Colour Terms, Syntax and Bayes
Modelling Acquisition and Evolution

Mike Dowman

February 2004

PhD Thesis,
School of Information Technologies,
Faculty of Science,
University of Sydney

Supervised by Associate Professor Judy Kay

Colour Terms, Syntax and Bayes: Modelling Acquisition and Evolution

Mike Dowman

Abstract

This thesis investigates language acquisition and evolution, using the methodologies of Bayesian inference and expression-induction modelling, making specific reference to colour term typology, and syntactic acquisition. In order to test Berlin and Kay's (1969) hypothesis that the typological patterns observed in basic colour term systems are produced by a process of cultural evolution under the influence of universal aspects of human neurophysiology, an expression-induction model was created. Ten artificial people were simulated, each of which was a computational agent. These people could learn colour term denotations by generalizing from examples using Bayesian inference, and the resulting denotations had the prototype properties characteristic of basic colour terms. Conversations between these people, in which they learned from one-another, were simulated over several generations, and the languages emerging at the end of each simulation were investigated. The proportion of colour terms of each type correlated closely with the equivalent frequencies found in the World Colour Survey, and most of the emergent languages could be placed on one of the evolutionary trajectories proposed by Kay and Maffi (1999). The simulation therefore demonstrates how typological patterns can emerge as a result of learning biases acting over a period of time.

Further work applied the minimum description length form of Bayesian inference to modelling syntactic acquisition. The particular problem investigated was the acquisition of

the dative alternation in English. This alternation presents a learnability paradox, because only some verbs alternate, but children typically do not receive reliable evidence indicating which verbs do not participate in the alternation (Pinker, 1989). The model presented in this thesis took note of the frequency with which each verb occurred in each subcategorization, and so was able to infer which subcategorizations were conspicuously absent, and so presumably ungrammatical. Crucially, it also incorporated a measure of grammar complexity, and a preference for simpler grammars, so that more general grammars would be learned unless there was sufficient evidence to support the incorporation of some restriction. The model was able to learn the correct subcategorizations for both alternating and non-alternating verbs, and could generalise to allow novel verbs to appear in both constructions. When less data was observed, it also overgeneralized the alternation, which is a behaviour characteristic of children when they are learning verb subcategorizations. These results demonstrate that the dative alternation is learnable, and therefore that universal grammar may not be necessary to account for syntactic acquisition. Overall, these results suggest that the forms of languages may be determined to a much greater extent by learning, and by cumulative historical changes, than would be expected if the universal grammar hypothesis were correct.

Preface

The aim of this thesis is to show how linguistic phenomena which have been identified by other researchers can be explained. The results reported in this thesis were obtained by running simulations on computers, and comparing the outputs of these simulations to data reported in published sources. Therefore, the research involved no empirical data collection. This thesis attempts to provide explanations of a number of phenomena which have been identified as a result of empirical investigations, and subsequently reported in the literature. However, while this thesis contains discussions of many empirical findings, and of analyses of empirical data, it is not concerned with these analyses in themselves, or with the correctness of the data on which they are based. For example, much of this thesis relies heavily on Kay and Maffi's (1999) analysis of the data of the World Colour Survey, which is discussed in detail in section 2.1. Kay and Maffi's analysis remains controversial (Levinson, 2001), but I do not attempt to analyse the primary data myself. That would go beyond the scope of this thesis, so instead I simply tried to create a computer model which would account for the emergence of colour term systems corresponding to those which Kay and Maffi reported were attested in the World Colour Survey.

While criticisms have been made of Kay and Maffi's work (many of which are discussed in section 2.1 below), a wealth of empirical data supports their general conclusions. The basis of most of the criticisms seems to be that Kay and Maffi's

analysis does not take account of the full complexity of colour term systems, especially in that it excludes non-basic colour terms, and that it ignores the secondary connotations of basic colour terms. However, neither of those factors would seem to in any way invalidate the results which Kay and Maffi did report. Hence it would seem that, even if some of the details of Kay and Maffi's analyses turn out to be incorrect, that is unlikely to greatly affect the validity of the work presented in this thesis. Many other issues relating to this thesis remain controversial, such as exactly what neurophysiological mechanisms underlie colour vision, and whether the dative alternation can be explained in terms of semantic regularities. However, like with the controversies surrounding colour term typology, resolving these issues concerning empirical findings goes beyond the scope of this thesis. Further empirical findings may necessitate a revision of this work, but at present it is only possible to work on the basis of the results which have been published up to this point.

While the research for this thesis relied entirely on computer modelling, its actual subject matter lies within linguistics, and to some extent related disciplines such as psychology. Hence the thesis has been written so that it should be, as far as possible, understandable by a general linguistic audience. At the same time, I have tried to explicate linguistic terminology whenever possible, for the benefit of non-linguists. In general, where specific technical issues arise, I have tried to make them as accessible as possible, but a description of some aspects of the research necessitates the use of mathematical concepts and related notation which are likely to be unfamiliar to many readers interested in the subject matter of the thesis. This thesis makes considerable use of techniques from statistics and machine learning, in particular Bayesian inference and minimum description length. However, while some attempt is made to justify why these techniques were chosen, actually demonstrating why they are

effective machine learning techniques, or why, for example, Bayes' rule is correct, goes beyond the scope of this thesis. It can be said that, while this thesis uses machine learning, it is not about machine learning itself, and hence I have tried to restrict discussion of the machine learning literature to the minimum necessary.

All of the thesis is my own original work, except where explicit reference is made to other work. However, portions of the thesis have been published, and I have presented results at a number of conferences, and given some other talks. A full list of this research activity appears below (on all of which I am the sole author, except where noted otherwise). In general, the publications do not correspond exactly to particular parts of this thesis, but Chapter 9 is essentially an expanded version of the 2000 Cognitive Science Society conference paper, while parts of the other five papers are reproduced in Chapters 1 to 8.

Publications

2003	Modeling Language as a Product of Learning and Social Interactions. *Cognitive Systems*, Volume 6, issue 1.
2003	Explaining Color Term Typology as the Product of Cultural Evolution using a Multi-agent Model. In *Proceedings of the Twenty-Fifth Annual Meeting of the Cognitive Science Society*. Cognitive Science Society.
2002	Modelling the Acquisition of Colour Words. In B. McKay and J. Slaney (eds.) *Advances in Artificial Intelligence*. Berlin: Springer-Verlag.
2001	A Bayesian Approach to Colour Term Semantics. *Lingu@scene*, Volume 1.
2001	*A Bayesian Approach to Colour Term Semantics*. Technical Report Number 528. Sydney: Basser Department of Computer Science, University of Sydney. (This is a more technical version of the lingu@scene paper above.)
2000	Addressing the Learnability of Verb Subcategorizations with Bayesian Inference. In Gleitman, L. R. & Joshi, A. K. (Eds.) *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*. Mahwah, New Jersey, USA: Lawrence Erlbaum Associates.

Other Conference talks and presentations

2002	Modelling the Evolution of Basic Colour Terms. Talk given as part of the *Language in Time Symposium*, University of Western Australia, 25-27 June.
2002	A Bayesian Model of Colour Term Acquisition and Typology. Talk given at *Australian Linguistics Society Conference*, Macquarie University, Sydney, 13-14 July.
2002	Bayesian Learning of Linguistic Categories. Talk given at *European Society for the Study of Cognitive Systems Special Workshop on Multidisciplinary Aspects of Learning*, Paris, 17-19 January.
2002	*Explaining Language Typology using a Multiple Agent Model*. Talk given at Department of Computer Science and Software Engineering, University of Western Australia, 28 June.
2000	A Bayesian Model of Syntactic Acquisition. Talk given at *Australian Linguistics Society Conference*, University of Melbourne, 7-9 July.
2000	*Learning Verb Subcategorizations - A Case Study for the Acquisition of Syntax*. Departmental Seminar, Department of Linguistics, University of Sydney, 3 March.
1999	A Minimum Encoding Inference Approach to Theoretical Syntax. Talk given at *Australian Machine Learning Workshop*, Australian National University, Canberra, 22-23 November. Co-authored with Jon Patrick.
1999	Syntax: Accounting for Acquisition and Change. Talk given at *Postgraduate Students' Conference*, Department of Linguistics, University of Sydney, 23 September.

During the course of my Ph.D. I have received help from an awful lot of people, and I'd like to thank them all. Some people have helped me out with relevant literature, others have given me feedback on my papers or have had helpful discussions with me about my research. Others have just given me much needed encouragement along the way. I will probably will forget to mention some people who should be here, but I would like to thank Judy Akinbolu, Stephen Anthony, Brett Baker, Tony Belpaeme, Nils Bruin, Cassily Charles, Garry Cottrel, Gerhard Dalenoort, David Dowe, Mark Dras, Michelle Ellefson, Jeff Elman, Eva Endrey-Walder, Bill Foley, Alexander Francis, Roslyn Frank, Yukari Fujiwara, John Goldsmith, Paul Green-Armytage, Tom Griffiths, Joost van Hamel, Catherine Harris, Timo Honkela, Jim Hurford, Simon Kirby, Bill Labov, Johan Lammens, Darren Ler, Andrew Lum, Sinead Lyle, Rob MacLaury, Que Chi Luu, Chris Manning, Vladimir Novikov, Adam Blaxter Paliwala, Jon Patrick, Sam Pickering, David Powers, Hong Liang Qiao, Christine Rakvin, Jacqueline van der Schaaf, Jane Simpson, Kenny Smith, Josie Spongberg, Anders Steinvall, Kimie Takahashi, Tania Tsatralis. I have had three supervisors during the course of my Ph.D., but Judy Kay deserves special thanks as she has supervised most of it, and I probably would never have completed without her help. I should also give special thanks to Mark Ellison, who supervised my undergraduate dissertation (Dowman,1998), and without whom I probably would never have got interested in minimum description length, Bayesian inference, or any other form of computer modelling. And I could not have done some of the hard maths without the help of Emmanuel Letellier. Of course any errors or omissions remain my own responsibility (and I have not always taken all the advice I was given).

My Ph.D. was supported by a scholarship funded by a grant from the Australian Research Council, an International Postgraduate Research Scholarship from the

Australian government, and an International Postgraduate Award from the University of Sydney.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis aims to increase our understanding of language. Perhaps the most important aspect of this process is concerned with understanding exactly what language is, and hence exactly what kind of explanation we need in order to understand it. Many researchers, notably de Saussure (1959/1916) and Chomsky (1986), have noted that the word *language* can mean several different things. Often researchers have focussed on one particular concept of language, and they have sometimes argued that the concept they use is the single correct one for use in linguistics. For example, Chomsky (1986) stressed the importance of viewing languages as psychological phenomena, while other researchers (for example Halliday, 1985) have preferred to highlight the fact that language is a communicative system. If we accept Chomsky's position, linguistics becomes concerned with identifying psychological mechanisms, but if we accept Halliday's, then linguistics is concerned with investigating how language can be used to achieve particular communicative functions. Many researchers, however, have simply focussed on the systems of rules which seem to underlie the examples of language which we can hear and read, and have tried to describe languages in terms of formal systems, without making explicit reference to the people who produce and listen to them, or the

purposes for which they are used. Still other researchers have argued that languages are built up and change over many generations, and so if we want to understand languages we should understand the factors which have caused them to evolve[1] in particular ways. These are just a few of the different perspectives on language which are common in linguistics. In practice, most linguists seem to draw upon aspects of a number of different concepts of language, and often they do not make it explicit exactly which concept of language their analysis relates to.

There is clear justification for using each of the above concepts of language, so it does not seem appropriate to claim that there is a correct approach which should be used in all linguistic investigations. This thesis integrates multiple concepts of language, and shows how quite different approaches can be coherently used to account for a variety of linguistic phenomena. It is shown that, if languages are to be fully understood, it is necessary to consider more than one concept of language. Many of the phenomena which have attracted the attention of linguists, can probably not be understood by considering only one specific concept of language. Much of the work presented in this thesis is based on Hurford's (1987, 1990) conceptualisation of language, which integrates individual and social concepts of language into a single unified theory.

However, as well as investigating which general paradigms are most suitable for explaining linguistic phenomena, this thesis also tests some more specific hypotheses.

---

[1] It should be noted that throughout this thesis, except where I explicitly note otherwise, when I refer to evolution, I mean cultural evolution, that is a process whereby individual languages change over time. This should not be confused with the phylogenetic (biological) evolution of human's linguistic capabilities.

Most importantly, the thesis tries to understand how children learn language. This issue has been central to linguistic theory, at least since Chomsky (1965). We should note that the primary justification for the Universal Grammar hypothesis (the claim that children are born with an innate knowledge of much of the structure of language), is that it has been argued that children need Universal Grammar to learn language. Hence all research into Universal Grammar can be seen as work aimed at developing a theory of language acquisition.

The key methodology of this thesis involves the construction of computer models. These aim to explain language acquisition by demonstrating how the structure of some aspects of a language can be determined. These models learn based on examples of utterances in the language, sometimes together with a representation of the non-linguistic context in which those utterances were made. These models hence specify a particular computation, which it is hypothesized is either the same, or in some way parallels, a computation that people perform in order to learn language. Clearly here I am not referring to some process or algorithm which people consciously carry out in a step by step fashion. The computer models instead aim to recreate a process which is hypothesized to match or in some way parallel that occurring in the brain, but about which people have no conscious awareness. Therefore, while this thesis is not concerned with computers themselves, it is concerned with computation.

More specifically, the thesis investigates the hypothesis that children learn language using Bayesian inference, or at least that Bayesian inference can accurately model children's learning. Bayesian inference is a method of learning which can be applied to almost any situation in which we want to make inferences based on empirical evidence. It is usually an implicit assumption of work in linguistic theory that

languages are not statistical, and this is occasionally argued for explicitly (for example, Chomsky, 1957). In this thesis, I suggest that language is statistical, by which I mean that the frequency with which particular words, constructions, or other aspects of language, occur, forms a part of our internalised knowledge of language. Bayesian inference is a statistical procedure, so employing it as a theory of language acquisition necessitates the adoption of the view that people make probabilistic (and hence statistical) inferences. They can only do this by taking account of the frequency with which particular types of construction are used, or the frequency with which words are used to denote particular meanings. Bayesian inference is a very general method, and so Bayesian modelling can encompass a wide range of very diverse approaches. The models used for explaining language acquisition in this thesis employ two different types of Bayesian inference, firstly Bayes' optimal classification, and secondly minimum description length, but we can see both of these as being motivated by a single hypothesis, which is that people learn language using Bayesian inference.

Most of this thesis is concerned with colour terms: with how they are acquired, how they change over time, how we should explain their typology, and what sort of representation is needed to account for their meanings. Chapter 2 outlines the relevant literature concerning colour terms, and reviews the kinds of approach which have been proposed for explaining linguistic phenomena. It then goes on to review research which has employed the methodology of expression-induction modelling, which I use to explain empirical data concerning colour terms.

Chapter 3 describes a model which was implemented in an attempt to account for colour term acquisition, and then goes on to demonstrate that it can account for how

colour terms are learned. However, in Chapter 4, the model is evaluated in terms of alternative approaches to linguistic theory, and it is shown how the model can only really account for the data concerning colour terms when it is placed in an evolutionary context.

Chapter 5 introduces a new model of the acquisition of colour term systems, which takes account of findings concerning colour vision and colour language, which were neglected in the design of the previous model. In Chapter 6, it is shown how the new model could account for typological data when it was used to simulate the evolution of colour term systems, something which was not possible with the previous model. In Chapter 7, attempts are made to make the simulations more realistic, by investigating whether the model is robust in the face of noisy data, and Chapter 8 evaluates the findings of the models, and discusses how the present models relate to previous attempts to explain the same data. It also discusses the shortcomings of the models, and makes suggestions as to how further work could improve on them.

Chapter 9 shows how Bayesian inference can also solve learnability problems in syntactic acquisition, and discusses the relevance of this to contemporary linguistic theory. Finally, Chapter 10 discusses the relevance of the results to a number of theoretical debates, concerning the nature of language, and what kind of concept is needed to explain linguistic phenomena. Overall, the thesis develops proposals which show how data which has been of interest to linguistic theorists, but which has resisted a coherent explanation in other frameworks, can be explained.

Chapter 2

Background

This first part of this thesis describes computational modelling experiments performed to explain empirical data concerning basic colour terms. Colour terms are simply words in natural languages which are used to denote the property of colour. In most, if not all, languages, a special subset of such words can be identified, which Berlin and Kay (1969) named *basic colour terms*.

Berlin and Kay (1969, p6) listed a number of criteria which they used to distinguish basic colour terms from other words used to denote colour. They considered colour terms to be basic only if they were known by all speakers of the language and were highly salient psychologically, and if they did not just name a subset of the colours denoted by another colour word and their meanings were not predictable from the meanings of their component parts. They also provided some further criteria to deal with any doubtful cases. Application of these criteria seems to distinguish clearly between basic and non-basic colour terms in most languages, although there can still remain some questionable cases[2]. The application of these criteria to English, results

[2] These include the Russian term *goluboy* 'light blue', which seems to be less salient than the other Russian blue term *siniy*, but for some speakers it seems that this term names a range of colours disjunct

in the set of 11 basic colour words, *red, yellow, green, blue, orange, purple, pink, brown, grey, black* and *white*, excluding terms such as *crimson*, *blonde* and *royal blue*[3] (Berlin and Kay, 1969). It is possible that some other words should be considered basic for some speakers (for example *turquoise*, *cream* or *beige*), but all of these terms could be excluded on grounds of salience, each of them being much less frequent than any of the words usually considered to be basic colour terms in English[4].

---

from *goluboy*, at least in some contexts, which is evidence that it should be considered basic. Another problematic case concerns the Hungarian red terms *piros* and *vörös*. MacLaury (1999) discusses these terms, both of which seem to be highly salient. V*örös*, however, names a much smaller range of colour than *piros* does, so it could be seen as a non-basic red term. This is similar to the situation seen in Japanese, where *ao* names both green and blue hues, while *midori* names a narrower range of blue colours, which might suggest that *midori* is a non-basic colour term. However, *midori* is more commonly used to name blue colours than *ao* is, and is also highly salient (Conlan, 2002), so perhaps both terms should be considered basic, which is how they were treated by Berlin and Kay (1969). Some linguists have even gone so far as to question the validity of the distinction between basic and non-basic colour terms (Levinson, 2001; Saunders, 1992), but it seems that the consensus of opinion is that the distinction is valid, at least in the majority of cases. The case of Hungarian *piros* and *vörös* is discussed further in Chapter 8.

[3] It should be noted that the following convention is used concerning linguistic examples: whenever they appear in the text they are italicized. This is especially important when English colour terms are discussed, because it makes it clear whether I am talking about an English word, such as *red*, or a particular colour, such as red.

[4] This is supported by evidence derived from the 100 million word British National Corpus (of which about 90% is written language, and 10% spoken). The least frequent basic colour term in this corpus is *orange*, with only 607 occurrences, but this was much more frequent than all of *beige* (174

There has been a considerable amount of research into the properties of basic terms, but perhaps the most important study was that of Berlin and Kay (1969). They examined a sample of 98 languages, and found that there was very wide variation between the colour terms of different languages, in that the actual ranges of colour denoted by each term differed between languages. However, they found that this variation was certainly not without limit, as had been presumed by earlier researchers (for example Gleason, 1961). While the number of colour terms varied between languages, which combination of colour terms exist in any given language seems to be at least partly predictable.

