreted as specifying the degree of membership of the colour in the semantic category labelled by the colour term. Hence these values may be used to define fuzzy membership in sets corresponding to each colour term.

There are a number of interesting properties of these fuzzy sets, most obviously that for each set, and hence for each colour term, some colours will be members with a greater certainty than other colours. However, there will be a probability associated with the membership of every colour in every set, so that, while it will be considered that some colours are almost certainly not members of the set, there will always be a small probability associated with the possibility that they are members. The implications of these properties of the colour term's denotations is discussed in detail below, but now I move on to consider how the model may be practicably implemented on a computer.

5 Implementation of the Model

While section 4 specifies the model in detail, it does not discuss how the model can be implemented in practice. In particular, equations (12) and (13) both contain integrations, but it remains to be shown that the relevant parts of these equations can in fact be integrated in practice, and how these integrated equations may be use to determine how likely it is that each colour of interest comes within the denotation of each colour term.

5.1 Calculating the Probability that a Colour is within the Denotation of a Colour Term

If we substitute the right hand side of equation (12) for the term $P(h_i \mid d)$ in (13), we obtain the equation (14).

(14) $$P(x \in C \mid d) = \int_{h_i \in H} P(x \in C \mid h_i) \frac{P(d \mid h_i)}{\int_{h \in H} P(d \mid h) dh} dh_i$$

The value of the integral of $P(d \mid h)$ with respect to dh is not dependent on the value of dh_i, and so is a constant term in the integration with respect to this variable. This allows equation (14) to be rewritten as (15).

(15) $$P(x \in C \mid d) = \frac{\int_{h_i \in H} P(x \in C \mid h_i) P(d \mid h_i) dh_i}{\int_{h \in H} P(d \mid h) dh}$$

As the term $P(x \in C \mid h_i)$ evaluates to one in the case where x comes within the denotation of the hypothesis h_i, and to zero in other cases, the top half of the fraction in (15), is effectively a sum over $P(d \mid h)$ for all ranges of the space of possible hypotheses in which x comes within the denotation of the hypothesis. In contrast the term on the bottom of (15) is a sum over $P(d \mid h)$ throughout the hypothesis space, regardless of whether x comes within the denotation of those hypotheses or not. Hence the equation may be written as in (16), where *Px* corresponds to the sum over $P(d \mid h)$ for hypotheses including x in their denotations, and *Pnotx* corresponds to the sum over this same term for hypotheses not including x in their denotations. The program will hence implement (14) by calculating *Px* and *Pnotx*, and substituting their values into (16).

(16) $$P(x \in C \mid d) = \frac{Px}{Px + Pnotx}$$

5.2 Derivation of the Integrals

Now that the task of determining the probability that a colour comes within the denotation of a colour term has been reduced to determining values of sums of $P(d \mid h)$ over specific ranges of hypotheses, and substituting these values into equation (16), it is necessary to consider how these sums can be calculated in practice. Consider again equation (9), repeated here as (17). This formula contains the symbols p, n, m, s, and e. What it is important to determine for the process of integration, is whether these values are constant across different hypotheses, h, or for those terms which are variables, in what way they will change.

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

First of all it may be noted that p is a constant term throughout the model. n and m correspond to how many of the example colours come within the range of the hypothesis under consideration, and how many outside of it. Hence these values will vary between hypotheses, depending on which colour examples each hypothesis includes in its range, and which it excludes. However, these values will change discretely at fixed points in the hypothesis space, and so integrations over the space of hypotheses cannot include hypotheses for which the corresponding values of n and m would vary. Integrating over the whole of the hypothesis space would only be possible if an equation relating the values of n and m to the hypotheses could be substituted for these values. Hence the value of the sum over $P(d \mid h)$ will have to be calculated in sections, with the values of the sum for different values of n and m considered separately.

At this point it is also noting that we wish to derive separately the sum over $P(d \mid h)$ for those hypotheses which include the point of interest, x, and those which do not. This is also a property which will change discretely at points in the hypothesis space, and so similarly it will be necessary to consider sums over areas of the hypothesis space where x is included in the range of the hypothesis separately to those where x is not included in the range of the hypothesis.

The final two terms to consider are s and e, which correspond to the location in the hypothesis space of the start and the end of a hypothesis. These values define the particular hypothesis under

consideration, and so when we sum over areas of the hypothesis space, we are in fact summing over equation (17) for ranges of the variables *s* and *e*. Recall from section 4.2.1 that the hypothesis space, *H*, is composed of hypotheses which may start at any point in the colour space, and may end at any point. Hence, when we sum over areas of the colour space, we must sum over equation (17) for a range of values of *s*, and for each value of *s*, for a range of values of *e*. Recall also from section 4.2.1 that for the purposes of implementation the variable specifying the end of a hypothesis, *e*, will always have a value greater than that specifying the start, *s*. So in some cases the value of *e* will be greater than the size of the colour space (100), in cases where the hypotheses include the origin in their range. In fact, in some cases we will consider continuous ranges of hypotheses where the range of start values also crosses the origin, and in these cases the upper limit of *s* will also be greater than 100, to indicate a location in the colour space clockwise from the origin.

As summing over ranges of the hypothesis space requires summing over ranges of two separate variables, this must be implemented using a double integration. (18) expresses how the probability of the data given a range of hypothesis is determined by summing over a range of the hypothesis space. However the expression $h \in H$ seen in earlier equations, indicating that the sum takes place over the full range of hypotheses, is replaced by two separate integrals specifying that the sum be taken over a specific range of hypotheses, which here are represented collectively as H_i. This specific area of the hypothesis space, H_i, contains the range of hypotheses which start anywhere between the points s_1 and s_2, and which end anywhere between the points e_1 and e_2. Throughout the rest of this paper, I will use H_i to refer to any particular range of hypotheses currently under consideration.

(18) $$\int_{h \in H_i} P(d \mid h)dh = \int_{s_1}^{s_2}\int_{e_1}^{e_2} P(d \mid h)deds$$

If we substitute for the expression *P(d | h)* in (18), using (17), we obtain equation (19), specifying exactly the integration which must be performed in order to derive an equation allowing the sum over the probabilities of a set of data to be calculated for a specific range of the hypothesis space. It is now necessary to perform first the integration over *e*, and secondly the integration over *s*, so as to allow this expression to be evaluated for specific values of the parameters.