This thesis attempts to address the issue of why there are such regularities. Berlin and Kay suggested that the regularities were the product of an evolutionary process in which languages gradually evolved from an initial state in which they had only two basic colour terms, and in which more terms were added in predictable orders. One of the goals of this thesis is to investigate whether this hypothesis could explain the typological patterns seen in colour term systems. However, while Berlin and Kay claimed that the patterns were due to an evolutionary process, they left the details of

---

occurrences), *turquoise* (64 occurrences), and *cream* (22 occurrences). Terms were only counted when they had been tagged as adjectives, which would, for instance, exclude *orange* when it was used to name the fruit, but it is possible that these figures where somewhat distorted by use of words to denote properties other than colour (for example the use of *red* to mean radical). However the general finding seems clear, which is that the eleven basic colour terms are much more frequent than other colour words, so even if some other colour words appear to satisfy some of the criteria for basic colour term status, we can exclude them on grounds of salience in the language as a whole. The British National Corpus is available on-line at http://www.natcorp.ox.ac.uk/

this process unspecified. Clearly cultural evolution of language[5] is realized through a process in which language is passed from speaker to speaker over a number of generations (although this is not to suggest that a person does not change the way they speak during their own lifetimes; that kind of change could also form part of such an evolutionary process). Any complete theory of the evolution of colour terms should

---

[5] It should be acknowledged that 'evolution of language' can also refer to real phylogenetic (biological) evolution, but this is clearly not the sense in which Berlin and Kay (1969) intended to use the term *evolution*. It seems that the differences between the colour term systems used by speakers of different languages are due primarily to the language to which those speakers are exposed, not to any difference in their genetic makeup as compared to speakers of other languages. Hence it would seem that the only evolutionary process through which the colour term systems of individual languages evolve is a cultural one, not a phylogenetic one. It is interesting to note, however, that many early studies did in fact presume a non-linguistic explanation of at least some of the differences between colour term systems, in that it was assumed that colour vision had only recently evolved, and that speakers of languages with fewer colour terms were in fact able to distinguish fewer colours (Gladstone 1858 and Geiger, 1880, both cited in Berlin and Kay, 1969).

Also, Bornstein (1973) argued that the presence of green-blue composites is, at least in part, due to a difficulty that speakers of languages with such terms have in distinguishing green and blue, but Maffi and Hardin (1997) have argued that such an interpretation is unlikely to be correct, because there is a consensus that the visual apparatus of all peoples is essentially the same. One exception to this generalisation, however, is that in some localities of the world, it is apparent that the incidence of colour blindness is much higher than the norm (Sacks, 1996), so in such communities we might well expect that this would have a significant impact on the languages spoken in those communities. If a relationship between incidence of colour blindness and colour term systems could be demonstrated, it would show a difference between languages due to genetic differences between the speakers of those languages, and so the explanation of the differences between those languages would partly be in terms of phylogenetic evolution.

specify exactly how colour term systems are transmitted between generations, and what properties of either people or their environment are responsible for creating the attested typological patterns.

The methodology which was used here in an attempt to provide a fully explicit and rigorous theory concerning the evolution of colour terms involved the construction of a computer model. This model aimed to implement the most essential elements of the processes through which colour terms evolve over a number of generations, including specifying how their denotations could be learned. This required making a number of assumptions to compensate for areas where Berlin and Kay, and subsequent researchers, have not made sufficiently precise claims[6]. It would be surprising if every

[6] I do not intend my assertion that Berlin and Kay's theory is in many aspects vague to be in any way a criticism of it. Clearly, if we are unsure of the details of some process then it might well be best to simply leave them unspecified, and instead to state only the general properties of the process. It is often said that one of the advantages of computer modelling is that it forces researchers to make their theories more explicit (Latimer, 1995). However, consideration of actual computer models would show that only certain aspects of the processes under investigation are ever made explicit. For example, in the models presented in this thesis, processes concerning how colour terms are articulated and perceived, as well as the phonological form of colour terms, are completely neglected. This is of course not in any way a claim that such processes do not exist, but these matters have simply been left unspecified, because it was felt that they were not essential to an explanation of colour term typology. Hence, in the computer models, colour words were simply represented by strings of letters, which models the fact that different colour words have to be phonologically distinguishable, but leaves vague the details of how this is achieved in practice in any particular language. We might find justification for this approach in term's of Occam's razor, which is stated by Heylighen (1997) as 'one should not increase, beyond what is necessary, the number of entities required to explain anything'. We might argue that this would provide support for a principle that theories should be left vague unless there is

detail of the implementation corresponded exactly to how colour term systems evolve in the real world, so it would be incorrect to suggest that the model replicates colour term evolution precisely. Instead, the aim of this thesis is to validate the evolutionary hypothesis, and to provide a simple and plausible theory which is able to account for the available typological evidence.

As will be shown below, the evolutionary model can account for much of the data concerning colour term typology, and this could be taken as evidence that the model, at least in broad outline, does reflect how colour term systems evolve. However, perhaps more importantly, it demonstrates that colour term typology could be the product of languages evolving under the influence of the human visual system, and so no other factors, for example innate knowledge or relativistic influences, are necessary in order to explain colour term typology. This is not to say that the present model proves that such factors do not play a role in shaping colour term systems, for it certainly does not do so, but if colour term typology can be explained without reference to such factors, then it would seem that we should not presume that they play a part in forming colour term systems unless further evidence can be found to support such a view.

---

evidence favouring one set of specifics over other possible ones. The assumptions made in the construction of the computer model of colour terms, and the justifications for them, are made explicit below, primarily in section 3.2.

2.1 Colour Terms across Languages

Before going on to consider how colour term typology can be explained, it is first necessary to obtain a clear understanding of the data concerning the properties of basic colour terms, both within individual languages and cross-linguistically. One of the properties which this thesis aims to explain is that, in all languages, basic colour terms have *prototype properties*. The central function of a colour term is clearly to identify a range of colours, so that the word can be used to distinguish these colours from those which the word does not denote. However, colour terms do not simply denote a uniform range of colours, but instead some colours are members of the colour category corresponding to the colour term to a greater extent than are other colours.

Typically, for each colour term, there will be a single colour which speakers of the language consider to be the best example of that colour term, and this colour is called the prototypical colour. Moving away from the prototype, colours become increasingly less good examples of the colour category as they become more dissimilar from the prototype. At a certain level of dissimilarity from the prototype, we will find colours for which it is difficult to determine whether they come within the colour category, or outside of it. Hence for these colours it is not clear whether the word can be used to denote them or not. This part of the colour space is known as a category's fuzzy boundary, where the exact limit of the category is unclear (Taylor, 1989).

There is considerable inconsistency between people as to where they place the limits of a colour category, because, if different speakers of the same language are asked to outline the boundaries of a colour term on an array of colour chips, then they are

likely to place the boundary in slightly different places. Furthermore, if the same person is asked to perform the task a second time, they are unlikely to place the boundary in exactly the same place, which shows that even for individual speakers there exists considerable inconsistency concerning category boundaries (Berlin and Kay, 1969). The existence of the prototype phenomena is also clearly demonstrated by expressions such as 'a good red', 'sort of red' and 'slightly red', none of which would make sense if all red colours were equally good examples of the colour term *red* (Kay and McDaniel, 1978). While colour terms are perhaps one of the best examples of prototype categories, Taylor noted that the meanings of many words, including most nouns, verbs and prepositions, also have prototype properties, and that prototype properties can be observed in many other aspects of language, including in syntax, morphology and phonology.

Taylor (1989) claimed that prototype categories have prototype structures because the extent of the category is determined by contrasting individual members of the category to the central prototype. However, this account seems somewhat problematic, because there is variation between colour terms as to how similar a colour must be to the prototype for it to be a member of the category. This clearly has to be the case, as some colour terms denote much larger parts of the colour space than others do. A further, and perhaps more important objection, is that the prototypical colour is not always in the centre of the region denoted by the colour term (as can be seen in the colour charts of Berlin and Kay (1969)), so in such cases knowledge of the prototype would not be sufficient to determine the range of the colour term, even if the size of the colour category was also known. In the Bayesian model used in the simulations reported here, the prototype is not used to define the extent of the

category, but instead arises as a by product of the acquisition mechanism, as will be shown below in section 3.6.

Berlin and Kay (1969) investigated the range of colour term systems in a wide range of languages. Their study gathered data from speakers of twenty different languages, using arrays of Munsell[7] chips. Firstly the basic colour terms of each language were determined, in terms of the criteria discussed above. This produced the first finding of Berlin and Kay's study, which was that all languages appear to have between 2 and 11 basic colour terms[8]. Berlin and Kay also used data from published sources such as

---

[7] Munsell chips are small pieces of cardboard which are painted in carefully controlled pigments, so that the colours of the chips are systematically spaced over the range of all possible colours, at least in as far as it is possible to create the appropriate pigments. While Munsell chips, and the Munsell system of ordering colours (Cleland, 1937), seem to be the only colour apparatus and colour order system used in linguistic research, it should be noted that there are many other colour systems available, some of which, such as the natural colour system (Hård, Sivik and Tonnquist, 1996), vary considerably in the structure they give to the colour space in terms of how far apart particular colours are placed. It is important to bear in mind that, while attempts have been made to standardize the Munsell system so that the distances between colours reflect psychological data concerning colour dissimilarity (Indow, 1988), it should be noted that there are many such methods for standardizing colour spaces, and no consensus has been reached amongst colour theorists as to establishing a single correct colour space.

[8] Although note the discussion above concerning the definition of a basic colour term. If the Russian term *goluboy* and the Hungarian term *vörös* were considered to be basic, then these languages would both have 12 basic colour terms. However, there does not appear to be any example of a language which has 12 colour terms which are all clearly and uncontroversially basic, despite there being a considerable number of languages with 11 basic terms. Hence there does seem to be a need for an explanation of why there is a limit of 11 terms, at least in most cases. It is possible that this phenomenon could be due, at least in part, to languages becoming increasingly similar due to contact

dictionaries and grammars to bring the total number of languages in their study to 98, and that data appeared to confirm the finding.

The next stage of the research involved asking each person to map both the outer boundary of each of the basic colour terms in their language on an array of Munsell chips, and to identify the best or most typical examples (the prototype) of each term. They discovered that the boundaries of the areas of colour denoted by colour terms varied greatly between languages, which was consistent with earlier findings. Clearly where a language has fewer basic colour terms then each colour term must denote a wider range of colours, at least if the language is to allow each part of the colour space to be named by one of the terms. However, even in languages which had the same number of colour terms, and in which each colour term in one language was roughly equivalent to a colour term in another, the locations of the boundaries of the colour term's denotations varied considerably.

This finding might have led to the view that there were few constraints on the kinds of colour term systems which could emerge in languages, but a quite different picture emerged when Berlin and Kay looked at the distribution of the prototypes of the colour terms. They found that most of these were placed in just a few areas of the colour space, either clustering on single Munsell chips, or a small number of nearby chips, leaving over 70% of the surface of the Munsell array clear of any prototypes at all. They showed that there were in fact 11 clusters of prototypes, corresponding to

---

with one another, as many of the languages with 11 basic colour terms are major languages not restricted to a small locality (although there remain many fairly widely used languages with large numbers of speakers which have less than 11 basic colour terms).

the locations in which English speakers would place the best examples of each of the basic colour terms in English. This clearly showed that the lexicalization of the colour space in unrelated languages was certainly not random or completely arbitrary, but appeared to conform closely to universal restrictions.

A further finding emerged when Berlin and Kay investigated the combination of colour terms existing in any particular language. They found that when terms were classified based on their prototypes, it was largely predictable which colour terms would exist in a language, if the number of basic colour terms in the language was known. Berlin and Kay expressed these regularities using the implicational hierarchy shown in Figure 2.1. This hierarchy was constructed partly using evidence derived from participants using Munsell arrays, though evidence for most of the languages used to construct this hierarchy came only from published sources such as dictionaries. All languages appeared to have a term with its prototype at white, and a term with its prototype at black, shown at the left of the hierarchy, but some languages had no other basic colour terms but these. However, if a language had a term for any of the colours further right in the hierarchy, it always had terms for all the colours further left in the hierarchy. For example, if a language had a term with its prototype at yellow, then it would also have terms with their prototypes at white, black and red.

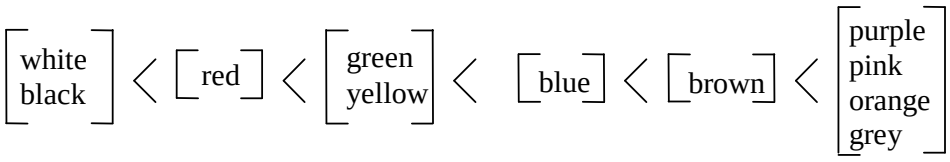


Figure 2.1. Berlin and Kay's (1969) Implicational Hierarchy.

It should be noted that while Berlin and Kay proposed that this hierarchy described the general patterns seen in colour term systems cross-linguistically, they did acknowledge the existence of some exceptions and problematic cases. Some of these problems revolved around cases where it was unclear which terms should have been classified as basic. For example the Catalan speaking informant considered *negre* 'black' to be a kind of *gris* 'grey', while acknowledging that English *black* was not a kind of *grey*. This indicated that *negre* might not be a basic colour term, and so that Catalan might lack a basic term for black, hence violating the hierarchy. Another problem concerned Cantonese, which had terms for white, black, red, green, yellow, blue, pink and grey, but which lacked a term for brown, again contravening the hierarchy. However, Berlin and Kay suggested the terms for pink and grey appeared to be recent additions to the language, and that they might not be basic terms at all.

Other languages which were problematic with respect to the hierarchy were Vietnamese, Western Apache, Hopi, Samal and Papago, all because they either lacked a basic colour term which the hierarchy predicted they should have, or because they had acquired a basic colour term before they had reached the stage in the hierarchy where it would normally be expected to emerge. However, given that only six out of the 98 languages in the study appeared to be seriously problematic, Berlin and Kay did not modify their theory, but instead simply decided to treat these languages as exceptions.

A number of criticisms can be made of Berlin and Kay's methodology, perhaps the most important of which is that most of their data concerning colour terms came from published sources or communication with other linguists, rather than directly from native speakers of the languages concerned. Furthermore, out of the twenty languages

for which interviews were conducted using arrays of Munsell chips, in all but one case the informants were living in America, and were bilingual in English, so it could have been the case that the results were partly a product of the influence of English colour categorization on the other languages. Also, for most of the languages, data was obtained from only a single informant, so it was not possible to be sure whether the results reflected the language as a whole, or simply the idiolect of one individual speaker.

Since Berlin and Kay published their original study, however, there has been a great deal of interest in basic colour terms, and much more data has been collected, generally using methods which addressed the deficiencies of Berlin and Kay's original study. These studies have in large part confirmed Berlin and Kay's original findings, though several modifications have been made to their hierarchy (including, for example, Kay, 1975 and Kay, Berlin and Merrifield, 1991), to accommodate some language types which were not attested in their original study, or which were originally treated as exceptional aberrations.

A very large survey of the colour term systems of 110 minor languages, the World Colour Survey (Kay, Berlin, Maffi and Merrifield, 1997), has now produced a wealth of high quality data, allowing us to get a much more complete picture of colour term systems worldwide. Using this new data, Kay and Maffi (1999) produced a new classification of colour term systems, which has modified the original hierarchy of Berlin and Kay (1969) considerably, but which still shows that the attested colour term systems are only a small subset of those which are logically possible. Kay et al (1997) noted that it appears that there are six fundamental colours, corresponding to the colours which would typically be the prototypes of red, yellow, green, blue, black

and white colour terms, and that the order of appearance of basic colour terms which do not include one of these colours in their denotations, such as the English terms *purple, orange, turquoise, brown* and *grey*, is less predictable. Their classification of colour term systems was made primarily by considering only terms whose denotation included at least one of the fundamental colours. These classifications were then simply augmented with a list of which other basic colour terms existed in the language.

However, Kay Berlin and Merrifield (1991) did note that while purple or brown terms may be seen in languages which do not have separate terms for both green and blue, contrary to Berlin and Kay's hierarchy, it seems that orange or pink terms do not normally appear unless a language has separate terms for both green and blue. Kay (1975) had already noted that grey terms sometimes appear in languages even when those languages have not developed terms for some of the other colours which Berlin and Kay predicted would normally appear before grey. The general conclusion that we can draw from these findings is that the order in which such terms emerge in a language does not seem to be completely predictable, though there appear to be general trends concerning the order in which these terms emerge. We can note that orange terms tend to be seen only in languages which have already developed purple terms, but that this is not always the case (MacLaury, 1997a), and that purple terms tend to emerge once a language has acquired separate terms for green and blue, but that this rule does not apply to all languages. These specific findings are among the data which the evolutionary computer model described here is able to explain.

Another important difference between the analysis of Berlin and Kay (1969) and that of Kay and Maffi (1999) is that, while Berlin and Kay classified terms simply in terms

of the locations of their prototypes, Kay and Maffi have paid more attention to terms' full denotational ranges. They classified colour terms depending on which fundamental colours they contained, rather than just in terms of which fundamental colour corresponded to the term's prototype, so that, for example, two terms which both had red prototypes, would be classified differently if one also named a range which included yellow, but the other did not.

Berlin and Kay had noted the existence of languages which had only two basic colour terms, and they assumed that such systems would cleanly divide the colour space up into light and dark colours, though they did not investigate any such language experimentally. However, when Heider and Olivier (1972) investigated the Indonesian language Dani, using Munsell chips, they found that the two colour words divided up the colour space so that one, *mola* denoted light colours, but also yellow and red hues of medium lightness, while the other, *mili*, denoted dark colours, but also blue and green hues of medium lightness, so that their denotations were complementary, essentially covering the whole colour space. (So, for example, *mili* would denote dark yellow hues, although, in English, we would call dark colours of the same hue as yellow *brown* or *khaki*.) Further research has shown that all languages with two colours, whilst being extremely rare, are either of this type, dividing the colour space up into a white-red-yellow category and a black-blue-green one, or else they simply make a dark light split, as was originally presumed by Berlin and Kay (MacLaury, 1997a). Initially it was presumed that in such systems, one term would have its prototype at white, and the other at black, but Kay et al (1997) noted that this was not always the case, and that such three way composite terms can have their focus on another of the fundamental colours, so, for example a white-yellow-red category might have its prototype at red, rather than at white.

In the colour charts published by Berlin and Kay (1969) which showed the colour names given to each colour chip on a Munsell array in each of the 20 languages investigated experimentally, many of the colour chips were left unlabeled, so it appeared that, in languages such as Mandarin Chinese, most colours could not be named by any colour term. This appeared to contrast with languages such as English, in which most people are able to specify a basic colour term which can describe every Munsell chip, although even in English it is difficult to decide on a name for some Munsell chips which are near the boundaries of the denotations of two or more basic colour terms. However, it seems that such situations were probably a product of the way in which some of the linguists involved in the study elicited their results, and that in almost all languages there are no areas of the colour space which cannot be named by a basic colour term.

Kay et al (1997) mapped out the areas of the colour space which could be named by each colour term, based on the reports of a number of informants. They showed that if several speakers of the same language were interviewed, and if only those colour chips which all of the informants think can be named by a colour term are mapped on a Munsell array, then there would typically be large gaps between the areas of colour denoted by each word, because not all speakers of the language agree on which word should be used to name the more marginal members of each colour category. However, if instead the criteria for considering a colour to be within the denotation of a colour term are reduced to 30% agreement between speakers, and the colour terms' denotations are again mapped on an array of Munsell chips, then the gaps between the colour terms largely disappear.

MacLaury (1997a) also noted that how widely an informant draws the boundary of a colour term can depend on exactly what instructions they are given. While initially informants may include only a fairly small number of chips within a colour term's denotation, if they are subsequently asked if any other colour chips could also be named by the colour term, they are likely to include many more colour chips within the word's denotation. So it seems that together all the basic colour terms in a language typically cover the full range of possible colours, but because some colours are only marginal examples of any basic colour term, people may be reluctant to include them within any word's denotation.

Kay and Maffi (1999) did not find any language in the World Colour Survey which left any region of the colour space unnamed, although in some languages naming of parts of the colour space is very inconsistent across speakers, and it is possible that in some such cases the colour terms for some parts of the colour space do not fulfil Berlin and Kay's (1969) criteria for basic colour term status. However, Kay and Maffi do acknowledge the existence of one well documented language which does appear to leave parts of the colour space unnamed, Yélî Dnye (Papua New Guinea), which was reported by Levinson (2001). This language appears to have only three basic colour terms, *kpêdekpêde* 'black', *kpaapîkpaapî* 'white' and *mtyemtye* or *taataa* 'red'[9]. There also appears to be one other colour term in the language, *wuluwulu* 'dark red', but Kay and Maffi analyzed this as a non basic term. What is really interesting about this language is that the denotations of the three basic colour terms do not extend to cover

[9] The two forms for red appear to be due to dialectal variation, although many speakers used both terms, and a few put the prototype of each in a different part of the colour space (Levinson, 2001).

the whole of the colour space, so that large areas of the colour space are left without any colour term which can name them[10]. However, what is very clear from Kay and Maffi, is that the language studied by Levinson is exceptional, and that almost all languages do partition the colour space so that there is a colour term which can name every colour. Hence any theory concerning colour terms must explain why partition is the norm, while still allowing for the existence of languages which are exceptions to this rule[11].

It now remains to specify exactly which types of colour term systems Kay and Maffi (1999) found to be attested in the World Colour Survey. They have proposed that languages evolve from a state in which they have only two colour terms, and that they then gradually add more terms over time, never losing colour terms once they have gained them. Their classification of languages therefore takes the form of an evolutionary sequence, which begins with two colour terms, and then progressively subdivides the areas of the colour space named by each of these terms until each of the fundamental colours is named by a separate colour term. They found that 83% of the languages in the World Colour Survey lie somewhere along the trajectory shown in Figure 2.2, where early stage languages with only two colour terms are at the top of the diagram, in which case one term names light colours together with yellow and red,

---

[10] It should be noted, however, that Levinson (2001) clearly shows that it is possible to form expressions which describe other colours, so it is not the case that speakers of Yélî Dnye cannot refer to particular colours simply because they do not have actual colour terms denoting those colours.

[11] We should note, however, that Levinson (2001) suggests that languages which do not partition the colour space are much more common than Kay and Maffi (1999) acknowledge.

and the other dark colours together with blue and green[12]. Languages in which each of the fundamental colours is represented with a separate term are at the bottom of the diagram, and intermediate languages lie somewhere in between. Languages were considered to lie on this trajectory if they appeared to be best classified either as being at one of the five stages, or as being in transition between stages.

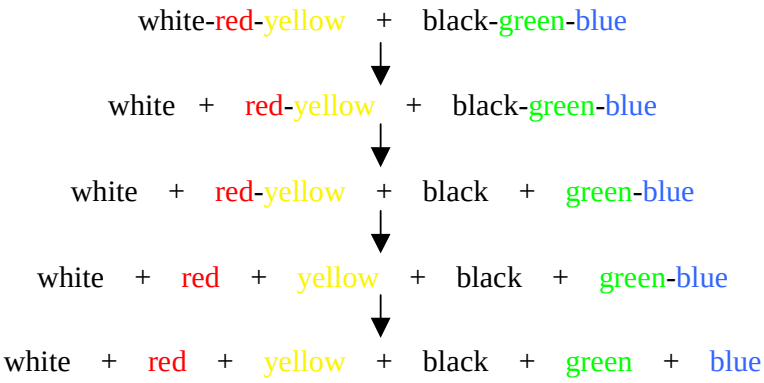


Figure 2.2 The Main Line of Kay and Maffi's Evolutionary Trajectory

In a language in transition, there would typically be disagreement between speakers as to how many basic colour terms the language possessed, usually because older speakers would not use a colour term which had entered the language during their lifetimes. In this case a language might best be classified as being at one stage for some speakers who did not use the extra colour term, and at a subsequent stage for those speakers who did use the term. For some speakers one of the colour words

---

[12] This trajectory appears to ignore languages with two basic colour terms, which simply make a split into dark and light colours, rather than grouping red and yellow with white and green and blue with black. This could be because such words may not be considered to be true colour terms, as the light/dark distinction on its own could be considered to be outside of the domain of colour.

might be only very weakly established, and so it could be unclear whether that term should be considered to be a basic colour term, leading to ambiguity as to at which stage the language should be placed. Out of 91 languages which Kay and Maffi (1999) placed on this trajectory, 18 were considered to be in transition.

Kay and Maffi (1999) acknowledged that not all languages seem to follow the above trajectory, at least not for the whole of their development. They placed seven languages either on a side branch of the main trajectory, or in transition into or out of the side branch. Once a language has reached the second stage of the trajectory shown in Figure 2.2, languages usually gain an extra colour term, so that a black term and a green-blue composite term come to replace the black-green-blue term. However, it seems that a few languages instead split apart the red-yellow composite and replace it with separate red and yellow terms, leaving the black-green-blue term intact. One language in the World Colour Survey seems to be in transition into this state, and two appear to be in transition out of it.

Once a language has reached this stage, it seems that there are two routes it can take. Firstly, the black-green-blue composite can split into a black term and a green-blue composite, in which case the language will return to the main line of the trajectory with a five term system. However, another possibility is that the black-green-blue composite splits to produce a black-blue composite and a separate green term. Three languages in the World Colour Survey were given this classification, while one was analyzed as being in transition into this state, and one as being in transition out of it. Following this stage, the black-blue composite splits, returning the language to a state on the main line of the trajectory.

In order to explain the existence of three other languages, Kay and Maffi (1999) had to postulate another branch to the evolutionary trajectory, although in this case its origin was somewhat unclear, as it could not have developed straightforwardly from the main line of the trajectory. This branch was proposed because two languages in the World Colour Survey had yellow-green-blue composite terms, together with separate black, white and red terms. At the first stage of the main line of the evolutionary trajectory, the yellow and green fundamental colours appear in separate composites, so languages of this type could not emerge simply through progressive splitting of composite colour terms. There was also one language which contained a yellow-green composite, together with black, white, red and blue terms. This language could be derived from the earlier type if the yellow-green-blue term split into a yellow-green term and a separate blue term, but this does not provide an explanation of how the yellow-green-blue composite arose in the first place.

I am aware of two hypotheses concerning the origin of yellow-green-blue composites. Kay and Maffi (1999) suggested that languages with these composite terms might be derived from languages such as Yélî Dnye, where there are black, white and red basic terms, but no basic term for the rest of the colour space. If the language then developed the principle of partition, so that the colour space was divided so that every colour could be named by a basic colour term, then it would seem that a yellow-green-blue term might emerge to fill the space of colours previously without a basic colour term. Kay and Maffi proposed that this was how yellow-green-blue, and ultimately yellow-green terms, emerge, although they did acknowledge that this hypothesis was somewhat speculative.

An alternative theory concerning the origin of yellow-green-blue composites was proposed by MacLaury (1997a). He suggested that these terms develop from languages which originally had only two colour terms dividing the colour space into light and dark. If a third term were then to emerge, it is possible that this term would then correspond to a middle brightness colour, and if a red term were also to emerge then the middle brightness colour would be left denoting the other colours of middle brightness: yellow, green and blue. This is obviously a very different explanation from Kay and Maffi's, but there does not appear to be really clear evidence favouring one over the other. The computer model which is the subject of this thesis does not propose that languages must always evolve by subdividing composite colour terms, so the existence of yellow-green and yellow-green-blue composites is not problematic. These colour terms could emerge simply as a product of random drift in the meanings of colour terms, so that a colour term which did not previously name both yellow and green might come to do so.

There were a few languages in the World Colour Survey which did not seem to fit well into Kay and Maffi's (1999) theory of a limited number of fixed evolutionary trajectories. Firstly there were three languages which appeared to be in transition directly from a state in which they had a black-green-blue composite term and separate white, red and yellow terms, to a state in which each of the fundamental colours was represented by a different colour term. This missed out the expected intermediate stages in which the black-green-blue term would be expected to split first into either black and green-blue or green and black-blue terms. It is quite possible that this phenomenon could be due to rapid change in the societies in which these languages were spoken, because in general there appears to be a positive correlation between level of technological development in a community, and the number of

colour terms present in the languages spoken in those communities. Therefore if the level of technological development in a community were to increase very rapidly, this might lead to a rapid increase in the number of colour terms in the community's language, which could lead to it jumping one of the stages in Kay and Maffi's trajectories. Of course, without further evidence, this proposal must remain speculative, but it does provide one possible explanation of what otherwise might appear to be fairly surprising data.

Kay and Maffi (1999) also noted that there were four languages which they could not place anywhere on the evolutionary trajectories. Each of these languages contained colour terms with their prototypes at black, white and red, but the way in which they divided up the rest of the colour space was inconsistent. Generally there would be a considerable amount of idiosyncrasy in the colour terms used by speakers of these languages, with different speakers using different terms. Even where speakers used the same colour terms, the areas of colour which they named with each term tended to be extremely variable. One such language is Culina (Peru, Brazil), which has white and red terms, a yellow term which extends into green and blue, and a black-green-blue composite. There are therefore two terms which seem to overlap in their denotations, both of which name green and blue colours. This language cannot be placed on the evolutionary trajectory, because it seems to be a mixture of a white, red, yellow and black-green-blue system with a white, red, yellow-green-blue one, and so there are two places on the trajectories where it could potentially be placed.

Another problematic language was Kuku-Yalanji (Australian), which again has black white and red terms, with the black term showing some extension into blue. Some speakers also used another term, *kayal*, to denote either blue-green or just green, but

because it was not used by most speakers it was not considered basic. Most speakers did, however, use another term, *burrkul* (or *burkul*), usually to denote all colours which were not black, white or red, though speakers who had well established green-blue or green terms did not used *burrkul* for those colours. Kay and Maffi (1999) did not classify *burrkul* as a basic colour term, and so this language did not correspond to any of the language types on Kay and Maffi's trajectories. The existence of so much inconsistency in the use of colour words amongst speakers of a single language might seem surprising, but it is a widely attested phenomenon (MacLaury, 1997a).

Some linguists have challenged the general findings of Berlin and Kay (1969) and Kay and Maffi (1999), suggesting that colour term systems do not conform to predictable rules to the extent that those researchers claimed. Saunders (1992) has gone so far as to suggest that some of Berlin and Kay's findings were a product of their methodology. She noted that many languages lack true colour words, that is words which denote purely colour. Typically many words which are often considered to be basic colour words have other connotations, for example they may have religious significance. She has claimed that such words can only be understood in relation to other words in the language, and within the context of the belief systems of which they form a part. These criticisms do appear to have a reasonable foundation, but it seems that regardless of the connotations surrounding colour words, part of their meaning corresponds to the range of colours which they denote. Hence I do not see how Saunders' criticisms invalidate Berlin and Kay's findings.

Levinson (2001) made similar criticisms of much of the work concerning colour terms. He has claimed that languages do not 'universally treat colour as a unitary domain, to be exhaustively named.' (p3). He suggested that the idea of colour as a

domain for linguistic categorization, in which we would expect to find a set of co-hyponyms each denoting specific ranges of colour might not be applicable to all languages. He noted that some languages conflate other properties, such as texture or variegation, with colour. This contributes to difficulties in determining which colour terms in particular languages are basic, because it makes it more difficult to decide which words are colour terms at all. It is, in any regard, often difficult to determine objectively which colour terms are basic, as the properties of some terms appear to place them at an intermediate level, in between that of basic and non-basic terms. Levinson noted that sometimes even whole expressions might seem to fulfil some of the basic colour term criteria better than individual colour words.

However, none of Levinson's (2001) findings seem to be at odds with the basic conclusions of Kay and Maffi (1999). Their data still demonstrate regularities in the way in which informants name colour chips, regardless of how many other aspects of the colour words' meanings or syntactic properties are disregarded. It may well be that Kay and Maffi's decisions concerning which colour terms should be considered basic were influenced to some extent by a reluctance to make classifications which would be inconsistent with their hypothesized evolutionary trajectories. However, the distinction between basic and non-basic colour terms is clearly not completely arbitrary, and while it seems likely that whether a marginal term was classed as basic or not might in some cases have been determined partly by whether it supported or contradicted Kay and Maffi's trajectories, such an issue can not be said to falsify Kay and Maffi's theory. Clearly, when reasonable distinctions were made to differentiate between basic and non-basic colour terms, the patterns reported by Kay and Maffi were observed. We should remember that, ever since Berlin and Kay (1969), exceptions to the hypothesized evolutionary trajectories have been acknowledged, and

so, even if classifying a term regarded as non-basic as basic would result in a language no longer fitting on an evolutionary trajectory, this would not be a serious problem for Kay and Maffi's theory of predictable trajectories.

MacLaury (1997a) has described some other phenomena concerning colour terms that Kay and Maffi (1999) did not mention. Firstly, he has noted the presence in some languages of basic colour terms naming broad ranges of desaturated colours, especially terms which include some range of dull brownish, lavender, grey and/or beige colours. These categories tend to be observed in languages with relatively few colour terms, and MacLaury reports that they are common. It seems that as languages gain colour terms these words become restricted to narrower ranges of colour, perhaps becoming terms like English *brown*. These colour terms do not appear on Kay and Maffi's (1999) evolutionary trajectories because they do not contain a fundamental hue, but any complete theory of colour term typology must explain how they could arise.

MacLaury (1997a) also discusses a phenomenon which he terms coextension. Coextension describes situations in which two colour terms name approximately the same range of colours, so that it seems as though they are in free variation, and that both are associated with a single colour category. However, in such cases, one of the colour terms, the *dominant* term will have its prototype near to the centre colour category, while the other, the *recessive* term, will tend to have its prototype near to the edge. Initially, when asked to map out the extent of the colours named by the recessive term, informants usually include only a small range of colours. However, if they are prompted to include all the colours which the term could possibly name, then they will extend its range so that it covers almost as many colour chips as the

dominant term. The existence of coextension is not widely discussed, but MacLaury reports that it is common, especially for composites such as green-blue terms, so it is important that an explanation of colour term typology allows for the existence of coextensive colour terms.

Overall we can characterize the empirical data concerning basic colour terms as firstly showing a great diversity of colour term systems across languages, but also revealing a lot of cross-linguistic regularities. All basic colour terms have prototype properties, usually with a single best example and fuzzy boundaries. The attested colour terms are all of types which make up only a small subset of those which are logically possible, and there are regularities in the observed combinations of types of colour terms that can exist in any language. Berlin and Kay (1969) and Kay and Maffi (1999) have explained this data by claiming that languages evolve along fixed evolutionary trajectories, and that they gradually add new basic colour terms over time. However, there appear to be some languages which are exceptions to the regular patterns, and not all researchers accept that there is a clear distinction between basic and non-basic colour terms. However, it seems that Kay and Maffi's theory of trajectories is successful in accounting for most of the data concerning most languages, and so it was taken as the benchmark against which the evolutionary computer model of this thesis was tested.

2.2 Modes of Explanation in Linguistics

Given the typological data reviewed above, we need to consider just what kind of explanation is appropriate to account for it. In order to determine what sort of theory is appropriate, we need to consider the nature of languages, and exactly what we mean by the word *language*. Languages can exist physically as speech or writing, but they

cannot be understood simply by collecting example sentences, because languages are productive systems and so there is an infinite number of possible sentences in all languages. Languages are known by individual people, and so can be seen as mental phenomena, in which individual people know the rules which constitute the language. However, languages are clearly not known just by individuals, but by a number of people who all use the same conventions to communicate, so perhaps are best understood as social phenomena. The first issue to be discussed in deriving an explanation for colour term typology is therefore whether the theory should be concerned with psychological knowledge, an abstract system, or some other concept of language.

Chomsky (1965, 1972, 1986) has emphasized that languages can be seen primarily as psychological phenomena, in that the ability to use and understand language is an ability which we have as individual people. Chomsky (1986) introduced the term *I-language*, meaning our psychological knowledge of language, and he argued that linguistics should be primarily concerned with the study of I-language, the form which language takes in the mind/brain. He considered *E-language*, actual examples of speech or writing, to be of only relatively peripheral interest, although of course acknowledging that the study of *I-language* is dependent on inferences made through observations of *E-language*.

Furthermore, Chomsky (1965, 1972, 1986) emphasized the need for a linguistic theory to account not only for static knowledge of language, but also to explain how children come to acquire I-language. Perhaps the most important observation supporting this viewpoint is that children are born without the ability to speak any language, but after exposure to a language for several years they gain the ability to

speak it fluently, almost without exception. Furthermore, all children are equally able to learn any language, the language which they learn being determined simply by the language to which they are exposed during childhood[13]. Chomsky has argued that the central goal of linguistics must be to explain how children come to acquire knowledge of language based on observations of other people's speech, a process which is represented in Figure 2.3. (Figure 2.3 is based on Chomsky, 1972, p119, but adapted in order to bring the terminology in line with modern usage.)

Figure 2.3. Chomsky's Conceptualization of Language Acquisition.

The central component of Chomsky's perspective on linguistics is what he has called a *Language Acquisition Device*, which is the part of the mind/brain that produces I-language. It learns based primarily on observations of E-language, but also could use any other evidence which might be available to the child. Chomsky (1986) placed particular emphasis on the central importance of understanding the Language Acquisition Device to gaining an understanding of language in general. Given this perspective, regularities across languages are generally seen as products of the innate Language Acquisition Device. The possible human languages are those which the Language Acquisition Device is able to learn, and so, if a particular linguistic

[13] There are of course some exceptions to this generalization, in particular that deaf children have no problem learning sign languages, but they do have problems learning spoken ones, and some disorders such as autism can prevent or impair the acquisition of any language at all.

structure is not attested, then this is presumably because the Language Acquisition Device is not capable of acquiring it.

The majority of Chomsky's work has been extremely nativist, in that it has assumed that the Language Acquisition Device supplies most of the structure of language in the form of *Universal Grammar*, and that learning plays a relatively minor role in simply choosing which of a constrained range of parameterized possible structures are present in the language being learned. Examples of this kind of theory are Government and Binding Theory (Chomsky, 1984) and the Minimalist Program (Chomsky, 1995), in which Universal Grammar, which is genetically[14] specified, consists of a limited number of innate grammatical categories and principles. Language universals and typological patterns can be explained as products of Universal Grammar. As all people share almost identical genotypes (there being only very limited genetic variation between individuals), they hence have almost identical Universal Grammars. We see similar structures in different languages, because these are part of Universal Grammar, and other possible, but unattested, structures are not seen, because they do not exist in Universal Grammar.

---

[14] It should be acknowledged that there is an increasing amount of evidence showing that the link between genes and neural structure is far from straightforward, and that development is often the product of interaction between genes and experience (Elman et al, 1996). However, while Chomsky does not go into the details of the mechanism, it is clear that the form of universal grammar is in some way determined genetically, and that Chomsky believes that all people develop essentially identical universal grammars (Chomsky, 1986, p25). Hence, for present purposes, it does not seem to be necessary to address the question of what form the mechanism through which universal grammars arise takes.