(19) $$\int_{h \in H_i} P(d \mid h)dh = \int_{s_1}^{s_2}\int_{e_1}^{e_2}\left(\frac{p}{(e-s)} + \frac{(1-p)}{100}\right)^n \left(\frac{(1-p)}{100}\right)^m deds$$

5.2.1 Integrating Over *e*

We can note first that we will only use integration to sum over areas of the hypothesis space in which *n* and *m* do not change, and hence, for the purposes of integration (over both *e* and *s*), these terms, along with *p*, are all constants. As the value of *s* is not dependent on the value of *e*, it too will be a constant for the purposes of the integration over *e*. As a first step in performing the integration, we may note that a constant term may be removed entirely from the scope of both the integrations, so as to derive equation (20).

(20) $$\int_{h \in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m \int_{s_1}^{s_2}\int_{e_1}^{e_2}\left(\frac{p}{(e-s)} + \frac{(1-p)}{100}\right)^n deds$$

Now it may be noted that the term to be integrated with respect to *e* is a binomial expression, so we can use the binomial expansion, as given in (21) to replace this term. (Note that the expansion of C_n^r

is as specified by equation (22).) If we equate *a* with the first part of the binomial, as in (23), and *b* with the second part as in (24), we can observe the equivalence of the term to be integrated and the left hand side of equation (21).

(21) $$(a+b)^n = \sum_{r=0}^{n} C_n^r a^{n-r} b^r$$

(22) $$C_n^r = \frac{n!}{(n-r)!r!}$$

(23) $$a = \frac{p}{(e-s)}$$

(24) $$b = \frac{(1-p)}{100}$$

When we use expression (21) to substitute for the appropriate term in (20), the result is equation (25).

(25) $$\int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m \int_{s_1}^{s_2}\int_{e_1}^{e_2} \sum_{r=0}^{n} \left(C_n^r \left(\frac{p}{(e-s)}\right)^{n-r} \left(\frac{(1-p)}{100}\right)^r \right) deds$$

By removing constant terms from the scope of the integrations, and removing the discrete summation from the scope of the continuous ones, (25) can be transformed into (26).

(26) $$\int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m \sum_{r=0}^{n} C_n^r p^{n-r} \left(\frac{(1-p)}{100}\right)^r \int_{s_1}^{s_2}\int_{e_1}^{e_2} \left(\frac{1}{(e-s)}\right)^{n-r} deds$$

We may now note that the integration to be performed is straightforward, but that it will have a special case when *n-r* is equal to one. This will be the when case *r* is equal to *n-1*, and so we will use the discrete summation up to only, *n-2*, and then include separate terms for when *r* is equal to *n-1* and when *r* is equal to *n*. The resulting equation, after these terms have been separated out is given in (27).

(27) $$\int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m \left(p^0 \left(\frac{(1-p)}{100}\right)^n \int_{s_1}^{s_2}\int_{e_1}^{e_2} \left(\frac{1}{(e-s)}\right)^0 deds + np^1 \left(\frac{(1-p)}{100}\right)^{n-1} \int_{s_1}^{s_2}\int_{e_1}^{e_2} \left(\frac{1}{(e-s)}\right)^1 deds \right.$$
$$\left. + \sum_{r=0}^{n-2} C_n^r p^{n-r} \left(\frac{(1-p)}{100}\right)^r \int_{s_1}^{s_2}\int_{e_1}^{e_2} \left(\frac{1}{(e-s)}\right)^{n-r} deds \right)$$

Simplifying terms which now are now raised to the power of one, or the power of zero, results in equation (28).

$$\textbf{(28)}\quad \int_{h\in H_i} P(d\mid h)dh = \left(\frac{(1-p)}{100}\right)^m \left(\left(\frac{(1-p)}{100}\right)^n \int_{s_1}^{s_2}\int_{e_1}^{e_2} deds + np\left(\frac{(1-p)}{100}\right)^{n-1} \int_{s_1}^{s_2}\int_{e_1}^{e_2} \frac{1}{(e-s)}\, deds \right.$$
$$\left. + \sum_{r=0}^{n-2} C_n^r\, p^{n-r} \left(\frac{(1-p)}{100}\right)^r \int_{s_1}^{s_2}\int_{e_1}^{e_2} \left(\frac{1}{(e-s)}\right)^{n-r} deds \right)$$

We must note, however, that when the binomial expansion was separated into three separate parts, an implicit assumption was made that *n* was equal to or greater than two, so that elements of the summation could be separated for the cases when *n* was equal to one, and when *n* was equal to zero, and that this would still leave at least one greater value of *n* which would be covered with the summation. However, in some cases *n* will be equal to one, or to zero, which will be so when the range of hypotheses under consideration contains only a single colour example, or no colour examples at all. Substituting the value of zero for *n* in equation (26) results in equation (29), which is the equation to be integrated in the case that the hypotheses span no colour term examples.

$$\textbf{(29)}\quad \int_{h\in H_i} P(d\mid h)dh = \left(\frac{(1-p)}{100}\right)^m \int_{s_1}^{s_2}\int_{e_1}^{e_2} deds$$

Substituting the value of one for *n* in equation (26), results in equation (30), which is the equation to be integrated in the case that the hypotheses span only a single colour term example.

$$\textbf{(30)}\quad \int_{h\in H_i} P(d\mid h)dh = \left(\frac{(1-p)}{100}\right)^m \left(\frac{(1-p)}{100} \int_{s_1}^{s_2}\int_{e_1}^{e_2} deds + p \int_{s_1}^{s_2}\int_{e_1}^{e_2} \frac{1}{(e-s)}\, deds \right)$$

We may note that equation (29) contains an instantiation of the term corresponding to the first integral in (28) and (30) contains a sum of instantiations of the first and second integrals in the same equation. Hence, rather than proceeding with the integration of these equations separately to that of the equation for when *n* is greater than or equal to one, they will be derived by substituting in the appropriate value of *n* to the terms in the final integrated form of equation (28). Hence we now proceed with the integration of this equation, first rewriting the final integral as in (31), so as to make the integration more transparent.

$$\textbf{(31)}\quad \int_{h\in H_i} P(d\mid h)dh = \left(\frac{(1-p)}{100}\right)^m \left(\left(\frac{(1-p)}{100}\right)^n \int_{s_1}^{s_2}\int_{e_1}^{e_2} deds + np\left(\frac{(1-p)}{100}\right)^{n-1} \int_{s_1}^{s_2}\int_{e_1}^{e_2} \frac{1}{(e-s)}\, deds \right.$$
$$\left. + \sum_{r=0}^{n-2} C_n^r\, p^{n-r} \left(\frac{(1-p)}{100}\right)^r \int_{s_1}^{s_2}\int_{e_1}^{e_2} (e-s)^{r-n}\, deds \right)$$

When the three integrations with respect to *e* are performed, the result is as given in (32).

$$\textbf{(32)}\ \int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m \left(\left(\frac{(1-p)}{100}\right)^n \int_{s_1}^{s_2} [e]_{e_1}^{e_2}\, ds + np\left(\frac{(1-p)}{100}\right)^{n-1} \int_{s_1}^{s_2} [\ln(e-s)]_{e_1}^{e_2}\, ds \right.$$

$$\left. + \sum_{r=0}^{n-2} C_n^r \frac{p^{n-r}}{r-n+1} \left(\frac{(1-p)}{100}\right)^r \int_{s_1}^{s_2} \left[\left(\frac{1}{e-s}\right)^{n-r-1} \right]_{e_1}^{e_2} ds \right)$$

This completes the integration over the variable *e*. However, before the integration over *s* is performed, it is necessary to consider what the values of the limits on the integration will be, as the result of the second integration will depend on whether e_1 and e_2 are constants with respect to the value of *s*.