Kay and McDaniel (1978) came close to Chomsky's position in proposing a limited set of universal colour categories, which were determined by the neural response functions of cells in the retina of the eye. They proposed that some of the universal properties of colour term systems were due to all the colour categories in the world's language being chosen from this universal inventory. This was clearly an attempt to explain properties of the colour term system in terms of innate structure[15], and so was in this way similar to Chomsky's Universal Grammar. However it appears that on the one hand Kay and McDaniel's proposal was too restrictive, because there is considerable variation in the exact denotations of similar colour terms in different languages, but also that it was not sufficiently constraining, in that it predicted the existence of types of colour categories which have never been observed[16].

While Chomsky has emphasized both a psychological perspective on language, and a strongly nativist approach, many approaches to linguistics retain the focus on psychology, but propose a much greater role for learning than is typical in Chomsky's work. Examples of this can be found in many of the connectionist approaches to linguistics, such as that of Rumelhart and McClelland (1986), who modelled the acquisition of the past tense of English verbs. Their neural network, together with the

---

[15] From Kay and McDaniel (1978) it is not clear exactly whether the categories themselves should be considered innate, or whether they are simply derived from innate structures. However, it would seem that the nature of the categories is determined by neural structures which are common to all humans (with the exception of colour blind individuals), and so it seems reasonable to consider them to be innate, even if there is some input from experience in determining exactly which categories emerge in each individual person.

[16] Kay and McDaniel (1978) is discussed again in section 2.3.

algorithm used to train it, can be seen as a Language Acquisition Device. At the end of the learning process, the final knowledge of language will correspond to the neural network and the learned connection strengths. (There is therefore no clear division in this theory between the acquired knowledge of language and the Language Acquisition Device, but that is also true of many strongly nativist theories.) Rumelhart and McClelland's model reproduced many of the patterns observed when children learn English, and so it was argued that these patterns could be explained as resulting from properties of the children's Language Acquisition Devices, which they have suggested may learn in a similar way to the neural network (Rumelhart and McClelland, 1986, p267).

All of the above approaches to explaining language have placed little emphasis on the social contexts in which languages are used, focusing instead on language in individual people. In contrast, de Saussure (1959/1916) proposed that language must be understood as a phenomenon which is simultaneously psychological and social. While the ability to speak and understand language is undoubtedly psychological, language is used principally for communication between people. Successful communication can occur only when more than one person shares the same language, and in general languages are shared by whole communities of speakers. Individuals will only be able to communicate successfully if they use at least approximately the same conventions for expressing meaning as other members of the community. Changes in language will be initiated by an individual person making up a new word or expression, or using an existing word to mean something new, but such innovations will only become part of the language if they are adopted as conventions by other members of the speech community. Hence, we can see that language exists as a

system which is shared by all members of a community, and so it may not be possible to fully understand language simply from a psychological perspective.

Hurford (1987, 1990) put concepts of language involving social dimensions on a more concrete footing, by placing Chomsky's (1972) concept of a Language Acquisition Device within a social context. Chomsky's conceptualization of language acquisition (Figure 2.3) neglects to specify how the primary linguistic data is produced. Hurford noted that this data is produced by other speakers of the language, and so Chomsky's diagram can be extended to produce Figure 2.4. (Figure 2.4 is adapted from Hurford (1987), p22.)

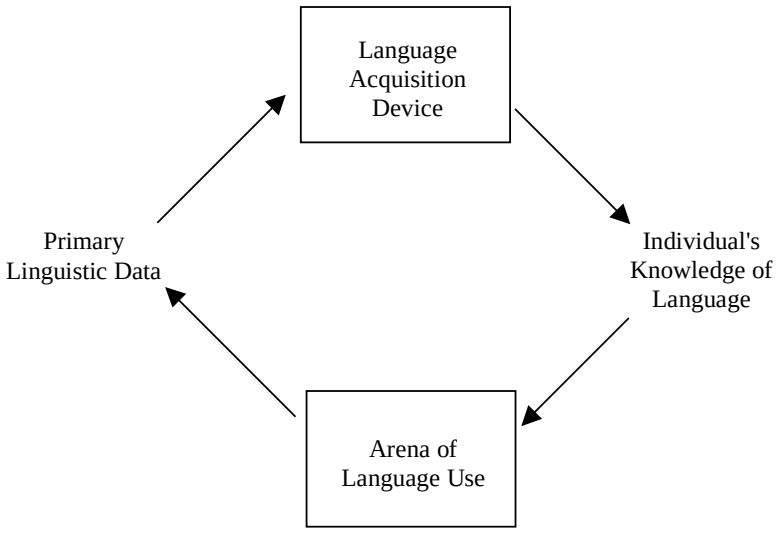


Figure 2.4 Hurford's Diachronic Spiral

The new component in Hurford's concept of language, which is missing from Chomsky's, is the arena of language use, through which language passes from one generation of speakers to the next. The arena of language use is partly psychological, but it also includes all those factors relating to the context in which we communicate, and the aspects of the world which determine what we talk about, and so what utterances (or written language) are produced to form the primary linguistic data for

the next generation of speakers. These factors all affect people's I-languages, because they determine what input is available to the Language Acquisition Device.

One way in which the arena of language use might influence a language, is related to the frequency with which particular aspects of language are used. If a construction is used frequently, then it is almost certain to be acquired by the next generation of speakers, but if it is used rarely (perhaps because there is little demand for its meaning to be expressed) then it is likely that that construction will be lost from the language. One example which might be a possible instance of this phenomenon is the loss of the irregular past tense form of *geld* in English. In Middle English this was *gelt*, but while other similar irregular past tense forms have been retained (such as *dwelt*, though some speakers might prefer *dwelled*), *gelt* is no longer a part of the English language. This is presumably because most present day speakers of English have little cause to refer to gelding, and so the past tense form has at some stage not been used frequently enough for it to be included in the language of the next generation of speakers (Pinker, 1994). Factors such as this will shape language, and will determine the range of human languages which exist in the world, but they are certainly not properties of the Language Acquisition Device.

Hurford's model makes explicit the route through which diachronic[17] change in languages must occur. Clearly, a new construction (or lexical item) must initially be

---

[17] *Diachronic* is used within linguistics to mean 'development through time', and usually refers to changes which affect a language (or a variety of a language) as a whole. (Changes to the languages of individual people, such as those which take place when a child learns a language, are not considered to be diachronic, even though language acquisition is a from of development over time.) Diachronic is the

used by one particular person, and it will then form part of the primary linguistic data from which other speakers learn language. However, it will only become part of the language if it is then incorporated into at least some of those people's I-languages, a process which would involve the Language Acquisition Device. Hurford's diagram most obviously corresponds to the situation where one generation of speakers is passing on language to a following generation, but that need not necessarily be the case. While the major change in a person's I-language clearly happens during acquisition, adults may modify their languages in response to the speech of other people, and children probably learn much of their language from their peers. Hence the linguistic data from which any one person acquires, and subsequently modifies, their language, is likely to be produced by people of a variety of ages, and in some cases two people will each learn from linguistic data produced by the other. However, all of these situations can still be considered to be part of the process depicted by Hurford's spiral.

Hurford's spiral retains Chomsky's I-language concept, but it allows for more factors to influence languages, and gives a wider definition of the scope of linguistic inquiry. Instead of simply studying the process of language acquisition, and the resulting I-languages, we can now extend the study of language to explain how language evolves under the influence of not only the mechanisms which individual people use for perceiving and producing E-language, and learning I-language, but also under the influence of non-linguistic factors, which will also affect the evolution of language.

---

opposite of *synchronic*, which is used to refer to linguistic approaches which do not consider language change, but simply study languages as they are at one point in time.

Explanations of linguistic phenomena under Chomsky's model are limited to a single generation, but Hurford's spiral allows for the possibility of explaining language universals and language typology in terms of diachronic processes. Hurford's diachronic spiral suggests that it may not be possible to understand language simply in terms of E-language or I-language, but that both concepts may be needed. Below I discuss a computational model, Hurford (2000), which shows that compositional regularities (and by analogy many other kinds of linguistic phenomena), may be apparent in E-language, even if they do not exist as part of the I-language of the person who produced that language.

Kirby (1999) discussed some aspects of linguistic typology which seem to be better explained within Hurford's diachronic model of language than in Chomsky's purely ontogenetic approach. Kirby sought to explain a number of typological implicational universals as the result of evolutionary processes. One such implicational universal, observed by Keenan and Comrie (1977), noted that there exists a hierarchy concerning which positions in a sentence structure are available for relativization in particular languages. This is shown in (2.1), where the syntactic positions most often available to relativization appear on the left, and where there exist increasingly fewer languages which allow relativization in each subsequent position in the hierarchy. In fact, all languages appear to allow relativization of subjects, and if they allow relativization of noun phrases in any other syntactic position, then they also allow relativization of all the syntactic positions left of that position in the hierarchy.

(2.1) Subject > Direct Object > Indirect Object > Oblique > Genitive > Object of Comparison

(2.3), (2.4) and (2.5), all of which are derived from (2.2), exemplify subject, direct object and oblique relatives respectively, demonstrating that English allows relativization in all of these positions. (The position from which the relativized noun phrase has been extracted is marked *t*.) English in fact allows relativization of noun phrases in any of the syntactic positions in the hierarchy.

(2.2) The linguist wrote the book about the language.

(2.3) The linguist who [*t* wrote the book about the language]

(2.4) The book which [the linguist wrote *t* about the language]

(2.5) The language which [the linguist wrote the book about *t*]

Keenan and Hawkins (1987) conducted a psycholinguistic study which suggested that people find it easier to parse sentences containing relative phrases when the relative phrase is further left in the hierarchy, rather than further right. In languages which do not allow relativization of noun phrases in some of the positions in the hierarchy, other mechanisms may be available which would allow equivalent meanings to be expressed. This might involve the use of syntactic processes such as passivization, which can promote noun phrases to positions in which they are relativizable, or speakers might resort to using circumlocutions. However, there is presumably some cost for the speaker in applying these extra operations when generating sentences, if only because they normally require the addition of extra morphemes such as passive markers. The actual cost of such processes could be expected to vary from language

to language, depending, for example, on how morphologically complex passives are in particular languages. Hence, whether the cost of such a transformation was greater than the added cost of relativizing a noun phrase in a syntactic position further right in the hierarchy, would vary from language to language.

We might expect that factors affecting the cost to speakers of forming particular kinds of relative clauses would influence which positions would most often be relativizable. Whichever position has the least cost for the speaker when it is relativized, could be expected to be relativizable in the greatest number of languages. However, there remains a need for an explanation of how an option which had a lower cost for speakers could come to be selected as the only permissible option in a language. It does not seem likely that people would not be able to learn languages of types unattested on the hierarchy, simply because those languages included constructions with a higher cost to speakers, when alternative constructions with lower costs were not permitted. For example it would seem to be unlikely that people would be unable to learn a language which permitted relativization of obliques, but not of subjects. This language might seem perverse, because it allows a construction with a high parsing cost (oblique relatives), while blocking one with a lower cost (subject relatives), but there does not seem to be a reason why such a language should be unlearnable[18].

---

[18] If we considered the hierarchy from the perspective of universal grammar, then we could hypothesize that the hierarchy was unlearnable because some innate principle of grammar did not permit it. It is quite possible that this is indeed the correct explanation for the hierarchy, but this is not an explanation in terms of cost to speakers of constructions, which is the kind of mechanism on which I wish to focus here. Of course, it is possible that an explanation of the phylogenetic evolution of universal grammar

Kirby proposed that the implicational hierarchy is the product of a cultural evolutionary process. He suggested that people are more likely to acquire constructions which have a lower cost, either in terms of parsing, or complexity of utterances. If this were the case, then we would expect there to be an evolutionary pressure which would influence those types of relative construction to the left of the hierarchy to occur more often than those further right. The types of relative constructions further right in the hierarchy would be included only if that method of expressing the required meanings had a lower cost than the alternative of using other syntactic structures to express the same meaning. Kirby implemented a computer model which incorporated the competing pressures regarding ease of parsing and morphological complexity, and showed that languages with subject relatives only, and languages with both subject and object relatives, tended to emerge, as did languages with no relative clauses at all, but that languages which had object relatives but not subject relatives tended not to occur.

If languages do indeed evolve culturally under the kinds of pressure which have been identified by Kirby, then the languages actually occurring in the world will be a subset of the logically possible languages. This is illustrated in Figure 2.5, which is reproduced from Kirby (1999, p121), and in which the set of logically possible languages is represented by box E. All of the languages which actually exist in the world will fall within the intersection of the learnable languages, (L), and those

---

would involve reference to the same kinds of costs which Kirby has used to explain the hierarchy as the product of cultural evolution. However, this thesis is concerned with cultural evolutionary processes, not phylogenetic ones, so I will not go into that issue here, though it is discussed in Kirby (1999).

languages which are preferred as a result of evolutionary pressures, (F). This illustrates clearly that both psychological and communicative processes can influence languages, and hence factors of both types can potentially explain aspects of language typology. The model of colour term evolution described in this thesis aims to explain colour term evolution from this kind of perspective. However, before constructing the model, it was necessary to determine what kind of psychological pressures might affect colour term evolution. The next section reviews research which gives some indication of likely psychological pressures.

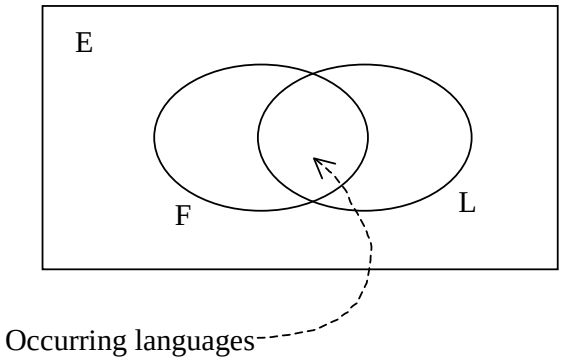


Figure 2.5 Interacting Constraints on Possible Languages

2.3 Psycholinguistic and Neurophysiological Findings

Attempts have been made to relate the findings concerning basic colour terms and their distribution across languages to more fundamental properties of the human mind and the human visual system. Hering (1964) noted that red and green appear to be opposite colours, and that the same is true for yellow and blue. This is because no colour can appear to be simultaneously red and green, or simultaneously yellow and blue. (For example, no colour can be described as reddish-green or yellowish-blue.) We can, however, perceive colours to be a mixture of two colours if those colours are not opposite, so we might describe orange as reddish-yellow, or lime as yellowish-green. Hering's observation established that the four colours, red, yellow, green and

blue, play a special role in the human visual system, and so we might expect them to have a special status in colour language. Hering proposed that the opponency of red and green, and of yellow and blue, was due to the physiology of the human visual system, although he did not have any direct evidence of this.

De Valois and Jacobs (1968) provided direct neurophysiological evidence to support Hering's proposal, by conducting experiments on Macaque monkeys. Macaque monkeys have visual systems very similar to those of humans. In the retina, both humans and macaque monkeys have three different types of light sensitive 'cone' cell, each of which responds maximally to light of one particular wavelength (Thompson, 1995). De Valois and Jacobs measured the outputs of cells in the presence of light of various wavelengths, and from these measurements they were able to infer the existence of two types of cell which processed the outputs from the cones. Firstly, they proposed that there existed *nonopponent cells*, which added together the outputs of the three types of cones, to indicate the blackness or whiteness of a light. Secondly, they proposed that other cells, which they termed *opponent cells*, subtracted the output of one type of cone from that of another. They proposed that there were four varieties of these opponent cells. The cells could either oppose red and green or yellow and blue, and each cell responded maximally in the presence of one of the colours which it opposed, and minimally in the presence of the other, so that for each opposition there were two polarities.

De Valois and Jacob (1968) hence identified six colours which would result in maximal or minimal firing rates for one type of opponent or nonopponent cell, and these colours appeared to correspond to the colours on which the prototypes of most colour terms are clustered (see section 2.1). This showed a correspondence between

neurophysiology and language, but it did not explain completely how the visual system affects language.

It was noted above that Kay and McDaniel (1978) attempted to make a direct link between the outputs of cells in the retina and the denotations of colour terms. They proposed that the output of the opponent cells would map directly to fuzzy set membership in colour term categories, but this suggested that, for example, all red terms in every language would have identical denotations, which is clearly not the case. It seems that it is only the prototypes of colour term categories which are consistent across languages, while category boundaries are much more variable. Hence most theories have simply assumed that neurophysiological processes give a special status only to the colours which produce maximal firing rates in opponent and nonopponent cells. The four such chromatic colours, red, yellow, green and blue, are generally referred to as *unique hues,* and the points in the colour space at which they occur as *unique hue points.*

The typological literature had already established the existence of unique hues, because it was noted that there are certain colours on which the prototypes of colour terms tend to be clustered. However, this does not necessarily entail that the prototypes occur in those places due to the influence of maximal firing rates in the retina. Indeed, it would seem that such an explanation is missing a number of steps which would be needed to explain exactly how a low level physiological response comes to influence language. A further problem arises because some researchers, including Kay and Maffi (1999), have stated that the locations for unique hue points which are predicted by the neurophysiological evidence are not consistent with those points at which the prototypes of colour terms tend to be clustered. However, Hardin

(1988) seems confident that colour term typology ultimately has its explanation in neurophysiology, and he discusses evidence which appears to show that the neurophysiological evidence does correctly predict the location of the unique hues.

Saunders and van Brakel (1997) have questioned the claim that there are exactly two types of opponent cells, suggesting that the evidence for yellow-green opponents in particular is not clear, and that there also exist cells focused on other hues such as orange. Both Kay and Maffi and Saunders and van Brakel's arguments suggest that the clustering of colour term prototypes in certain parts of the colour space might in fact not be due to neurophysiological effects. In this thesis I leave open the question of exactly what causes unique hues to have a special status. Regardless of whether or not the location of unique hues is determined by opponent cells, there is plenty of evidence to support their existence, coming not only from the typological literature discussed above, but also from a wide body of psychological research, which I briefly review below.

Firstly, some simple psychophysical experiments appear to demonstrate the existence of unique hues, and the opponency of red and green and blue and yellow. De Valois and De Valois (2001) report that if a unique red and a unique green light are added together in equal proportions, they will cancel each-other out, so that a neutral grey colour is perceived. Furthermore, after staring at a red surface, a green after image will be seen. Similar effects are observed for yellow-blue opponency.

More recent evidence has come from work aimed specifically at explaining properties of colour language. A number of psychological experiments have been conducted which have demonstrated the special status of the unique hues, especially those studies conducted by Rosch (some published under her earlier name of Heider),

including Heider (1971, 1972) and Rosch (1973). Heider (1971) investigated whether the colours which were consistently chosen as the prototypes of colour categories in Berlin and Kay's (1969) study were more salient than other colours. She did this by showing three year-old children rows of colour chips, which were all either of the same lightness or saturation, and asking them to pick out any chip. She found that the prototype colours were picked out much more often than would be expected just by chance, so she argued that those colours were more salient because they attracted the children's attention. In a further study, this time using mainly four year old children, a child would be shown colour chips one at a time, and asked to point each one out on an array of Munsell chips. The children pointed out the correct chip on the array most often when the colour they were trying to match was a prototype colour, which again demonstrated the special properties of those colours.