5.2.2 Identifying the Limits on the Integration

The nature of the limits on the integrations will be considered with reference to an example of generalising from a specific set of colour examples, to a previously unlabeled colour. Figure 5 below shows a representation of the phenomenological colour space on which are shown the colours associated with three instances of the use of the colour term *yellow*, along with the point, *x*, about which the probability of it being within the denotation of the term *yellow* will be calculated.

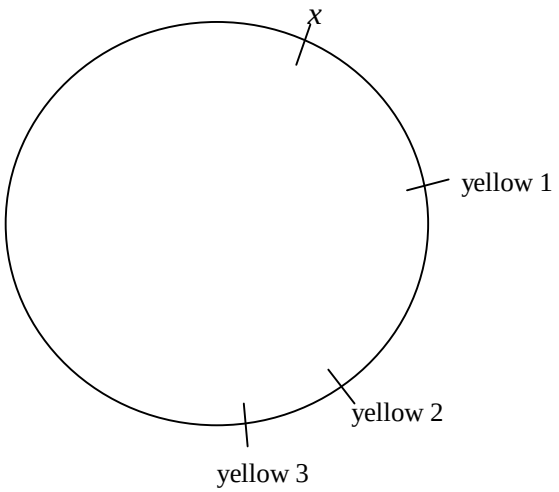

Figure 5. The Phenomenological Colour Space with Observed Colour Term Examples

The values of the terms *n* and *m* will differ depending on how many examples of the colour term *yellow* are within the scope of the hypothesis. Hence, when we consider continuous ranges of the hypothesis space, these values will change as either the position of the start or the end of the hypothesis being considered passes one of these points. Similarly, as the position of the start or the end of the hypothesis crosses the point *x*, the property of whether the hypothesis includes or excludes the point about which we wish to calculate the probability of it coming within the denotation of yellow changes. As the integral will not work over areas of the hypothesis space which have different values for the parameters *n* and *m*, and we wish to find separately the probability of the data summed over all hypotheses including *x*, and all hypotheses excluding *x*, we cannot use the equation for ranges of the colour space which cross the location of a colour example, or the point *x*. However, as we wish to consider the probability of the data given all the hypotheses, but, for reasons of efficiency and simplicity, making as few calculations as possible, with each calculation, we will always consider as large a range of hypotheses as is possible within these constraints. In most cases this will

involve setting the limits on the integration to points in the colour space either where there is an instance of an example of the colour term, or at the location of the point *x*.

As an example, we can consider calculating the probability of the data given all of the hypotheses which include the example colours labelled *yellow 1* and *yellow 2* in Figure 5 above, but which exclude the point *x*, and the example labelled *yellow 3*. This is a case of calculating the value of $\int_{h \in H_i} P(d \mid h) dh$, where H_i corresponds to this range of hypotheses. These hypotheses are those the starts of which are after *x*, but before *yellow 1*, and which have their ends after *yellow 2*, but before *yellow 3*. Hence the values on the limits on the integrals will correspond to these four points. The first position at which a hypothesis may start, s_1, will be the location of *x*, and the last position at which a hypothesis may start, s_2, will be the location of *yellow 1*. The first position at which hypotheses may end, e_1, will be *yellow 2*, and the last position at which hypotheses may end, e_2, will be *yellow 3*. When these values are substituted into the final form of the equation, its evaluation will determine the value of $\int_{h \in H_i} P(d \mid h) dh$.

In the above example, the limits on both *s* and *e* were all constants. This will be the case whenever there is at least one colour example or the point *x* separating the range of positions in which the hypothesis may start, from the range of positions in which it may end. The integration in such cases is completed in section 5.2.3. However there are some instances in which this condition does not hold, so the following paragraphs consider integration in these circumstances.

Let us now consider another example, that of calculating $\int_{h \in H_i} P(d \mid h) dh$ when the set of hypotheses under consideration, H_i, corresponds to a case where there is neither a colour example, nor *x* separating the range of the starts of hypotheses and the ends of hypotheses. This will be so whenever we consider hypotheses that contain all the colour examples and the point *x*. If we consider this with respect to Figure 5, there are four separate ranges of such hypotheses. Those which both start and end between *x* and *yellow 1*, between *yellow 1* and *yellow 2*, between *yellow 2* and *yellow 3* and between *yellow 3* and *x*. Considering as an example the case of hypotheses starting and ending between *x* and *yellow 1*, it might at first seem that the value of s_1 would be *x* and s_2 *yellow 1*, and that e_1 would be *x*+100 and e_2 *yellow 1*+100. (100 would be added to these latter values to indicate that the hypothesis would go all the way around the colour space and past the origin.) However, it is clear that these values are not correct when we consider the hypothesis which starts at the earliest possible point, *x*, and finishes at the latest, *yellow 1*. This hypothesis would go all the way around the colour space from *x*, but the last section, from *x* to *yellow 1*, would overlap itself. This is problematic, as the hypothesis is larger than the whole of the colour space, and includes some of the colours twice, which does not have a coherent meaning within the model.

What is wrong with the above values on the limits of the integration, is that they do not take account of the fact that the end of the hypothesis may not appear more than one full circle around the hypothesis space from the start. It is easy to incorporate this restriction into the limits on the integration by setting the upper limit on the end of the hypotheses, e_2, to *s*+100, rather than to *yellow 1*+100. Now the end of a hypothesis may reach all the way to *yellow 1* only in the case that this is also exactly where the hypothesis started. Incorporating these new limits on integration into equation (32) results in equation (33).