Heider (1972) was concerned that some of the effects showing the special status of prototype colours, might be due to their status as the prototypes of linguistic categories, and hence be a product of colour terminology, rather than a cause of typological restrictions on colour term systems. She sought to investigate whether this was the case, by performing experiments testing the memorability of colours with mono-lingual speakers of Dani, which, as noted in section 2.1, has only two basic colour terms[19]. Each subject was shown a test chip for five seconds, and then that chip was removed. After an interval of 30 seconds, subjects were asked to pick out the

[19] A problem with this approach is that, while Dani has only two *basic* colour terms, it also has non-basic colour terms, and so these terms could well interfere with the results of the experiments. However, probably no language really has only two colour terms, so it is likely that linguists will never investigate a language which approximates a two colour term language better than Dani does.

same colour from an array of Munsell chips. Heider found that the subjects were most accurate at choosing the correct colour when they had been shown a prototype colour chip, which replicates the results obtained for American children in the similar experiment described above. This was taken as evidence confirming the hypothesis that these colours are more memorable for all people, regardless of what language they speak.

Heider (1972) also investigated whether Dani speakers would find it easier to learn names for prototype colours than for other colours, by conducting an experiment in which they were taught to associate words with both prototype and non-prototype colour chips. Subjects were shown 16 colour chips, eight of which corresponded to the eight chromatic prototype colours, while the other eight were from parts of the colour space which generally did not form the prototypes of colour terms. The experimenter gave a name to each chip, which the subject was asked to repeat. The chips were then presented to the subject in a random order, and he or she was then asked to name each chip. If they gave the wrong name for a chip, then they would be told the correct one. The subjects were tested in this way five times a day until they got all of the colour names correct. The mean number of errors made by subjects was significantly greater for non-prototype than for prototype colours, showing that people do find it easier to learn words for prototype rather than non-prototype colours. Heider suggested that this showed that these colours were more easily kept in long term memory, which is presumably why they come to form the prototypes of colour terms. In the studies, Heider looked at the red, yellow, green, blue, purple, orange, brown and pink prototypes, but she suggested that the prototypes corresponding to unique hues were more memorable than the other ones, which seemed to suggest that their special status was due to the influence of the neurophysiology of colour vision.

Rosch (1973) investigated people's ability to learn artificial colour categories made up of five adjacent chips in the Munsell array. In some categories, prototype colours were central, but in others they were either more peripheral, or the categories did not include any prototype colour at all. In order to avoid interference from other colour terms, Rosch again performed her experiments with mono-lingual Dani speakers, using the same procedure as for the task of learning names for single colour chips reported in Heider (1972), but presenting each chip in a category separately during the repeated testing phase. It was found that the subjects learned categories in which prototypes were central with the fewest errors, and those which did not contain a prototype worst. Learning of sets in which the prototype was peripheral was intermediate. Again Rosch found that there were fewer errors made in learning when the prototype colour corresponded to a unique hue than for the other four prototypes which she investigated, demonstrating the primacy of those four hues.

Roberson, Davies and Davidoff (2000) sought to replicate some of Rosch's work, but this time using speakers of Berinmo, a language with just five basic colour terms, as well as British English speakers, as participants. Firstly, they repeated an experiment from Heider and Olivier (1972), in which desaturated colour chips (which generally would not be expected to contain category prototypes), were shown for five seconds, and 30 seconds later participants were asked to pick out the same colour from a Munsell array. Roberson et al found that when subjects made errors they chose adjacent chips with the same linguistic label significantly more often than adjacent chips with a different linguistic label, which showed that the language spoken by the participants was having an influence on their responses.

Roberson et al then sought to replicate the memory task of Heider (1972), by again showing test chips for five seconds, and 30 seconds later asking participants to pick out the chip from a Munsell array of maximally saturated colour chips. They found that both English and Berinmo speakers chose the correct chip more often for prototype colours than for other colours. However, it was shown that Berinmo speakers had a tendency to choose prototype chips, regardless of whether the test chip was a prototype one. When the bias towards choosing prototype chips was taken into account, it was shown that Berinmo speakers did not remember prototype colours more accurately than any other ones; it just seemed that they did because they tended to pick out prototype colours from the array, regardless of whether the test chip was a prototype one or not.

Roberson et al then conducted a further study to test the discriminability of prototype colours. The study was similar to Heider's (1972) experiment with four year old children. Participants were shown colour chips, and, while they could still see the chip, they were asked to point them out on the array. Both English and Berinmo speakers performed better at this task for prototype colours than for other colours, suggesting that these chips were more discriminable in the array. This suggests that all of the colour chips were not equally perceptually spaced in the Munsell array, and hence that many of the effects reported by Rosch might have been produced by properties of the array, rather than for any more fundamental reason.

Lucy (1992) reported that, if a different Munsell array was used, which had been corrected so that all chips were equally discriminable, then in most cases prototype colours were not remembered better than other colours. However, even with the corrected array, some groups of subjects did still perform better for the prototype

chips in memory tests, so the hypothesis concerning the special memorability of these chips still receives some support. Roberson et al found that when the discriminability advantage for focal colours was removed, neither English nor Berinmo speakers performed better in the memory task. However, it seems difficult to conclude that any one of these colour arrays is the correct one to use, as there are many ways of measuring distances between colours (MacLaury, 1997a), and there is no obvious reason why one of these should be the single correct method. Hence it would seem that we can produce or remove effects concerning the special status of the prototype colours simply by changing the colour array used in experiments.

The problems resulting from the relative discriminability of different colours could be avoided by performing experiments in which it was not necessary to discriminate between colours. Roberson et al repeated Heider's (1972) experiment in which participants were taught names for individual colour chips, but found, contrary to Heider, that Berinmo speakers did not perform better at learning names for prototype colours than for other colours. However, this result may have been obtained simply because most of the participants performed very poorly at the task. In a variation of the experiment, in which Berinmo speakers were taught to associate colours with pictures of nuts, the red prototype was consistently learned better than the non-prototype colours. This could be taken as evidence supporting the hypothesis that it is a neurophysiological factor which gives unique hues their special salience, if it was not the case that the red term was the Berinmo colour term with the most consistent prototype, and that this prototype corresponds to the universal red prototype. Hence, this suggests that this result was a product of colour vocabulary, and not of any pre-linguistic property of prototypical red.

Roberson et al's study has cast doubt on much of the evidence concerning the universal special status of prototype colours. However, while the results of the psychological experiments may be inconsistent, and the neurophysiological evidence only suggestive of a special status for certain colours, the cross linguistic evidence clearly demonstrates the universal properties of the prototype colours, because, cross-linguistically, most people place the prototypes of their basic colour terms on the universal prototype colour chips, or else chips immediately adjacent to them. The evolutionary computer model described in this thesis aims to give an explanation of colour term typology, suggesting that it is the product of evolutionary processes occurring under the influence of neurophysiological biases. The model rests on the assumption that the red, yellow, green and blue unique hues are especially salient, which is supported by some neurophysiological, psychological, and linguistic evidence, even though not all of the studies are in complete agreement.

2.4 Expression-Induction Models of Language

The evolutionary computer model is, to use a term introduced by Hurford (2002), a kind of expression-induction model. These models aim to simulate the process of language change, usually over a number of generations. They contain a number of artificial people[20], each of whom is capable both of learning some aspect of language, and also of using the language which they have learned to express themselves, hence

[20] Artificial people are more commonly referred to as *agents*. However, I believe that the term *artificial people* is preferable, partly because its meaning is more transparent, but also because *agents* is a much less specific term, being used to refer to quasi-autonomous parts of computer programs, most of which are not supposed to simulate real people at all.

creating some example utterances from which other artificial people can learn. Usually expression-induction models are run several times, so that the general properties of the languages which emerge in them can be observed. If all the emergent languages have a particular property which is also a universal in real languages, or if the emergent languages show a limited range of variation, reflecting typological patterns seen in real languages, then the models can be said to explain why these universals or typological restrictions exist[21].

Probably the first computer model which could be classified as an expression-induction model is that of Hurford (1987). In this model, there were ten individual people, and the meanings which they tried to express were the numbers between 11 and 20. At the start of the simulation each person knew the numerals for 1 to 10, so the aim of the simulation was to investigate in what way the language would develop to allow larger values to be communicated. Numbers were expressed by combining two individual digits to make a phrase, the meaning of which would be the values of each digit added together. So, for example, 15 could be expressed as 'seven-eight' or as 'six-nine'.

---

[21] An interesting exception to this generalization is the work of Harrison, Dras and Kapicioglu (2002), who created an expression-induction model that simulated the evolution of vowel harmony in Turkic languages. The aim was to initialize the model with the artificial people knowing a harmony system similar to that of an early form of Uzbek, and then to investigate what factors could have led to the growth and then subsequent decay of a vowel harmony system in this language. (Such changes are attested by modern and historical texts showing present day and early forms of Uzbek.) This simulation was therefore concerned with changes in one specific language, as opposed to most (if not all) other expression-induction models, which are concerned with the generic case of human language in general.

The simulation proceeded in a number of stages. In each stage one person would be chosen to be the speaker, and another the hearer. A number between 11 and 20 would then be chosen, and the speaker would communicate this number to the hearer. Initially the speaker would be equally likely to use any combination of numbers which express the right value, but if they had heard a numeral being used to form such expressions more frequently than other numerals, then they would use that numeral whenever possible. It was discovered that, after a short period of time, all speakers in the community would express all the numbers between 11 and 20 using the numeral ‘ten’ and one other numeral. This is exactly the system which is found in real languages, where these values are typically expressed with a morpheme which appears to be derived from the word ten, together with one other digit. (English displays this pattern for the ‘teen’ numbers, such as ‘fourteen’, which appears to be derived from ‘four’ and ‘ten’.)

Hurford’s model (Hurford, 1987) is important because it shows how a universal rule could evolve as a result of a diachronic process, even though individual speakers have no obligation to follow the rule. By the end of the simulations, a standard language had emerged which was shared by all speakers, simply due to each person listening to the language of the other people, and trying to express himself in the same way. The induction part of Hurford’s model was extremely simple, as the artificial people learned simply by keeping count of how often they had heard each numeral. More recent expression-induction models, including the model of colour term evolution described here, have begun to use much more sophisticated learning techniques, and I will briefly review some of the most important of those models. However, first I will mention another approach to modelling the evolution of languages over time, one that does not simulate evolution at the level of individual conversations between speakers,

but which can in some circumstances produce more rigorous results than those achieved with the expression induction methodology.

Nowak, Komarova and Niyogi (2002) describe work in which they try to determine some of the general properties that languages, and the mechanisms that humans use to learn them, must have, in order for coherent languages to emerge. They created a formalism to specify the scope of possible human languages, and created a measure that could calculate the communicative payoff when speakers of any two languages communicated. (If the languages were more similar, successful communication was more likely, and so communicative payoff would be higher.) They then introduced, a measure of how likely it would be that each learner would acquire each particular language, given the language spoken by the learner's parents. How likely a learner would be to acquire the same language as its parents would depend on the acquisition mechanism that he or she used. Nowak et al considered the case of a very poor learner, who was worse at learning than people are, and a very good learner, who learned, in some sense, perfectly. They then argued that people's ability to learn languages must come somewhere in between these two extremes, and so considering these cases allows us to place bounds on the actual mechanism that people use to learn languages. Using this technique, Nowak et al were able to show how restrictive universal grammar needs to be in order for coherent languages to emerge, and to demonstrate that this is related to the number of example sentences to which learners are exposed.

Perhaps the main criticisms that I would make of Nowak et al's work, is that the learning mechanisms that they consider are non-statistical, and hence they do not make use of any of statistical information in the input data. It would be surprising if

people learned languages in this way, as statistical patterns can provide a rich source of information about the target language, and so any person who ignored them could be disadvantaged. However, some other assumptions made by Nowak et al are also somewhat problematic. For example they assume the differences between languages can all be reduced to a series of binary parameters, but it seems difficult to see how some aspects of languages, such as lexical entries or phonetics, can be represented in this form. Furthermore, they assume that, in terms of a similarity measure that they define, all distinct languages are equally dissimilar, something that in reality is certainly false. Such mathematical models allow some bounds to be placed on the properties that languages can have, and on the learning mechanisms that people use to learn them, but the validity of any such findings is dependent on the assumptions made when constructing the model. Furthermore, the range of problems that can be addressed with such models is probably smaller than that which expression-induction models, more of which are reviewed below, can address. In particular, it would probably not be possible to construct such a model to explain the data concerning colour term typology that is the subject of much of this thesis.

Kirby (2002) created an expression-induction model to investigate compositionality. In all languages, the meaning of an utterance as a whole can be derived from the meanings of its individual parts, and Kirby sought an explanation of how languages having this property arise. Most of his simulations contained only two artificial people at any one time. One of these would be an adult, who had completed the process of language acquisition, and the other would be a child, who would try to learn the language spoken by the adult. In the initial state of the simulation, neither the adult nor the child knows any language, so when the adult first speaks he will have to begin by making up some new words. The meanings to be expressed take the form of simple

first order predicate logic formulae, such as *eats(tiger, sausages)*, which would correspond to the meaning ‘the tiger eats sausages’, or slightly more complex formulae such as *thinks(gavin, loves(gavin, mary))* which would mean ‘Gavin thinks he loves Mary’.

The simulation would begin with the adult expressing random meanings. Whenever the adult did not know how to express a meaning, they made up a random string of letters, and simply used that. (They also remembered the correspondence between the letter string and the meaning for later use). The child would observe all these letter strings together with the corresponding meanings, and would then derive a grammar from them. A child’s grammar consisted simply of a long list of all the meanings which he had observed, together with the strings of letters which were used to express the meanings, except that, where possible, he would try to form more general rules. If two strings with similar meanings both contained repeated letter sequences, then it could be possible to replace each of these with a variable, and add two new rules to expand this variable in such a way that it could express either of the two meanings. Initially, such similarities in any two such strings would appear simply by chance, but the creation of such a rule begins to add hierarchical phrase structure into the language, so such regularities would be likely to be preserved in future grammars.

After the process of induction was complete, the child then became an adult, and began the process of talking to a new child, who would similarly start from the point of not knowing any language at all. Eventually small compositional grammars emerged which were able to express an infinite number of meanings through recursion. Kirby (2002) suggested that this could explain why languages show compositionality. We should note that there is nothing in the model which prevents

non-compositional languages from being learned, and that neither adults nor speakers ever receive any reward for successful communication. Compositionality emerges not because the Language Acquisition Device constrains the learnable languages in such a way that only this kind of language can be learned, but because these languages are passed on more easily. Because each child hears only a finite number of example sentences, compact compositional languages are more likely to be passed on accurately to the next generation than larger non-compositional ones.

Hurford (2000) developed a similar model, but his model had one key difference. Instead of the artificial people always forming general rules whenever they encountered utterances supporting the formation of such generalizations, they would only do so on 25% of such occasions. This created an inbuilt bias to conservativeness, as the people did not always generalize as much as they potentially could. Hurford's simulations contained four adults and one child at any one time, and periodically the child would be promoted to be an adult, the oldest adult removed, and a new child added. This meant that each child would learn its language based on input from four different adults. We might expect that if there was only a weak bias towards forming general rules, then idiosyncratic lexicalizations would tend to emerge, where instead of complex meanings being expressed by general rules, they would be expressed using a single lexical item, in which the complex meaning had been paired with a phonological form.

Whether we find such non-systematic sound-meaning pairings in the emergent languages depends on whether we look at emergent E-languages or I-languages. If the utterances produced in the community (after a sufficient number of generations had elapsed), were examined, then it was clear that the language had a compositional

structure without any redundant lexicalizations, so that complex meanings always conformed to rules, and the meanings were a function of the meanings of their component parts.

However, these regularities were not always mirrored in the I-languages of the individual speakers. Typically I-languages contained some general productive rules, but they also contained many rote learned lexical items pairing whole utterances with their meanings. If the I-languages of each of the adults were compared, then presumably each adult would have learned a somewhat different I-language, because which constructions had been analyzed, and which simply memorized, was determined randomly. However, these people all spoke the same E-language, and these E-languages were compositional, because compositional rules were internalized by the speakers sufficiently often for the compositional structure of the languages to be maintained. I believe that these simulations are extremely important, because they demonstrate very clearly divergence between E-language and I-language, and show that rules apparent in E-language may not have any psychological reality for the speakers who produce that E-language. They hence demonstrate the importance of conceptualizing language within a framework which acknowledges both E-language and I-language, and they show that some aspects of language cannot be understood if either only E-language or only I-language is considered.

Another recent expression-induction model is that of de Boer (1999). This model is, in many ways, similar to the one presented in this thesis, in that it aims to explain cross-linguistic typological patterns as a product of a cultural evolutionary process. However, de Boer's model is concerned with the typology of a very different domain, namely that of vowel systems. Instead of simulating evolution over several

generations, as Kirby (2002) did, de Boer created 20 artificial people who were present for the whole of the simulation. At the beginning of the simulation, each of these people would know several random vowels. The initial vowels would be different for each person, but each person always knew the same number of vowels, usually between three and nine[22]. The vowels were represented in terms of the positions of the first and second formants[23] which reflects the most important acoustic properties by which the vowels of real languages are distinguishable. However, this representation is somewhat simplified, especially because it does not take account of lip rounding, which forms another dimension in the vowel systems of many languages.

The basic assumption behind the simulations was that, if the artificial people were made to imitate each other, then, as a result of this process, coherent vowel systems would emerge. Imitation proceeded by first choosing one person to be the initiator, and choosing one of the vowels which he knew. That person would then try to express a vowel similar to this, but altered slightly, to another person. That person would then find the vowel in their language which was most similar to the vowel which they perceived, and then express that same vowel back to the initiating person. In each case, the presence of random noise in the environment, which might distort the speech

---

[22] I should note that de Boer (1999) describes several versions of his simulations, in some of which the number of vowels was not fixed beforehand.

[23] *Formants* are frequencies at which the voice resonates, and so which are characteristic of the overall sound of the vowel. The first and second formants correspond to the two lowest such resonant frequencies (Ladefoged, 1975).

signal, was simulated by slightly altering the vowel being expressed. If the vowel which was most similar to that now perceived by the initiator was the same as that which the initiator initially expressed, then the process of imitation would be judged to be successful. This *imitation game* would be played with every other simulated person, and from this an overall score would be obtained for how successful the people were at imitating that vowel. If the actual vowel expressed was more successful than the one on which it had been based had been in the past, then the initial vowel would be altered to be a bit more similar to the expressed vowel. In any case, the scores for how successful the vowel had been were updated.

This imitation process was repeated many times, typically about 25000 times, at which point the vowel systems know by each person tended to be very similar. Most of the vowel systems also appeared to conform to most of the typological patterns reported in the literature, especially in that they tended to be symmetrical (with there being a corresponding back vowel for each front vowel[24]), and the vowels tended to be evenly distributed in the vowel space, in such a way as to be maximally acoustically distinct. De Boer concluded that his model showed that cross-linguistic typological patterns, which it had been previously argued were the product of innate properties of the human language faculty, actually emerged through a process of self-organization within a population.

---

[24] The terms *front* and *back* refer to the relative position in which the tongue is placed in the mouth in order to pronounce the vowel. Front vowels are pronounced with the tongue further forwards than when back vowels are pronounced .