(33) $$\int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m \left(\left(\frac{(1-p)}{100}\right)^n \int_{s_1}^{s_2} [e]_{s_1+100}^{s+100}\, ds + np\left(\frac{(1-p)}{100}\right)^{n-1} \int_{s_1}^{s_2} [\ln(e-s)]_{s_1+100}^{s+100}\, ds \right.$$
$$\left. + \sum_{r=0}^{n-2} C_n^r \frac{p^{n-r}}{r-n+1} \left(\frac{(1-p)}{100}\right)^r \int_{s_1}^{s_2} \left[\left(\frac{1}{e-s}\right)^{n-r-1} \right]_{s_1+100}^{s+100} ds \right)$$

However, (33) only applies in the case that all the colour examples are within the range of the hypotheses. Hence, in such cases there will be no colour examples outside of the range of the hypotheses, and so *m* will be equal to zero. Substituting this value into the equation results in (34). Integration with respect to *s* will now produce a different equation, as the upper limit on *e* is now no longer constant with respect to this integration, but is a variable dependent on *s*. The completion of the integrations in this case is presented in section 5.2.4.

(34) $$\int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^n \int_{s_1}^{s_2} [e]_{s_1+100}^{s+100}\, ds + np\left(\frac{(1-p)}{100}\right)^{n-1} \int_{s_1}^{s_2} [\ln(e-s)]_{s_1+100}^{s+100}\, ds$$
$$+ \sum_{r=0}^{n-2} C_n^r \frac{p^{n-r}}{r-n+1} \left(\frac{(1-p)}{100}\right)^r \int_{s_1}^{s_2} \left[\left(\frac{1}{e-s}\right)^{n-r-1} \right]_{s_1+100}^{s+100} ds$$

The only other situation in which there is neither a colour example nor the point *x* separating the range of possible values for the starts of the hypotheses and for the ends, is when there are no colour examples nor the point *x* within the range of the hypotheses. In Figure 5 above, there are four situations which correspond to this case. These are when the whole of each hypothesis is between *x* and *yellow 1*, between *yellow 1* and *yellow 2*, between *yellow 2* and *yellow 3*, or between *yellow 3* and *x*. We can first consider as an example the hypotheses within the part of the colour space between *yellow 2* and *yellow 3*. It is not possible to simply set the lower limits on both *s* and *e* to *yellow 2*, and the upper limits on both these same variables to *yellow 3*, as this would allow cases in which the end of the hypothesis came before the start.

What is needed, is to constrain the range of permissible endpoints of hypotheses, such that these may only occur in positions between the start of the hypothesis and *yellow 3*. This can be achieved by making the lower limit on *e* equal to *s*. Making this change to equation (32) results in equation (35).

(35) $$\int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m \left(\left(\frac{(1-p)}{100}\right)^n \int_{s_1}^{s_2} [e]_{s}^{s_2}\, ds + np\left(\frac{(1-p)}{100}\right)^{n-1} \int_{s_1}^{s_2} [\ln(e-s)]_{s}^{s_2}\, ds \right.$$
$$\left. + \sum_{r=0}^{n-2} C_n^r \frac{p^{n-r}}{r-n+1} \left(\frac{(1-p)}{100}\right)^r \int_{s_1}^{s_2} \left[\left(\frac{1}{e-s}\right)^{n-r-1} \right]_{s}^{s_2} ds \right)$$

We may note, however, that in such cases, as there are no colour examples within the range of the hypotheses, *n* will be equal to zero. Hence, in this case only the first integration should be included. Making this change, and setting *n* equal to zero, results in equation (36). The integration over *s* in this case is presented in section 5.2.5.

(36) $$\int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m \int_{s_1}^{s_2} [e]_s^{s_2}\, ds$$

These three cases for different types of limits on the values of *e* cover all the possible limits on the integrations, so it is now possible to proceed with the integration of equations (32), (34) and (35) with respect to *s*.

5.2.3 Integrating over *s* when the limits of *e* are constants

This section completes the derivation of an equation for $\int_{h\in H_i} P(d \mid h)dh$ when the range of locations for the end of the hypotheses is separated from the range of locations for the start by at least one colour point or the point *x*, and hence the limits on the integrations are all constants. Firstly, the limits on the value of *e* in equation (32) are substituted in to give equation (37).

(37) $$\int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m \left(\left(\frac{(1-p)}{100}\right)^n \int_{s_1}^{s_2} e_2 - e_1 ds + np\left(\frac{(1-p)}{100}\right)^{n-1} \int_{s_1}^{s_2} \ln(e_2 - s) - \ln(e_1 - s) ds \right.$$
$$\left. + \sum_{r=0}^{n-2} C_n^r \frac{p^{n-r}}{r-n+1} \left(\frac{(1-p)}{100}\right)^r \int_{s_1}^{s_2} \left(\frac{1}{e_2 - s}\right)^{n-r-1} - \left(\frac{1}{e_1 - s}\right)^{n-r-1} ds \right)$$

Rewriting the third integration results in (38).

(38) $$\int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m \left(\left(\frac{(1-p)}{100}\right)^n \int_{s_1}^{s_2} e_2 - e_1 ds + np\left(\frac{(1-p)}{100}\right)^{n-1} \int_{s_1}^{s_2} \ln(e_2 - s) - \ln(e_1 - s) ds \right.$$
$$\left. + \sum_{r=0}^{n-2} C_n^r \frac{p^{n-r}}{r-n+1} \left(\frac{(1-p)}{100}\right)^r \int_{s_1}^{s_2} (e_2 - s)^{r-n+1} - (e_1 - s)^{r-n+1} ds \right)$$

We should note that the integration of the third term will have a special case when *r* is equal to *n-2*, and hence it is necessary to separate this case out of the discrete summation. Making this change results in equation (39). Rewriting the equation in this way assumes that *n* is greater to or equal to three, so it will be necessary to consider another special case of the equation for when *n* is equal to two.

(39) $$\int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m \left(\left(\frac{(1-p)}{100}\right)^n \int_{s_1}^{s_2} e_2 - e_1 ds + np\left(\frac{(1-p)}{100}\right)^{n-1} \int_{s_1}^{s_2} \ln(e_2 - s) - \ln(e_1 - s) ds \right.$$

$$- \frac{n(n-1)}{2} p^2 \left(\frac{1-p}{100}\right)^{n-2} \int_{s_1}^{s_2} (e_2 - s)^{-1} - (e_1 - s)^{-1} ds$$

$$\left. + \sum_{r=0}^{n-3} C_n^r \frac{p^{n-r}}{r-n+1} \left(\frac{(1-p)}{100}\right)^r \int_{s_1}^{s_2} (e_2 - s)^{r-n+1} - (e_1 - s)^{r-n+1} ds \right)$$

Performing the integrations results in (40).

(40) $$\int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m \left(\left(\frac{(1-p)}{100}\right)^n \left[e_2 s - e_1 s\right]_{s_1}^{s_2} \right.$$

$$+ np\left(\frac{(1-p)}{100}\right)^{n-1} \left[(e_1 - s)\ln(e_1 - s) - (e_2 - s)\ln(e_2 - s)\right]_{s_1}^{s_2}$$

$$+ \frac{n(n-1)}{2} p^2 \left(\frac{1-p}{100}\right)^{n-2} \left[\ln(e_2 - s) - \ln(e_1 - s)\right]_{s_1}^{s_2}$$

$$\left. + \sum_{r=0}^{n-3} C_n^r \frac{p^{n-r}}{(r-n+1)(r-n+2)} \left(\frac{(1-p)}{100}\right)^r \left[\left(\frac{1}{e_1 - s}\right)^{n-r-2} - \left(\frac{1}{(e_2 - s)}\right)^{n-r-2} \right]_{s_1}^{s_2} \right)$$

When the integrals are expanded, by substituting the values of the limits on the integration for *s*, the resulting equation is given in (41). This equation is correct when there are at least three colour examples within the range of the hypotheses being considered, but we also need to derive equations for when there are only two colour examples, when there is only a single colour example, or no colour examples at all, within the range of the hypotheses being considered.

$$\textbf{(41)}\ \int_{h\in H_i} P(d\mid h)dh = \left(\frac{(1-p)}{100}\right)^m \left(\left(\frac{(1-p)}{100}\right)^n (e_2s_2 - e_1s_2 - e_2s_1 + e_1s_1) + np\left(\frac{(1-p)}{100}\right)^{n-1}\right.$$

$$\left((e_1 - s_2)\ln(e_1 - s_2) - (e_2 - s_2)\ln(e_2 - s_2) + (e_2 - s_1)\ln(e_2 - s_1) - (e_1 - s_1)\ln(e_1 - s_1)\right)$$

$$+ \frac{n(n-1)}{2} p^2 \left(\frac{1-p}{100}\right)^{n-2} \left(\ln(e_2 - s_2) + \ln(e_1 - s_1) - \ln(e_2 - s_1) - \ln(e_1 - s_2)\right)$$

$$+ \sum_{r=0}^{n-3} C_n^r \frac{p^{n-r}}{(r-n+1)(r-n+2)} \left(\frac{(1-p)}{100}\right)^r$$

$$\left.\left(\left(\frac{1}{e_1 - s_2}\right)^{n-r-2} - \left(\frac{1}{(e_2 - s_2)}\right)^{n-r-2} + \left(\frac{1}{(e_2 - s_1)}\right)^{n-r-2} - \left(\frac{1}{e_1 - s_1}\right)^{n-r-2}\right)\right)$$

Firstly an equation will be derived for the case where there are no example colours within the range of the hypotheses, and hence *n* is equal to zero. As noted above, this equation will contain only a term corresponding to the first integral. The second integral is applicable only when *n* is greater than or equal to one, the third when *n* is greater than or equal to two, and the fourth when *n* is greater than or equal to three. When only the first integral, and the constants it is multiplied by, are included in the equation, and the value of zero substituted in for *n*, the result is equation (42). This equation applies in all instances where the limits on integration are constants, and the range of hypotheses contains no colour examples. As the range of start and end points for the hypotheses must be separated by discontinuities, the discontinuity after the start but before the end must be the point *x*, as there cannot be a colour example within this range.

$$\textbf{(42)}\ \int_{h\in H_i} P(d\mid h)dh = \left(\frac{(1-p)}{100}\right)^m (e_2s_2 - e_1s_2 - e_2s_1 + e_1s_1)$$

Next an equation will be derived for the case that there is only a single colour example within the range of the hypotheses. As noted above this equation can be derived by including only the first two integrals. Taking the corresponding integrals from equation (41), and substituting the value one for *n*, results in equation (43). This equation applies in all cases where the limits on integration are constants, and the hypotheses contain a single colour example.

$$\textbf{(43)}\ \int_{h\in H_i} P(d\mid h)dh = \left(\frac{(1-p)}{100}\right)^m \left(\left(\frac{(1-p)}{100}\right)(e_2s_2 - e_1s_2 - e_2s_1 + e_1s_1)\right.$$

$$\left. + p\left((e_1 - s_2)\ln(e_1 - s_2) - (e_2 - s_2)\ln(e_2 - s_2) + (e_2 - s_1)\ln(e_2 - s_1) - (e_1 - s_1)\ln(e_1 - s_1)\right)\right)$$

Finally it is necessary to derive one more equation, for the case when the hypotheses include exactly two colour examples within their range. Here we take the first three integrals of equation (41), and set *n* equal to two. This results in equation (44), which applies in all cases where the limits on integration are constants, but the hypotheses contain two colour examples.

$$\textbf{(44)}\ \int_{h\in H_i} P(d\mid h)dh = \left(\frac{(1-p)}{100}\right)^m \left(\left(\frac{(1-p)}{100}\right)^2 (e_2 s_2 - e_1 s_2 - e_2 s_1 + e_1 s_1) + 2p\frac{(1-p)}{100}\right.$$

$$\left((e_1 - s_2)\ln(e_1 - s_2) - (e_2 - s_2)\ln(e_2 - s_2) + (e_2 - s_1)\ln(e_2 - s_1) - (e_1 - s_1)\ln(e_1 - s_1)\right)$$

$$\left. + p^2\left(\ln(e_2 - s_2) + \ln(e_1 - s_1) - \ln(e_2 - s_1) - \ln(e_1 - s_2)\right)\right)$$

Four final equations have now been derived, equations (41), (42), (43) and (44). These hypotheses cover all the cases in which the range of start values and the range of hypotheses are separated by at least one colour example or the point *x*, and will be used in the final implementation of the model.