While both de Boer's model, and the evolutionary model of colour terms described in this thesis, aim to explain language typology using same kind of computer modelling methodology, the expression-induction model which is most relevant to the model of colour term evolution presented in this thesis is that of Belpaeme (2002). Belpaeme's model also looks at the cultural evolution of colour term systems over time[25], although the details of the model are somewhat different from the model presented here, as are the aspects of colour term systems for which it is able to account. Belpaeme's simulations typically contained ten artificial people, each of whom was able to represent colour categories using adaptive networks, a kind of neural network. Colour in the model was represented in terms of the CIE-LAB space, which represents colour in terms of three dimensions, one of which corresponds to its degree of redness or greenness, one to the degree of yellowness or blueness, and the third to the lightness or darkness of the colour[26]. The networks acted as fuzzy membership functions, allowing colour categories corresponding to a volume of the three dimensional CIE LAB space of almost any size or shape to be represented. Each artificial person could also remember a number of word forms, each of which could be paired with a particular colour category.

---

[25] Belpaeme also models the phylogenetic evolution of innate colour categories, but this work is less relevant, mainly because it is less similar to my model of colour term evolution. However, because there is considerable variation in the ranges of colours denoted by similar colour terms in different languages, it seems unlikely that all these colour terms could be innate. Hence, I will proceed on the assumption that colour categories are learned, and ignore Belpaeme's work on innate colour categories.

[26] This colour space was chosen because Lammens (1994) showed that his computer model of colour naming worked best in this space.

In the initial state of the simulation, the artificial people did not know any colour categories or colour words, so, the first time one of them spoke, he would have to create a new category and corresponding word. In general, communication proceeded by first choosing one colour to be a topic, and another to a context, and then choosing one person to be a speaker, and another to be a hearer. The speaker would then try to communicate to the hearer which of the words was the topic, and which was the context, by choosing a word which included the topic, but not the context, in its denotation. If the word that the speaker used was known by the hearer, and the colour category which the hearer had associated with that word included only one of the colours, then the hearer would understand that that colour was the topic. If this was correct, then communication would have been successful, and the association between the topic colour and the colour word would be strengthened. If communication was not successful, then the hearer would be shown the correct topic, and the word's colour category would be adapted, so that it would be a better match for the topic colour. Categories and words which were persistently not useful in communication would eventually be forgotten.

In some simulation runs, the same artificial people would exist for the whole of the simulation, though in others evolution over a number of generations would be simulated, by periodically replacing one of the people with a new one who had not learned any colour words. However, similar results were obtained in both these conditions. The most important result was that, over a period of time, coherent colour lexicons emerged which were shared by all the artificial people. The colour lexicons would divide the colour space into a number of colour regions, each of which would be associated with a particular colour word. The people never agreed completely about the exact meaning of each colour word, but their languages were consistent

enough for them to achieve rates of communicative success in excess of 85%. However, the colour categories emerging in Belpaeme's model did not resemble the colour terms of real languages, as they did not conform to the typological restrictions observed in colour term systems cross-linguistically[27].

The model of colour term evolution presented in this thesis builds on the work of Belpaeme (2002). It uses a similar methodology, in that it is also a kind of expression-induction model, but there are a number of key differences. The learning mechanism used is Bayesian inference, not adaptive networks, and there is no feedback mechanism which informs the artificial people whether communication has been successful. Instead they simply try to mimic each other's language, though they receive no reward for successful imitation. Some of the findings of the psychological studies of colour perception and colour naming reviewed above were incorporated into the model, and this allowed it to account for much of the typological data concerning colour words, as will be shown below.

---

[27] Belpaeme (2002) did suggest that the split into light and dark colours seen in languages with only two colour terms might be explainable in terms of his model, because this might be the easiest way to divide up the colour space, but, in its present form, the model would not be able to account for any other aspects of colour term typology.

Chapter 3

A Bayesian Model of Colour Term Acquisition

The acquisitional part of the expression-induction model of colour term evolution shows how children can learn the denotations of colour words based on a number of examples of colours which those words have been used to denote. The model demonstrates that, if colour words were learned using Bayesian inference, then this would cause them to have prototype properties. When the model was designed, attention was paid to the psycholinguistic, neurophysiological and typological evidence, in the hope that this would enable it to account accurately for the empirical data concerning colour terms.

3.1 Bayesian Inference

The acquisitional model learns using Bayesian inference, which is a statistical procedure that allows empirical evidence to be used to determine how likely it is that hypotheses are correct. It derives its name from Bayes' rule of Bayes (1763), given in (3.1) below, although Barnett (1982) notes that Bayesian inference itself is a more recent development, which cannot readily be attributed to any single person.

$$(3.1)\quad P(h \mid d) = \frac{P(h)P(d \mid h)}{P(d)}$$

Bayesian inference has previously been used as the basis of psychological models of learning, including such works as Anderson and Matessa (1991), who modelled categorization, Griffiths and Tenenbaum (2000), who looked at how people can predict the frequency of periodic events, and Tenenbaum and Xu (2000), who modelled language acquisition. All of these models produced results that closely parallel human's performance on the same tasks, suggesting that people use Bayesian inference to accomplish those tasks. Studies such as these are the primary reason for presuming that colour words are learned using Bayesian inference – if people use Bayesian learning in one domain, it seems likely that they will use it for learning in other domains as well.

Bayes' rule, (3.1), allows the probability that a hypothesis is correct with respect to some relevant data to be calculated. (This is called the *a posteriori* probability of the hypothesis, and is written as $P(h \mid d)$). However, Bayes' rule can be applied only if we know the correct values for all the terms on the right hand side of the equation. Firstly we must know how likely the hypothesis was before we considered the data (its *a priori* probability, written $P(h)$), and how likely we would be to observe that data if the hypothesis were correct, (the probability of the data in terms of the hypothesis, written $P(d \mid h)$). We also need to know how likely the data was anyway, not with respect to just one particular hypothesis, but in terms of all possible hypotheses. (This is the probability of the data, written $P(d)$.) In general we cannot really know for sure exactly what value we should assign to each of these probabilities, so instead we must make reasonable estimates.

We should note, however, that here the Bayesian model of colour term acquisition is being used not as an objective inference procedure, but as a psychological theory. Hence, when an assumption is made in the design of the Bayesian model, it is effectively a proposal that people implicitly make that assumption when they learn colour words. Clearly in some cases, given the totality of our knowledge of colour term systems cross-linguistically, more accurate assumptions could be made, but in the present context 'more accurate' does not necessarily correspond to 'more correct'. In general, when assumptions are made, the aim is either to make the assumption which it seems most likely that children must make in order to learn colour words, or else just to make the simplest assumption which seems reasonable, in accordance with the principle that we should prefer simpler theories over more complex ones, unless there is evidence to the contrary[28].

Looking at equation (3.1) in more detail, we can see that the *a posteriori* probability of a hypothesis is directly proportional to both its *a priori* probability, and the probability of the data with respect to that hypothesis. In other words, hypotheses which are likely to be correct before we have observed any data, are still more likely to be correct once we have seen the data, all other things being equal, while hypotheses which predict the occurrence of the kind of data which has been seen, are more likely to be correct than those which predict that such data is unlikely to be observed.

---

[28] This principle, which is sometimes termed Occam's Razor, is widely supported in the literature on philosophy of science, and was discussed above in Chapter 2.

The third term on the right hand side of the equation is the probability that the data has, before any particular hypothesis is considered. So long as the available data is constant, the value of this term does not change, even when we consider alternative hypotheses, and so, if we were only interested in the relative probabilities of two or more hypotheses, we could ignore this term. However, so long as we know the full range of possible hypotheses, and how the data can be assigned a probability with respect to each one, the value of this term can be calculated using a standard Bayesian procedure, allowing the exact *a posteriori* probabilities of hypotheses to be determined, not just their relative probabilities. Implementing Bayesian inference consists simply of performing the calculation specified by equation (3.1), but in order to be able to do that, we need to identify how the values of the terms on the right hand side of the equation can be determined.

Perhaps Bayesian inference can be made somewhat clearer through the use of an example. Suppose I was to see a person in the street who had green skin and antennae growing out of his head. There are two possibilities which could explain his odd appearance, each of which is a hypothesis, and which I will term h_1 and h_2. Firstly the person might be an alien from outer space *(*hypothesis h_1*)*, or secondly the person might be a human (hypothesis h_2). (Of course there could be other possibilities, but for reasons of simplicity, I will assume that there are only two possible hypotheses.) In general, I expect that most people I see on the street are not aliens, so I will estimate that the probability of someone I see on the street being an alien to be one in a billion, so $P(h_1)=0.000,000,001$. Given that we are considering only two probabilities, this means that the chances that the person is really a human are 999,999,999 in a billion, so $P(h_2)$=0.999,999,999.

Now we have to consider the available data, which is that the person has green skin and antennae protruding from his head. Firstly, we need to consider the case of the person being an alien, and determine a probability for how likely we would have been to observe the data (the green skin and antennae), if he were an alien. I would guess that, if a person was an alien, it would be fairly unlikely, but not inconceivable, that he would have green skin and antennae growing out of his head. Hence, I will estimate that the probability of making these observations if we know the person to be an alien, is one in a thousand, so $P(d|h_1)$=0.001. If the person is really a human and not an alien, then it would seem that he would be very unlikely to have green skin and antennae, but these are still possible, because he could simply be dressed up in an alien costume. Hence I will estimate the probability of a human I met on the street having green skin and antennae to be one in a million, so $P(d|h_2)$=0.000,001.

Now, if we want to determine how likely it is that the figure is an alien, we must start by determining the probability of the data, $P(d)$. We do this by working out the probability of observing the data regardless of which of the two hypotheses is true, by calculating the probability of both the first hypothesis being true and of us observing the data if that hypothesis was true, and adding this probability to the equivalent one for the second hypothesis, as shown in (3.2).

(3.2) $P(d)=P(h_1)P(d|h_1)+P(h_2)P(d|h_2)$

Substituting the values identified above into (3.2), as in (3.3), produces a probability for the data which is approximately equal to one in a million. Such a very low probability should not be a surprise – after all it is not very often that we see people with green skin and antennae on the street.

(3.3) P(d)=*0.000,000,001*×0.001+0.999,999,999×0.000,001=0.000,001,000,000,999

The final stage of the Bayesian inference procedure is to calculate the actual probabilities of each of the hypotheses, by substituting into equation (3.1), as in (3.4) and (3.5).

(3.4) $P(h_1 \mid d) = \frac{P(h_1)P(d \mid h_1)}{P(d)} = \frac{0.000,000,001 \times 0.001}{0.000,001,000,000,999} = 0.000,001$

(3.5) $P(h_2 \mid d) = \frac{P(h_2)P(d \mid h_2)}{P(d)} = \frac{0.999,999,999 \times 0.000,001}{0.000,001,000,000,999} = 0.999,999$

We can see that we have inferred that it is much more likely that the person is in fact a human dressed up rather than an alien, even though humans very rarely dress up as aliens. This is because the *a priori* probability of seeing an alien on the street is so much lower than of seeing a human on the street. Even though it is much more likely for an alien than a human to have green skin and antennae, the calculations reveal that the probability of the person being an alien is only one in a million. We should note that as we assumed *a priori* that either h_1 or h_2 must be correct, the sum of the *a posteriori* probabilities of these hypotheses is one, indicating that one or the other one must be correct.

We should perhaps consider what it would take to make me believe that what I had seen was in fact a real alien. Suppose that, after passing the person, a flying saucer landed in the street behind me and picked him up. This would provide more evidence concerning whether the person was really an alien or not. It would seem that the probability of a flying saucer landing in the street and picking up a human would be very low (let us say one in a billion), but if there is an alien in the street, the probability of him being picked up by a flying saucer would be much higher (say one

in a thousand (0.001)[29]). Using this new data, we can update the probabilities we earlier calculated for hypotheses h_1 and h_2, which will now be the *a priori* probabilities, because they were determined before consideration of the new data. Firstly the probability of the new data must be calculated using equation (3.2), as in (3.6), and then this value, can be used to calculate the *a posteriori* probabilities for each hypothesis, given both the data about the person's appearance, and the data about the flying saucer. These calculations are shown in (3.7) and (3.8).

(3.6) $P(d)$=0.000,001×0.001+0.999,999×0.000,000,001=0.000,000,001,999,999

$$(3.7)\quad P(h_1 \mid d) = \frac{P(h_1)P(d \mid h_1)}{P(d)} = \frac{0.000,001 \times 0.001}{0.000,000,001,999,999} = 0.5$$

$$(3.8)\quad P(h_2 \mid d) = \frac{P(h_2)P(d \mid h_2)}{P(d)} = \frac{0.999,999 \times 0.000,000,001}{0.000,000,001,999,999} = 0.5$$

We can now see that it is equally likely that the person is a real alien as that he is just a human dressed up as an alien. This seems to be reasonably in accord with intuitive judgments. Of course we do not normally expect to pass aliens on the street, but then we do not normally expect to see flying saucers either. Having observed such a highly unlikely event, we might be prepared to consider the possibility that the person was in fact a real alien. More importantly, this example has illustrated how Bayesian inference can be applied in practice, and how the overall *a posteriori* probability that is assigned to a hypothesis is determined both by its *a priori* probability, and by the data which has been observed. While Bayesian inference may seem a very technical

[29] I have still assigned quite a low probability to this data because, assuming that aliens do exist and visit the Earth, I still think that it is quite unlikely that they travel by flying saucer.

and abstract procedure, people make inductive inferences all the time, so we must have some mechanism for doing this, even if we are not consciously aware of it, and in this thesis I am suggesting that this mechanism is Bayesian. While the alien example probably does not demonstrate clearly what advantage Bayesian inference would have over other, perhaps simpler, mechanisms which might be used to perform inference, we can understand some of the key benefits of Bayesian inference by looking more closely at the work of Tenenbaum and Xu (2000) and Griffiths and Tenenbaum (2000).

Tenenbaum and Xu (2000), in common with the approach of this thesis, used Bayesian inference to model the acquisition of word meanings. Their model learned meanings from examples of objects which words were used to refer to, which is similar to how the Bayesian model of colour terms learns, though Tenenbaum and Xu's model was concerned with concrete nouns. 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 which they had observed. If a person had observed only a small number of examples of a word, then they would be unsure of just how far beyond those examples the word's denotation extended, but as a larger number of examples were seen, the learner would become more confident about the word's exact denotational range. Importantly, Tenenbaum and Xu were able to demonstrate, using psychological experiments, that the generalizations made by their model were very similar to those made by people when presented with the same evidence. This provides some of the clearest empirical support for the hypothesis that Bayesian inference is an important mechanism in language acquisition, and hence is very supportive of the potential of the Bayesian model of colour terms.

Griffiths and Tenenbaum's (2000) model is even more closely related to the work in this thesis, although their work is not itself concerned with language. Griffiths and Tenenbaum investigated how people can predict the frequency with which some event occurs, based on observations of the time since its last occurrence. Through psychological experiments, they found that 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. 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, they are likely to believe that trains run with a frequency much closer to 103 seconds. Tenenbaum and Griffiths attempted to account for this data using a Bayesian model similar to the one presented in this thesis, but in which 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, by averaging over all possible hypotheses, the model arrives at a best guess for the actual interval between train arrivals.

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 both have continuously varying values. Hence, it was possible to 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. Griffiths and Tenenbaum found that their model accurately replicated the data obtained from human subjects, and so this again is strongly supportive of Bayesian models of learning, and hence of the Bayesian model of colour terms.

The model of colour term semantics also has many similarities with some of the models which Tenenbaum (1999) used to explain the acquisition of concepts, especially one of those models, which is described in Appendix C.2 of Tenenbaum (1999), and which was used to model the acquisition of 'disjunctive concepts' in a one dimensional space. Tenenbaum and Griffiths (2001) have proposed that these models might exemplify a universal principle which people may use whenever they generalise from examples.

Their work has built on the work of Shepard, especially Shepard (1987), who sought to develop a universal law for generalisation. His models only applied to generalisation between a single exemplar and a novel stimulus, while the model of this paper, and the models of Tenenbaum (1999), Tenenbaum and Griffiths (2000) and Tenenbaum and Xu (2000) could all generalise based upon multiple exemplars. However, Shepard's (1987) model was, as will be seen below, similar to the model of this thesis, in that it proposed that people make the *a priori* assumptions that categories are all of equal size, and are equally likely to occur anywhere in the conceptual space. Further discussion of the differences between the model of Shepard (1987) and the more recent Bayesian models can be found in Tenenbaum and Griffiths (2001). The examples of other models of Bayesian learning all provide evidence to support the Bayesian model of colour term semantics, but the primary motivation for that model is of course that it accounts well for the specific empirical data concerning colour terms themselves.

3.2 Axioms of the Acquisitional Theory

This section outlines the assumptions made in the construction of the Bayesian model of acquisition, which correspond to the axioms of a theory of how people learn the

meanings of colour terms. The first of these axioms has already been introduced, because that is simply that the meanings of colour words are learned using Bayesian inference[30]. However, this still leaves open the question of what knowledge children have pre-linguistically, what evidence is available to them from which to learn, and what implicit assumptions they make about the possible denotations of colour words. The assumptions (or axioms) made in the construction of the model, concerning each of these of these questions, are the topic of this section. These axioms are summarised in Table 3.1, and each of them is then discussed in turn in the subsequent subsections.

[30] We might also describe this 'axiom' as the hypothesis under test. However, it is not possible to test one component of the model in the absence of other parts of the model, as the results obtained are based on the output of the model as a whole. Hence, it is difficult to determine the effect of each axiom, and to predict how the results might have been different if an alternative assumption had been made. Therefore, it might be best to regard the hypothesis under test as the whole theory, including all of its axioms. Chapter 8 discusses how we should interpret the results of the model, and to what extent they can be taken as supportive of the axioms, and of the theory as a whole.

1	Children can see colour, and have available a conceptual colour space, before they begin learning colour terms.
2	Children learn colour terms by observing other speakers use of these terms.
3	The evidence from which children learn is unreliable.
4	Unique hues are more salient to children than other colours. Hence children are most likely to remember examples of colours named by colour terms, when those colour terms are used to name unique hues. The unique hues are not evenly spaced in the colour space, with the green and blue unique hues being closest together, and the red and blue ones the furthest apart of any of the adjacent pairs.
5	Children assume that colour terms denote contiguous regions of the colour space.
6	Children learn colour term denotations using the Bayes' optimal classification form of Bayesian inference.

Table 3.1. Axioms of the Acquisitional Theory.

3.2.1 A Conceptual Colour Space

The first axiom of the model is that, before people begin to learn the meanings of colour terms, they must be able to see colours, and to understand the relationships between different colours. It seems that in order to accomplish this we must have some sort of conceptual colour space in which we can think about colour, and in which some colours will be closer to each other than are others.

Whilst the colour of light, which is what ultimately gives rise to our experiences of colour, is dependent on the wavelength of the light, where red light has the longest wavelength and purple the shortest, this does not reflect how we experience colour, because perceptually red and purple are similar colours. Colour, as we experience it, is best understood as having a three-dimensional structure, where it can vary on any of the dimensions of *hue*, *saturation* or *lightness* (Thompson, 1995). *Hue* is a circular dimension, in which the colours vary from red to orange, yellow, green, blue, purple

and finally back to red again, as shown in Figure 3.1, which has been labelled with the basic colour terms of English. The other two dimensions are both orthogonal to hue. *Lightness* simply corresponds to how light or dark a colour is, while *Saturation* corresponds to the degree to which a colour is free from dilution by grey, so that very bright colours are high in saturation, while black, white and grey have a zero degree of saturation. The Bayesian model is not at present concerned with the dimensions of saturation or lightness, so the relevant colour space is the one dimensional circular hue space of Figure 3.1.