5.2.4 Integrating over *s* when the upper limit on *e* is dependent on *s*

In this section, equations are derived for situations in which the hypotheses under consideration include all of the colour examples and the point *x*. First, substitutions are made of the limits on *e* in equation (34), which results in (45).

$$\textbf{(45)}\ \int_{h\in H_i} P(d\mid h)dh = \left(\frac{(1-p)}{100}\right)^n \int_{s_1}^{s_2} s - s_1 ds$$

$$+ np\left(\frac{(1-p)}{100}\right)^{n-1} \int_{s_1}^{s_2} \ln(100) - \ln(s_1 + 100 - s)ds$$

$$+ \sum_{r=0}^{n-2} C_n^r \frac{p^{n-r}}{r-n+1}\left(\frac{(1-p)}{100}\right)^r \int_{s_1}^{s_2}\left(\frac{1}{100}\right)^{n-r-1} - \left(\frac{1}{s_1 + 100 - s}\right)^{n-r-1} ds$$

As was the case for the integrations when the limits on *e* were both constants, the integration of the term in the discrete summation will have a special case when *r* is equal to *n-2*, and so this case must be separated out of the summation so as to enable the integration to be performed. Making this change results in equation (46).

$$\textbf{(46)}\ \int_{h\in H_i} P(d\mid h)dh = \left(\frac{(1-p)}{100}\right)^n \int_{s_1}^{s_2} s - s_1 ds$$

$$+ np\left(\frac{(1-p)}{100}\right)^{n-1} \int_{s_1}^{s_2} \ln(100) - \ln(s_1 + 100 - s)ds$$

$$- \frac{n(n-1)}{2} p^2 \left(\frac{1-p}{100}\right)^{n-2} \int_{s_1}^{s_2} \frac{1}{100} - (s_1 + 100 - s)^{-1} ds$$

$$+ \sum_{r=0}^{n-3} C_n^r \frac{p^{n-r}}{r-n+1}\left(\frac{(1-p)}{100}\right)^r \int_{s_1}^{s_2}\left(\frac{1}{100}\right)^{n-r-1} - \left(\frac{1}{s_1 + 100 - s}\right)^{n-r-1} ds$$

When the integrations over *s* are performed the result is equation (47).

(47) $$\int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^n \left[\frac{s^2}{2} - s_1 s\right]_{s_1}^{s_2}$$

$$+ np\left(\frac{(1-p)}{100}\right)^{n-1} \left[s\ln(100) + (s_1 + 100 - s)\ln(s_1 + 100 - s) + s\right]_{s_1}^{s_2}$$

$$- \frac{n(n-1)}{2} p^2 \left(\frac{1-p}{100}\right)^{n-2} \left[\frac{s}{100} + \ln(s_1 + 100 - s)\right]_{s_1}^{s_2}$$

$$+ \sum_{r=0}^{n-3} C_n^r \frac{p^{n-r}}{r-n+1} \left(\frac{(1-p)}{100}\right)^r \left[s\left(\frac{1}{100}\right)^{n-r-1} + \frac{1}{r-n+2}\left(\frac{1}{s_1 + 100 - s}\right)^{n-r-2}\right]_{s_1}^{s_2}$$

Finally, substituting in the limits on *s* produces equation (48). This is the equation which will be used to calculate $\int_{h\in H_i} P(d \mid h)dh$ for ranges of the hypothesis space which include all of the colour examples and *x*, where there are at least three colour examples. However, we also need to consider the cases when there are only two colour examples, where there is only a single colour example, or when there are no colour examples at all.

(48) $$\int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^n \left(\frac{s_2^2}{2} - s_1 s_2 + \frac{s_1^2}{2}\right)$$

$$+ np\left(\frac{(1-p)}{100}\right)^{n-1} \left((s_2 - s_1 - 100)\ln(100) + (s_1 + 100 - s_2)\ln(s_1 + 100 - s_2) - s_1 + s_2\right)$$

$$- \frac{n(n-1)}{2} p^2 \left(\frac{1-p}{100}\right)^{n-2} \left(\frac{s_2}{100} + \ln(s_1 - s_2 + 100) - \frac{s_1}{100} - \ln(100)\right)$$

$$+ \sum_{r=0}^{n-3} C_n^r \frac{p^{n-r}}{r-n+1} \left(\frac{(1-p)}{100}\right)^r$$

$$\left((s_2 - s_1)\left(\frac{1}{100}\right)^{n-r-1} + \frac{1}{r-n+2}\left(\left(\frac{1}{s_1 - s_2 + 100}\right)^{n-r-2} - \left(\frac{1}{100}\right)^{n-r-2}\right)\right)$$

First the case where hypotheses contain no colour examples, and hence *n* is equal to zero will be considered. The equation will contain only the first integral, and when the value of zero for *n* is substituted in, the result is as given in (49).

(49) $$\int_{h\in H_i} P(d \mid h)dh = \frac{s_2^2}{2} - s_1 s_2 + \frac{s_1^2}{2}$$

However, in the case that equation (49) applies, that is when there are no colour examples within the range of the hypotheses, and hence *n* is equal to zero, no colour examples can have been observed at all, because the hypotheses must include any colour examples which do exist. This means that there

will only be a single discontinuity in the hypothesis space, at the point *x*. So, as we will always consider the largest possible range of the hypothesis space that is possible with each use of an equation, in this case we will consider the range of hypotheses which start anywhere in the range from just after the point *x*, right round to the other side of this same point. Hence, the end of this range will be 100 units after the start, expressible with the equation $s_2=s_1+100$. Substituting into (49) using this equation results in equation (50).

(50) $$\int_{h \in H_i} P(d \mid h)dh = 5000$$

Equation (50) tells us that the sum of the probability of the data given a hypothesis, over the full range of hypotheses containing any single point, *x*, is the same regardless of where in the colour space that point is, and is equal to 5000. This equation may seem to be fairly meaningless, given that it applies only in the case that there isn't any data, but it will be useful in calculating the probability that particular colours may be within the denotation of a colour term in the case that no colour examples have yet been observed.

The second case to consider is when there is only a single colour example. In this case *n* will be equal to one, and the corresponding equation will contain terms corresponding to only the first two integrals in equation (48). When these changes are made to the equation the result is as given in (51).

(51) $$\int_{h \in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)\left(\frac{s_2^2}{2} - s_1 s_2 + \frac{s_1^2}{2}\right) + p\left((s_2 - s_1 - 100)\ln(100) + (s_1 + 100 - s_2)\ln(s_1 + 100 - s_2) - s_1 + s_2\right)$$

Lastly it remains to derive an equation for when there are only two colour examples, and hence *n* is equal to two. This equation will include terms corresponding to the first three integrals of equation (48). Including just these integrals and setting *n* equal to two results in equation (52).

(52) $$\int_{h \in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^2\left(\frac{s_2^2}{2} - s_1 s_2 + \frac{s_1^2}{2}\right) + 2p\left(\frac{(1-p)}{100}\right)\left((s_2 - s_1 - 100)\ln(100) + (s_1 + 100 - s_2)\ln(s_1 + 100 - s_2) - s_1 + s_2\right) - p^2\left(\frac{s_2}{100} + \ln(s_1 - s_2 + 100) - \frac{s_1}{100} - \ln(100)\right)$$

The four final equations derived in this section, (48), (50), (51) and (52) cover all cases in which the range of the hypotheses include all the colour examples and *x*, and will be used in the final computer implementation of the model.

5.2.5 Integrating over *s* when the lower limit on *e* is dependent on *s*

This section derives an equation for $\int_{h \in H_i} P(d \mid h)dh$ for continuous ranges of the hypothesis space which contain no colour examples nor the point *x*. Starting with equation (36) derived above, substituting in the values for the limits on *e* results in equation (53).

(53) $$\int_{h \in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100} \right)^m \int_{s_1}^{s_2} (s_2 - s)ds$$

Integrating over *s* produces equation (54).

(54) $$P(d \mid H_i) = \left(\frac{(1-p)}{100} \right)^m \left[s_2 s - \frac{s^2}{2} \right]_{s_1}^{s_2}$$

And substituting in the limits on *s* results in equation (55). This equation will be used to calculate the value of $\int_{h \in H_i} P(d \mid h)dh$ for all ranges of hypotheses which contain neither any colour examples nor the point *x*.

(55) $$\int_{h \in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100} \right)^m \left(\frac{s_2^2}{2} - s_1 s_2 + \frac{s_1^2}{2} \right)$$

5.3 Applying the Equations

The nine final equations derived in section 5.2 can now be used to calculate $\int_{h \in H_i} P(d \mid h)dh$ for all regions of the hypothesis space. The total of the results of these calculations for all the ranges of the hypothesis space containing *x*, and the total for all the ranges not containing *x* can be calculated. These values can be substituted into equation (16) to determine the probability that the point *x* comes within the denotation of the colour term ($P(x \in C \mid d)$). Calculating this probability for all possible values of *x* will give the degree of fuzzy membership of each hue within the denotation of the colour term, thus defining a fuzzy set.

6 Learning the Denotation of English Colour Terms from Examples

In order to investigate how the model would perform in practice, and what properties the learned denotations of the colour terms would have, the model was trained on the six chromatic basic colour terms of English which are distinguished principally on the basis of their hues. Using data from Berlin and Kay (1969), estimates were made of the width of the phenomenological colour space corresponding to each colour (the width being measured along the hue dimension only), and these estimates were used to map each colour term to a section of the phenomenological colour space in the model, such that these six colours partitioned the colour space. Berlin and Kay show the extent of these colours on an array of Munsell colour chips. As the Munsell system of colour chips attempts to space the chips in phenomenologically even gradations of colour, the number of colour chips within the range of each colour term should at least approximate the size of the phenomenological colour

space denoted by that colour. However, whether or not Berlin and Kay's data is an accurate measure of the size of the denotation of each colour term is not important for present purposes, as the purpose of the current experiments was simply to show how any such colour system can be learned from examples.

Example colours for each colour term were generated within the range of the phenomenological colour space corresponding to each colour. In order to simulate possible inaccuracies in the naming or perception of these colours, each was then randomly adjusted within the range of plus or minus five units. The model was initially trained on five examples of each colour term. The resulting fuzzy denotation of each colour term is represented graphically in Figure 6. The horizontal axis covers the range of hues in the phenomenological colour space, from red to orange, yellow, green, blue and finally back to red. (As the colour space is circular, the left and right edges of the graph represent adjacent points in the colour space.) The vertical axis represents the probability with which it is believed that each hue comes within the denotation of each colour term, ranging from zero, indicating that a hue is definitely not within the range of a colour term, to one, indicating that it definitely is. These values may alternately be interpreted simply as the degree of membership of each hue within the categories corresponding to each colour term.

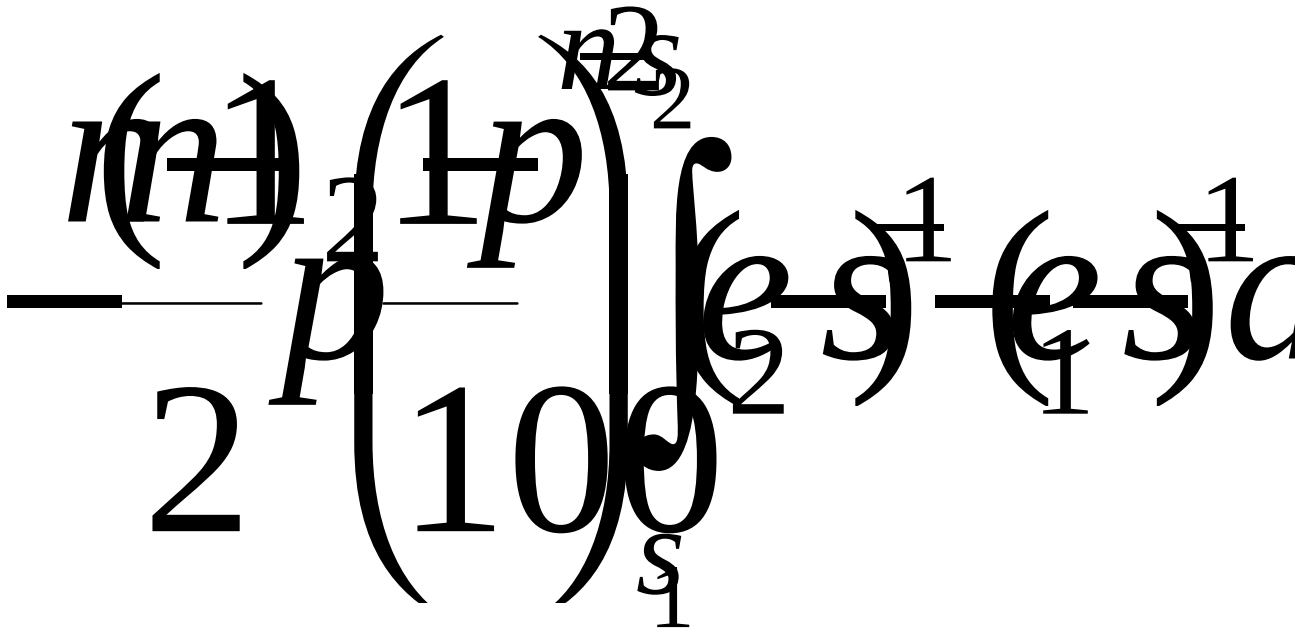

Figure 6. The Fuzzy Denotation of English Basic Colour Terms after 5 Examples

In Figure 6 we can observe some important properties of the learned denotations. Firstly we can observe that each colour category has a prototype structure. Consider for example the curve corresponding to the colour term *yellow*. This curve rises to a single peak near the middle of the graph, with the probability of a hue being a member of this category decreasing the further the hue is away from this point. There is a section of the colour space about which the model is almost certain that it comes within the denotation of *yellow* (where the *yellow* curve is very high), and there is a section of the colour space which the model thinks is unlikely to come within the denotation of *yellow* (where the curve drops low down, getting below 0.2). In between these two areas are colours which may be considered marginal examples of the colour term, especially where the curve is near the 0.5 level, where the model considers it equally likely that a hue may or may not be within the denotation of the colour term.