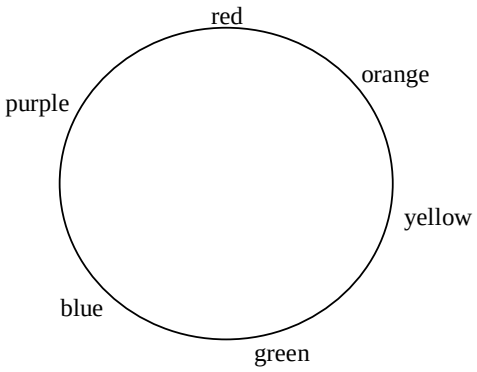


Figure 3.1. The Conceptual Colour Space.

It is proposed that people have such a conceptual colour space before they begin the process of learning colour words. This proposal is supported by evidence from Xiao, Wang and Felleman (2003), who showed that there are parts of Macaque monkey's brains in which cells responding maximally to different colours of light are spatially organized in the same order as in the conceptual colour space, starting at red, and finishing at purple, although is not clear that they demonstrated the existence of a circular colour space, in which purple was adjacent to red. The circularity of the hue dimension is, however, well supported by psychological evidence. Such a conceptual

space would allow inferences to be made, such as that red is more similar to orange than it is to yellow, and that we can have an 'orangey-red', but not a 'reddish-green' (because orange and red are adjacent in the colour space, while red and green are not). Within this colour space, it is possible to model the meanings of the English basic colour terms *red, orange, yellow, green, blue* and *purple*, but not terms which are distinguished from these in terms of lightness or saturation, such as *pink* or *brown*, or the achromatic terms, *black, grey* and *white*.

We should note, however, that the relationship between the physical properties of light entering the eye and perceived colour is not straightforward, but is moderated by a number of intervening processes. Much is known about the physiology of the colour vision system, and there is a number of theories to account for phenomena such as perceived colour constancy despite varying illumination[31]. However, consideration of

---

[31] One of the best known theories concerning colour constancy is that of Land (1977). Shepard (1994) has even suggested that humans may have developed a colour system in which colours can vary along three separate dimensions because such a system makes it easier to maintain colour constancy. Because natural illumination also varies primarily in three ways (strength of illumination, degree of redness as opposed to greenness, and degree of blueness as opposed to yellowness), it is easier to determine the colour of objects, despite seeing them in varying light conditions, if our vision system also measures light in terms of those same three dimensions. Therefore 'such a three-dimensional representation of colour may have emerged as an adaptation to a pervasive and enduring feature of the world in which we have evolved' (p20).

Shepard (1992) describes experiments which investigated whether the colour space used by colour blind people who lacked the ability to distinguish between red and green was also three dimensional, or whether, as might be expected as they are unable to discriminate colours which differ on one of the three dimensions of colour, only two dimensional. Firstly he showed subjects sheets of coloured paper,

such issues is outside of the scope of this thesis, so here it is simply assumed that such processes exist, and that they are able to convert the physical rays of light entering the eye into an internal representation which can then be categorized linguistically. The details of these processes are not relevant to the issues surrounding colour vocabulary which are being considered here.

A related problem concerns objects which are not composed of a single colour, but instead have varied colouration, or when different parts of the same object have different colours. Most objects in a child's environment probably have such characteristics, for example trees, houses and animals. However, basic colour terms

---

two at a time, and asked them to rate the similarity of each pair. Then he repeated the experiments, but this time using the names of two colours, such as *red* and *orange*. Analysis of the subjects' judgements as to how similar each pair of colours was, was then used to determine the shape of the colour space that the subjects were using. The first experiments, using coloured paper, produced evidence of a colour space, in which red and green were collapsed together. This is what would be expected, as the subjects lacked the perceptual apparatus necessary to distinguish red and green. In contrast, in the second experiments, the subjects appeared to be using a colour space in which red and green were distinct, and which had the same general properties as that used by non-colour blind people. Shepard took this as evidence that the full three dimensional colour space is innate, as the colour blind subjects presumably could not have learned it from perceptual input. However, I discuss in Chapter 8 work by Landau and Gleitman (1985) that appears to show that much knowledge of colour can be gained from linguistic context, so it could be the case that colour blind individuals are able to construct a three dimensional colour space using such evidence, and that the three dimensional colour space is not innate.

denote ranges of pure hues[32], so a child trying to infer a term's denotation must be able to determine either which part of an object is being referred to, or to abstract an overall colour from a part of an object which has varied colouration. It would seem that such processes must be in place before a child can learn the meanings of colour terms, but this thesis does not seek to explain how those processes work, but again simply assumes that they are in place before the process of colour term acquisition begins[33].

3.2.2 Evidence from which Children Learn

The next issue to be considered is what data is available from which people can learn colour words. It seems a reasonable assumption that children are not taught the full range of the denotations of each word they know explicitly, in terms of exactly what it can and cannot be used to denote. Instead, it seems more plausible that children learn the meanings of colour words primarily by observing other people' uses of those words. If a person utters a phrase such as 'that red chair', and a child listening to this is able to determine the referent of the expression, then they will be able to tell that

---

[32] We should qualify this statement somewhat, because, as was noted above, authors, such as Levinson (2001) have noted that, in some languages, other properties are conflated with colour, so that words do not refer just to hue. This would appear to make the child's learning task more difficult, because they must determine whether a word refers just to colour, to a mixture of colour and other properties, or perhaps not even to colour at all.

[33] It should be noted that it is also possible that, to some extent, the colour vision system develops whilst colour terms are being acquired.

the colour of the chair is within the denotation of the colour term *red*[34]. From a number of such observations, it would be possible to obtain several examples of colours that the word *red* can be used to denote, and so the input to the process of acquiring the meaning of each word will consist of a number of example colours. Learning will then consist of generalizing from these examples to the full range of colours which come within the word's denotation.

This approach of using meaning-form pairs as input data has been followed by many other researchers when creating computer models of the acquisition of meaning. For example Morris, Cottrell and Elman (2000) trained a neural net to acquire grammatical relations by pairing sentences with representations of their meanings. The neural net was then able to learn to interpret novel sentences, by generalizing from the example sentences. Kobayashi, Furukawa, Ozaki and Imai (2002) used a very different learning mechanism, inductive logic programming, in a computational model of word meaning acquisition, but they also trained their model with form-meaning pairs. Kobayashi et al assumed that as well as getting positive examples concerning what words mean, children also received negative examples, specifying what words do not mean. However, this assumption is problematic, because many researchers report that some children do not receive any such corrections, and yet they still acquire language successfully (Guasti, 2002). That is why the present model learns using only positive examples.

---

34 This ignores the problem of determining that the word red is a colour term at all, which might be as complex a problem as actually determining the word's denotation, but this issue is outside of the scope of this thesis (though it is discussed further in Chapter 8).

3.2.3 Unreliability of Data

We should note that there is a lot of potential for error in the procedure by which a child attempts to determine the colour which a colour term has been used to identify. For example, children might incorrectly infer the referent of a colour term, or the speaker who they were observing might use the wrong colour term to identify a particular colour. Hence, it would seem that if children are to be successful in their acquisition of colour words, they must assume that the data from which they learn is unreliable. This means that there will be a possibility that any particular example which they observe is erroneous. Hence another axiom of the acquisitional model is that children will believe that there is only a certain probability that each example they observe is accurate.

3.2.4 Salience, Memorability and Location of Unique Hue Points

It was noted above (in section 2.3) that there is a number of colours known as *unique hues*, which have a special status psychologically. These hues are at the points in the colour space corresponding to the best examples (prototypes) of the English words *red, yellow, green* and *blue* (and are commonly the prototypes of colour terms in other languages, as discussed in section 2.1). We should note that, in Figure 3.1, these colours are not evenly spaced in the conceptual colour space. This is because another axiom of the acquisitional theory is that the unique hues are not all equidistant. More precisely, it is proposed that pre-linguistically, the green and blue unique hue points are conceptually close together, whilst the red and blue unique hue points are the furthest apart of any two adjacent unique hue points, and that the red-yellow and yellow-green distances are intermediate.

The motivation for this proposal is primarily that it results in an explanation of the typological patterns, as will be shown below. MacLaury (1997a) reports that there is some evidence to suggest that the green and blue unique hues are in some way closer than any of the other hues are to each other, although evidence concerning the conceptual distances between the other unique hue points is less clear. This issue is problematic, because we are trying to measure distances in a subjective conceptual space, and there is no clear objective way to do this. MacLaury surveyed a range of literature, which attempted to address this issue, but from his survey it is clear that there is no generally agreed method for measuring distances in colour spaces.

Furthermore, it is well established that, how people perform on some cognitive tasks related to colour, is affected by the way languages categorize colour. Roberson et al (2000, p394) state that distances within the conceptual colour space 'are stretched or distorted by the influence of linguistic categories'. This indicates that relativistic effects might interfere with any attempt to establish the exact location of the unique hue points, because performance on any psychological task concerning colour cognition could be affected by the language spoken by the subject. Hence, the primary motivation for these conceptual distances is the typological evidence, because, as will be shown below, if these distances are incorporated into the model, the model is then able to explain the typological patterns.

It is, however, worth noting that at least one measure of the distance between unique hues does provide some support for the locations implemented in the acquisitional model. Boynton and Olson (1987) gave values for the distances between the

centroids[35] of the areas on the surface colour space named by each of the English basic colour terms *red*, *yellow*, *green* and *blue*. These were 6.5 between green and blue, 7.3 between yellow and green, 11.2 between blue and red, and 12.2 between red and yellow. The rank of these distances is identical to the rankings used in the model, except that, as noted above, in the model, blue and red are placed somewhat further apart than red and yellow. Boynton and Olson's results might therefore be seen as providing support for the model, but we should be cautious about making such an interpretation, because there exist a number of different methods for measuring such distances, each of which produces somewhat different results (MacLaury, 1997a).

Above it was noted that the unique hues have special properties, in that there is a considerable body of evidence to suggest that colours corresponding to the unique hues are especially salient, and are especially well remembered. This suggests that people would not be equally likely to remember examples of all colours when learning colour words, but would be more likely to remember colours when they were unique hues. This corresponds to another axiom of the acquisitional model, which is that people will forget, or will never remember, a certain proportion of example colours, and that the proportion of colours remembered will be greater for colours at unique hue points than for colours in other parts of the conceptual colour space.

---

[35] The centroids are measures of the central tendencies of all the colours which would be assigned a given name, so they do not correspond exactly to category prototypes, as the prototypes are not always in the centres of the ranges of colour named by each colour term. Boynton and Olson (1987) do not give equivalent figures concerning category prototypes, but the distance values for the centroids should give some indication of the approximate values for the distances between prototypes.

3.2.5 Possible Colour Term Denotations

Finally, before we can apply the Bayesian procedure, we need to specify a range of possible hypotheses. In the case of learning colour words, a hypothesis will correspond to the range of colour that the word denotes. Such a hypothesis can vary in size from taking up almost none of colour space, to including the whole of the colour space, and can correspond to any contiguous range of colours. Gärdenfors (2000) has suggested that it is a general property of concepts used by humans that they do not denote disjoint sections of conceptual spaces, so the restriction that hypotheses must correspond to continuous ranges of the colour space seems reasonable[36]. It is

[36] This might appear to be problematic, as some researchers have claimed that colour terms exist which include yellow and blue but not red or green in their denotations, or which include red and green but not yellow or blue (McNeill, 1972; Saunders and van Brakel, 1997). The denotation of such a colour term would not correspond to a continuous section of the hue circle, because red and blue, and yellow and green, are not adjacent to each other. However, in a thorough review of the languages which McNeill and Saunders and van Brakel claimed contained such terms, Bailey (2001) shows that there are in fact no such terms in any of these languages. There is a number of different explanations as to why it was thought that these languages contained either yellow-blue or red-green terms. Firstly, some of the languages contained yellow-green-blue terms, which are not problematic because their inclusion of green means that the range of colours which they denote is not discontinuous. Secondly, a colour term might be applicable to more colours in an extended sense, but not in its basic sense. For example, white wine is yellow, but this does not mean that the English word *white* denotes both white and yellow, simply that it has been semantically extended so that it can be used to describe a type of wine. Also, colour terms may have undergone another type of semantic extension so that they have secondary connotations. For example, English *green* can mean ‘inexperienced’, and so could potentially be applied to both yellow and blue coloured objects, but that does not mean that its meaning has been extended to include both yellow and blue. Bailey noted that some of the supposed red-green or yellow-blue terms, were simply words which had undergone one of these types of semantic extension, and so

proposed that children will consider all such hypotheses to be equally likely *a priori*, so that the model has no inbuilt bias to prefer colour terms corresponding to one part of the colour space as opposed to any another.

This final assumption might seem problematic, because the typological evidence (reviewed in section 2.1) reveals that colour terms are not equally likely to denote all regions of the colour space. However, we should note that this axiom concerns an assumption which children make in order to learn colour term denotations. Children would not have available information concerning colour term typology, and so would not know which ranges of the colour space most often correspond to colour term denotations. In the absence of any relevant information, it seems most likely that children assume all denotations to be equally likely.

3.2.6 Bayes' Optimal Classification

The final axiom of the theory is that people will decide which colours come within the denotation of a colour word using *Bayes' optimal classification* (Mitchell, 1997). Bayesian inference was outlined above, but in order to achieve optimal performance, we should not just consider the single most likely hypothesis. If we want to calculate

---

could be used to refer more widely than their core meaning would allow, giving the impression that they denoted discontinuous ranges of colour. Thirdly, the meanings of the terms could really be 'pale' or 'faded', so that they were not really basic colour terms at all. Finally, the Ainu (Japan) word *hu*, described by McNeill as a red-green term, is not a colour term at all, but in fact means 'raw', so this is a simple case of misanalysis of data. In general, the literature on colour term typology, reviewed above in section 2.1, does not report the existence of colour terms which have discontiguous denotations. Hence, there does not appear to be any clear evidence supporting the existence of colour terms denoting discontinuous regions of the colour space.

the probability that a colour term can be used to name one particular colour then we should add the probabilities of all the hypotheses which include that colour within the word's denotation, which will determine the probability that that colour is within the denotation of the word. This is because the word will be able to denote that colour, no matter which one of the hypotheses which include the colour within their ranges is the correct one. So long as our assumptions are correct, this will determine the probability as accurately as is possible given the available data, which is why this procedure is termed Bayes' *optimal* classification.

3.3 The Bayesian Model

In order to incorporate all of the above axioms into a Bayesian model, it was necessary to write equations for each of the terms on the right hand side of Bayes' rule in equation (3.1) on page 68 above. In order to do this, it was first necessary to specify more precisely what form the conceptual colour space takes, and what properties it has. Two separate computer models were in fact implemented, both based on the above axioms. The key difference is that the first did not give any special status to the unique hue points, and so did not implement axiom four. This section completes the specification of this first model, leaving the details of the second model, which incorporates all of the axioms, until Chapter 5.

3.3.1 Calculating Probabilities

For the purposes of constructing the model, hues will be numbered using an arbitrary numbering scheme, which has its origin (zero) in the red space, and which increases through orange, yellow, green, blue and purple, up to 100, where we return to the origin, as shown in Figure 3.2. When a learner observes a colour term example, it can be represented simply as a point in this colour space, so in the implementation of the

model, colour examples will be represented simply as (possibly fractional) numbers between 0 (inclusive) and 100 (exclusive).

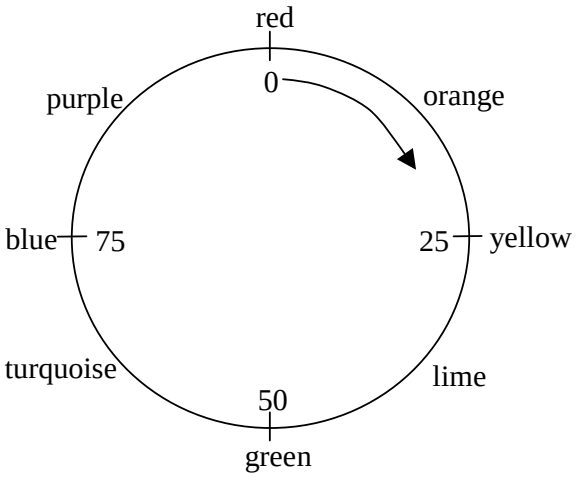


Figure 3.2. Indexing Colours in the Colour Space.

Using this coordinate system, hypotheses will be indexed as shown in Figure 3.3 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. 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. Such a situation is illustrated in Figure 3.4.

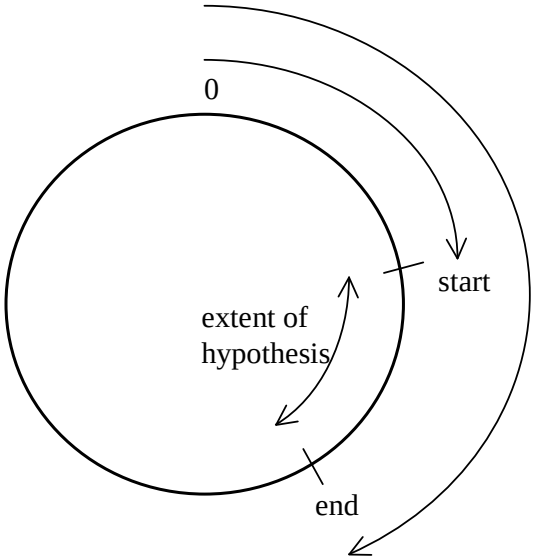


Figure 3.3. A Hypothesis as to the Denotation of a Colour Term.

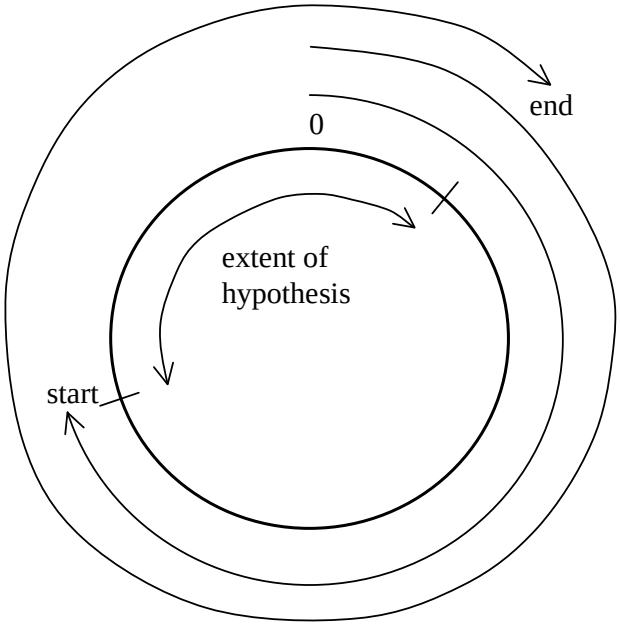


Figure 3.4. A Hypothesis which Crosses the Origin.

Given the above specification of hypotheses, it is now necessary to specify how probable it is that the examples of the colour term would have been observed if a particular hypothesis were correct, or in other words how to calculate values for the term *P(d | h)* from equation (3.1) on page 68 above. Firstly we need to assume that

children make no *a priori* assumptions that some colours are named by colour words more often than other colours are, and that a colour word is equally likely to be used to identify colours anywhere within the range of colours corresponding to its meaning. Even though some colours might be better examples of a word like *red* than others, this does not mean that the colours which are the best exemplars of the word will be named most often, so the second of these assumptions seems reasonable. However, the assumption that all colours are named with equal frequency would seem to be somewhat more problematic, as certain colours are almost certainly more frequent in a person's environment than others, and also linguistic reference to some colours is likely to fulfil greater purpose than reference to others. Children learning a language could compensate for such biases, by monitoring the abundance of particular colours in their environment, and so ideally the model would do so too, but the added complexity that this would entail seems unnecessary for present purposes.