Hence we can see that the denotations have the key properties of prototype categories. Firstly some members of the category are better examples of it than others are. Secondly there is a single best example of each category. Finally there are marginal members of each category, about which it is difficult to be certain whether or not they are members of the category.

If we look at areas close to the boundaries of two colour terms, for example the boundary between blue and purple towards the right of Figure 6, we can see that their are colours which are considered more likely than not to be a member of more than one colour term. This has occurred because the model has tended to overextend the colour categories, and this effect has been most severe for the smallest categories. This is because *a priori* the model considers the denotation of all colour terms to be equally likely to be of any possible size, but the average size for all possible hypotheses is equal to half the colour space. In contrast there are no hues which are consider more likely than not to be a member of no colour term at all. This has occurred because, given only five examples of each colour term, the model has been influenced to a very large extent by its *a priori* assumptions.

Fifteen more examples of each colour term were added to the model, and the resulting denotations are shown in Figure 7. The main difference between this graph and Figure 6 is that the model is now much more certain about which hues come within the denotation of each colour term, and which do not. There are areas where the curves come very close to the top of the graph, and where they are very flat, because in these areas the model is almost completely certain that the hues come within the denotation of the corresponding colour term. (While on the graph it may look as though these curves are completely flat and that they have reached all the way to the top of the graph, this is simply a consequence of the accuracy with which the graphs have been drawn. Each curve still rises to a single point, and the degree of membership decreases very slowly on each side of this point.)

Figure 7. The Fuzzy Denotations of English Colour Terms after 20 Examples

The model is now also almost completely certain that some hues cannot be labeled with each colour term. This can be seen in the areas of the graph as areas where the curves are very close to zero. As the model now has more data available from which to learn, it is able to determine the correct denotation of the colour terms with a greater degree of accuracy, and so the range of hues about which there is uncertainty about whether they come within the denotation of a colour term is much

smaller. However each category still retains the overall prototype structure, with varying degrees of membership, a single best example, and some marginal examples of the category.

7 Conclusions and Future Directions

The bayesian model described in this paper provides an account of the semantics of basic colour terms. This account is flexible, so that it can account for the wide variation seen in the colour systems of different languages. Not only does it provide an account of the nature of the semantics of colour terms, it also demonstrates how those semantics can be learned by speakers of a language. While the model has so far been used only to account for the acquisition of basic colour terms, there seems no reason why it should not be equally good as a model of the acquisition of secondary colour terms.

Of course, there are aspects of basic colour terms for which the model does not provide an explanation. Probably most importantly the model does not give any preference to learning colour terms with a prototype in one part of the colour space as opposed to any other. This conflicts with experimental evidence from sources such as Rosch (1973) who found that colour terms focussed in certain areas of the colour space were easier to learn than those focussed in other areas. For example, almost all languages have a colour term with a prototype near the prototype of *red* in English. Rosch found that colour terms where this colour was near the centre of the category were easier to learn than colour terms where this colour was near the periphery. Berlin and Kay (1969), and many studies since, have found very marked typological patterns in colour terms across languages. This suggests that there must be some form of universal biases which direct people to have a preference for some colour categories as opposed to others.

Hence it seems that, while the model presented here has the advantage over some other approaches of being flexible enough to account for the full variety of basic colour term systems seen in the world's languages, it is in fact too flexible, as it fails to account for typological patterns, and observed preferences in learning some colours as opposed to others. Work is on progress in investigating how such biases can be introduced into the model, so that it is able to give a fuller account of the phenomenon of basic colour terms across languages.

One hypothesis about why typological patterns are observed in basic colour terms systems is that they are produced as colour term systems evolve over time, and those systems gradually add new colours (Kay and Maffi, 1999). An extension of the current model which will investigate such hypotheses, is to simulate populations of speakers over time as colour terms are acquired over a series of generations. Such computational evolutionary models, based on work such as Kirby (2000), will form a later stage of the current research project.

Overall, the model is a proposal as to the general nature of the process of acquiring the meanings of basic colour terms. It also suggests that prototype effects more generally may be due to similar processes of Bayesian inference. However, further investigations will be needed to investigate how well such an approach can account for the full range of psychological and typological data concerning basic colour terms and prototype categories.

8 Acknowledgements

I would like to thank everyone with whom I have discussed any of the ideas in this paper, but in particular Eva Endrey-Walder, who first suggested to me that adjectives could be given a Bayesian interpretation, Emmanuel Letellier for help with what he described as 'high school level maths',

Yukari Fujiwara for good cooking, and my PhD supervisor, Judy Kay without whose encouragement I would probably never have written this paper in the first place. I am also grateful to the Australian government and the University of Sydney for providing me with IPRS and IPA scholarships respectively.

9 References

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.

Brent, M. R. & Cartwright, T. A. (1997). *Distributional Regularity and Phonotactic Constraints are useful for Segmentation.* In M. R. Brent (Ed.) *Computational Approaches to Language Acquisition*. Cambridge, MA: MIT Press.

Dowman, M. (2000). Addressing the Learnability of Verb Subcategorizations with 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.

Ellison, T. M. (1992). *The Machine Learning of Phonological Structure*. Doctor of Philosophy thesis, University of Western Australia.

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.

Honkela, T. (1997). *Self-organizing Maps in Natural Language Processing*. Doctor of Philosophy thesis, Helsinki University of Technology.

Huttenlocher, J., Hedges, L. V. & Vevea, J. L. (2000). Journal of Experimental Psychology: General, Volume 129, Number 2, pages 220-241.

Kay, P. & McDaniel, K. (1978). The Linguistic Significance of the Meanings of Basic Colour Terms. *Language*, Volume 54, Number 3.

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

Kirby, S. (2000). Syntax without Natural Selection: How Compositionality Emerges from Vocabulary in a Population of Learners. In C. Knight, M. Studdert-Kennedy and J. Hurford (eds.) Cambridge: Cambridge University Press.

Land, E. H. (1977). The Retinex Theory of Colour Vision. *Scientific American*, Volume 237, Number 6, pages 108-128.

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.

Rosch, E. H. (1973). Natural Categories. *Cognitive Psychology*, Volume 4, pages 328-350.

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

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: Routledge.