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

probability of an example being observed in each section[37]. Each section is of equal size, and the sections do not overlap (but neither is there any gap left between neighbouring sections). We will use q to represent the number of such sections into which the colour space is divided.

If we can be sure that all the examples are accurate, then there is an equal probability of observing an example in any of the sections of the colour space within the range of the hypothesis. This probability, *P(example | example is accurate)*, is given by equation (3.9). We should note that this assumes that each hypothesis begins and ends at the boundary between sections, rather than somewhere within a section, so that each section is either wholly within, or wholly outside of, the hypothesis. It will be shown below that the number of such sections can be made to tend to infinity, thus making the colour space continuous, and so this assumption is unproblematic.

(3.9) $P(\textit{example} \mid \textit{example is accurate}) = \dfrac{100}{(e-s)q}$

However, it would seem likely that some of the examples which a child observes might not be accurate, and so, if learning is to proceed successfully in the presence of such examples, children must have some degree of expectation that any particular example might not be accurate. If an example were not accurate, then a child would have no way of knowing whereabouts in the colour space it would be observed, and so it is assumed that such examples would be equally likely to occur anywhere in the

[37] If we regarded the colour space as continuous, then there would be a zero probability of observing an example at any particular point, because there is an infinite number of points at which examples could appear, if we measured with sufficient accuracy.

colour space, and this probability, *P(example | example is not accurate)*, is given by equation (3.10). (We should note firstly, that, even if an example is erroneous, it could nevertheless come within the hypothesis simply by chance, and secondly, that equation (3.10) is independent of the hypothesis under consideration.)

(3.10) $P(example \mid example\ is\ not\ accurate) = \frac{1}{q}$

A child will not know which examples are accurate examples of the colour word, and which are simply random, so it is necessary to introduce a parameter, *p*, which corresponds to the degree to which a child believes an example to be accurate. This parameter can vary from 1, when the child will be completely certain that all examples are accurate, to 0, when the child believes that all examples are random. In the first of these situations, a single erroneous example could have a catastrophic effect on learning, because no matter how much it is at odds with the other examples, the model will never consider the possibility that it is erroneous. However, if *p* is set to zero, then no learning would occur at all, as the child would not believe that the examples gave any indication of the meaning of the colour word whatsoever. In this thesis, it is assumed that this parameter is always set to a value between these extremes, so that the model will believe that the examples are indicative of the meaning of the colour word, but it will still be able to learn even if some of the examples are misleading.

Examples which fall outside of the hypothesis must be inaccurate, and so their probability, *P(example outside of hypothesis)*, is given by multiplying together the probability that an example is not accurate, *(1-p)*, by the probability of observing an

inaccurate example given in equation (3.10). The equation resulting from this operation is given in (3.11).

(3.11) $P(\textit{example outside of hypothesis}) = \frac{(1-p)}{q}$

If a colour example comes within the range of the hypothesis, then it may be accurate, in which case the equation for its probability could be derived from equation (3.9), but it could also be inaccurate, so that it is only within the hypothesis due to chance, in which case its probability could be derived from equation (3.10). However, when a child is learning, they will not be able to be sure which of these two situations applies, and so must consider each possibility according to its probability as defined by the parameter *p*. We must derive an equation for the overall probability of an example, based on the possibilities of it being either accurate or inaccurate, with both of these possibilities being weighted in accordance with their probabilities, which are *p* and *(1-p)* respectively. The total probability of such an example, *P(example within range of hypothesis)*, will be found by adding its probability under each of these possibilities, which produces the equation given in (3.12).

(3.12) $P(\textit{example within range of hypothesis}) = \left(\frac{100p}{(e-s)} + (1-p)\right)\frac{1}{q}$

The equations so far are all concerned with only a single example. The probability of all the observed examples given a particular hypothesis (*P(d | h)*) can be found by multiplying together the probabilities of each individual example. Where there are *n* examples which come within the scope of the hypothesis, and *m* examples which come outside of the hypothesis, this probability can be calculated using equation (3.13).

$$(3.13)\quad P(d \mid h) = \left(\left(\frac{100p}{(e-s)} + (1-p)\right)\frac{1}{q}\right)^{n}\left(\frac{(1-p)}{q}\right)^{m}$$

We can extract *q* from the terms put to the power of *n* and the power of *m*, to create a new term of *q* to the power of *n* plus *m*. However, *n* plus *m* is always the total number of examples observed, and so this value will be constant across all hypotheses. Equation (3.13) is rewritten as (3.14) below, where *r* is used to represent the total number of examples.

$$(3.14)\quad P(d \mid h) = \left(\frac{100p}{(e-s)} + (1-p)\right)^{n}(1-p)^{m}\frac{1}{q^{r}}$$

Before going any further, we should perhaps consider what effect the incorporation of the accuracy parameter will have. Consider the situation represented in Figure 3.5, where the purple dots represent example colours, and where each arrow corresponds to a hypothesis. The larger hypothesis might be expected to have a higher probability, because it accounts for the location of all the examples, although it does not do so very precisely, because it includes such a large part of the colour space. However, so long as the model has a reasonably high expectation that some of the examples will be erroneous, the other hypothesis would in fact have a higher probability. This is because, although it does not predict correctly the location of two of the examples, it predicts where the other three examples are accurately. Incorporating an accuracy parameter into the model means that hypotheses which are able to predict accurately the location of most of the examples will generally have the highest probabilities, even if there are a few examples outside of their range.

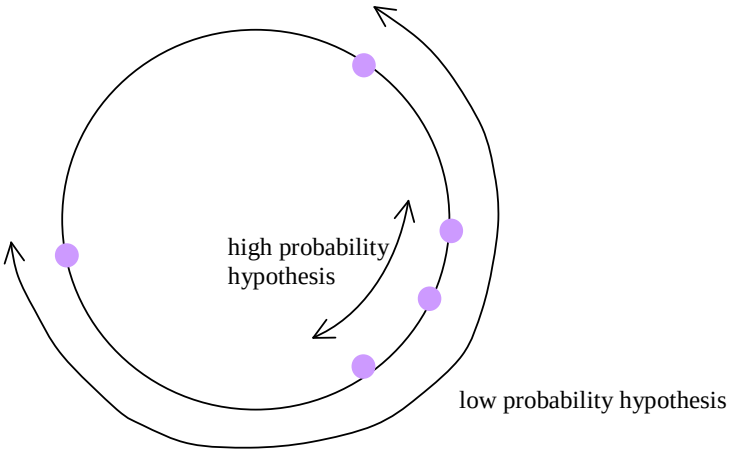


Figure 3.5. Hypothesis Probabilities with Erroneous Data.

So far, we have specified how probabilities will be assigned to two of the three terms on the right hand side of Bayes' rule (equation (3.1)), but, in order to calculate the probability of a hypothesis, we must also be able to calculate the *a priori* probability of the data, *P(d)*. We can do this by taking the product of the probability of the data given a hypothesis, and the *a priori* probability of the hypothesis, and finding the total of all these products for all the hypotheses. This is the same procedure as was applied in the alien example of section 3.1, though it is more complex, because we are considering more than two possible hypotheses.

Ideally we do not want to divide up the colour space into a number of arbitrarily sized sections, as there does not appear to be any empirical motivation for doing so. Hence it would seem desirable that we increase the number of sections in the colour space, *q*, so that *q* tends to infinity, and the colour space effectively becomes continuous. *P(d)* can then be calculated using calculus, as in (3.15), where *H* is the set of all possible hypotheses. Equation (3.15) simply states that, for each hypothesis, we should multiply its probability by the probability of the data in terms of it, and add together

all the resulting values. However, we do this over an infinite number of hypotheses. (We will see below that the term q will cancel out of the equations, and so its exact value is unimportant.)

$$(3.15)\quad P(d) = \int_{h \in H} P(d \mid h) P(h) dh$$

We now have all the terms which we need in order to calculate the *a posteriori* probability of a hypothesis using Bayes' rule. Substituting equation (3.15) into Bayes' rule we obtain equation (3.16), where the hypothesis whose probability is being calculated is now labelled h_i. All hypotheses have equal *a priori* probability, so the terms $P(h_i)$ and $P(h)$ will cancel out. The terms $P(d \mid h_i)$ and $P(d \mid h)$ also both contain the constant term q^r, and so this term will also cancel, as was noted above.

$$(3.16)\quad P(h_i \mid d) = \frac{P(d \mid h_i) P(h_i)}{\int_{h \in H} P(d \mid h) P(h) dh}$$

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

What is really needed is an expression for the probability that a colour can be named by the colour word which takes account of all the possible hypotheses. This can be achieved by using the procedure of hypothesis averaging, where the probability that a

colour can be named by the colour word if a particular hypothesis is correct, is multiplied by the probability of that hypothesis given all of the observed data. Equation (3.17) shows how the overall probability that the colour can be named by the colour word given all the data, $P(x \in C \mid d)$, can be found by summing over these products for all the hypotheses.

$$(3.17)\quad P(x \in C \mid d) = \int_{h_i \in H} P(x \in C \mid h_i) P(h_i \mid d) dh_i$$

The summation can be performed using calculus, because the colour space is continuous. However, the number of examples which are within the hypotheses changes discretely at the locations of the colour space where the examples occur, and the value of $P(x \in C \mid h_i)$ changes discretely at the location of the colour under consideration. For these reasons, the value of the sum must be calculated separately for sections of the colour space between such points, and then these values added together. Integrating the equations is fairly straightforward, and the full derivations, together with a description of how they are applied are given in section 3.4. However, firstly some other issues concerning the design and application of the model are discussed.

3.3.2 More than One Colour Term

So far the model has been described with respect to only a single colour term. This is because the denotation of each colour term is considered independently of all the others. Every observation is remembered in exactly the same way. The information recorded is always the name of the colour term, and the colour which the term was used to name. In considering the denotation of each colour term, account will be taken only of those colour examples which the model has observed it being used to name.

For every possible colour, a probability can be obtained for how likely it is 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, in most languages, basic colour terms have a property which might aid a child to acquire them, and that is that they partition[38] the colour space. Hence a child might be able to learn the denotation of one colour term, helped by examples of the use of another colour term, as they could infer that, where the denotation of one colour term ended, the denotation of another would begin. However, a major problem with this account is that it requires children to know that the colour terms they are learning conform to the partition principle. Clearly once children have learned the basic colour terms of their language, they will be able to observe with a reasonable degree of confidence that they do partition the colour space, and so they may be able to use the partition principle to consolidate their knowledge of the denotations of basic colour terms, but only when the process of acquisition is nearing completion.

If we were to propose that the partition principle played a more major role in the process of acquisition, then we would have to assume that children intuitively assumed it from the start of the acquisition process. This view would be consistent with the widely held belief that children are innately endowed with a Universal Grammar which specifies the general structure underlying all languages (Chomsky,

---

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

1986). However, if children do have available innate knowledge of the partition principle, this raises a number of further problems. Firstly, while basic colour terms typically partition the colour space, non-basic colour terms certainly do not. Hence a child would not only have to determine which words were colour terms, but also which of those colour terms were basic before they could use the partition principle to help in learning their denotations[39]. We would also then need to provide a separate account of how non-basic colour terms were learned. While this thesis is concerned principally with basic colour terms, the Bayesian model could equally well be used as an explanation of how non-basic colour terms are acquired. As there does not seem to be any reason to believe that non-basic colour terms are learned in a fundamentally different way from basic ones, it would seem preferable to have a single model which could account for the acquisition of both types of colour term[40].

A further objection to the use of the partition principle is that, as was noted above, there exist a very few languages where the basic colour terms do not partition the colour space, and hence where some colours cannot be identified by any basic colour term (Kay and Maffi, 1999; MacLaury, 1997a; Levinson, 2001). Hence it would seem

---

[39] One of Berlin and Kay's (1969) original criteria for distinguishing basic colour terms from non-basic ones was that the denotation of a basic colour term was not included in the denotation of any other colour term. Hence, it might seem more likely that the partition principle is used to differentiate between basic and non-basic colour terms once the denotations of those terms has been learned, rather than that the partition principle is used to learn those denotations.

[40] We could justify this using Occam's Razor, which was discussed above in Chapter 2. There would seem to be no benefit in advocating two separate mechanisms to explain the acquisition of colour terms, if the acquisition of all types of colour terms could be satisfactorily explained with a mechanism.

that children who relied on the partition principle would simply be unable to learn such a language correctly, or at least that the use of the principle would have a detrimental effect on their performance in correctly acquiring the language. Given all of these arguments, it seems preferable to proceed with a model which treats each colour term independently.

3.3.3 Deriving Fuzzy Sets

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, but so far no consideration has been given to how a semantic representation of a colour term might be derived. However, the Bayesian model implicitly defines a fuzzy set representation (Zadeh, 1965) for the denotations 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 corresponding to how likely it is 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[41]. Hence these values may be used to define fuzzy membership in sets corresponding to the colour terms. The use of fuzzy sets as a theoretical tool in

---

[41] Kosko (1994) stresses that the degree of membership in a fuzzy set is not the same as the probability of membership in that set. Hence, we should, perhaps, add another axiom to the theory. This is that people determine how good an example of a colour word a particular colour is, by considering the probability that that colour can be denoted by the colour word.

psychology is now well established (Stelmach and Vroon, 1988), and Kay and McDaniel (1978) had already used them to model colour term denotations (see section 2.3), both of which provide support for the proposal that colour terms denotations have fuzzy set representations.

There is 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[42]. The implications of these properties of the colour term's denotations are discussed in detail below, but now I move on to consider how the model may be practicably implemented on a computer.

3.4 Implementing the Model

While sections 3.2 and 3.3 specify the model in detail, they do not discuss how the model can be implemented in practice. In particular, equations (3.16) and (3.17) both

[42] We should note that this is a property of the fuzzy sets learned by this model, rather than a definitive property of fuzzy sets as such. Because the membership of the set is learned probabilistically, it is not possible to be entirely certain about whether any colour comes, or does not come, within the denotation of a particular colour term. However, in practice, people probably treat probabilities over a certain level as completely certain, and do not distinguish between being almost completely sure, and being completely sure. In reality, there is probably nothing of which we can be completely certain, as, for example, the whole world might be a hallucination, but this does not in practice seem to stop people from being 'sure' or 'certain' about many aspects of the world.

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 used to determine how likely it is that each colour of interest comes within the denotation of each colour term.

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

If we substitute the right hand side of equation (3.16) for the term $P(h_i \mid d)$ in (3.17), we obtain the equation (3.18).

$$(3.18)\quad P(x \in C \mid d) = \int_{h \in H} P(x \in C \mid h_i) \frac{P(d \mid h_i)}{\int_{h \in H} P(d \mid h) dh} dh_i$$

The value of the integral of $P(d \mid h)$ with respect to dh is not dependent on the value of h_i, and so is a constant term in the integration with respect to this variable. This allows equation (3.18) to be rewritten as (3.19).

$$(3.19)\quad P(x \in C \mid d) = \frac{\int_{h \in H} P(x \in C \mid h_i) P(d \mid h_i) dh_i}{\int_{h \in H} P(d \mid h) dh}$$

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 (3.19), 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 (3.19) 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 (3.20), 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 (3.18) by calculating *Px* and *Pnotx*, and substituting their values into (3.20).

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

3.4.2 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 | h)* over specific ranges of hypotheses, and substituting these values into equation (3.20) above, it is necessary to consider how these sums can be calculated in practice. While the symbols used in this section are defined in the text, for convenience, their meanings are summarised in Table 3.2 below. Consider again equation (3.14) (from page 96), repeated here as (3.21), in which some constants have been factored out. We should note that above (on page 98), it was shown that any constant terms in this equation will cancel out, including the term q^{-r}, and so throughout the rest of the derivations the term $\left(\dfrac{100}{q}\right)^r$ will simply be omitted. The 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.

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

Symbol	Meaning
n	The number of examples which come within the hypothesis.
m	The number of examples which come outside of the hypothesis.
r	The total number of examples (equal to $n+m$).
d	The data which is being learned from. (The set of all the examples.)
P	Used to stand for probability in general.
p	Each example is assumed to have a p probability of being accurate.
q	The number of sections into which the hypothesis space is divided. (This value tends to infinity.)
h	A hypothesis.
H	The set of all possible hypotheses.
H_i	Any subset of the set of all hypotheses.
s	The coordinate at which a particular hypothesis begins.
e	The coordinate at which a particular hypothesis ends.
s_1	The lowest value of s which any hypothesis in a particular subset of hypotheses can have.
s_2	The highest value of s which any hypothesis in a particular subset of hypotheses can have.
e_1	The lowest value of e which any hypothesis in a particular subset of hypotheses can have.
e_2	The highest value of e which any hypothesis in a particular subset of hypotheses can have.

Table 3.2. Summary of Symbol's Meanings

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. 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 worth 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 (3.21) for ranges of the variables s and e. Recall from section 3.2.5 (on page 87 above) 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 (3.21) for a range of values of s, and for each value of s, for a range of values of e. Recall from section 3.3.1 (on page 89 above) 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 (see Figure 3.3 and Figure 3.4 on page 91 above). So in some cases the value of e will be greater than the size of the colour space (100) . In most such cases, 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 (which is labelled s_2 below) 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. (3.22) 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 thesis, I will use H_i to refer to any particular range of hypotheses currently under consideration.

$$(3.22) \quad \int_{h \in H_i} P(d \mid h)dh = \int_{s_1}^{s_2}\int_{e_1}^{e_2} P(d \mid h)deds$$

If we substitute for the expression *P(d | h)* in (3.22), using (3.21), we obtain equation (3.23), which specifies 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.

$$(3.23) \quad \int_{h \in H_i} P(d \mid h)dh = \int_{s_1}^{s_2}\int_{e_1}^{e_2} \left(\frac{p}{(e-s)} + \frac{(1-p)}{100} \right)^n \left(\frac{(1-p)}{100} \right)^m deds$$

3.4.2.1 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 (3.24).

$$(3.24)\quad \int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m \int_{s_1}^{s_2}\int_{e_1}^{e_2}\left(\frac{p}{(e-s)}+\frac{(1-p)}{100}\right)^n deds$$

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 (3.25) to replace this term. (Note that the expansion of C_n^r is as specified by equation (3.26).) If we equate *a* with the first part of the binomial, as in (3.27), and *b* with the second part as in (3.28), we can observe the equivalence of the term to be integrated and the left hand side of equation (3.25).

$$(3.25)\quad (a+b)^n = \sum_{r=0}^{n} C_n^r a^{n-r} b^r$$

$$(3.26)\quad C_n^r = \frac{n!}{(n-r)!r!}$$

$$(3.27)\quad a = \frac{p}{(e-s)}$$

$$(3.28)\quad b = \frac{(1-p)}{100}$$

When we use expression (3.25) to substitute for the appropriate term in (3.24), the result is equation (3.29).

$$(3.29)\quad \int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m \int_{s_1}^{s_2}\int_{e_1}^{e_2}\sum_{r=0}^{n}\left(C_n^r\left(\frac{p}{(e-s)}\right)^{n-r}\left(\frac{(1-p)}{100}\right)^r\right)deds$$

By removing constant terms from the scope of the integrations, and removing the discrete summation from the scope of the continuous ones, (3.29) can be transformed into (3.30).

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

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 case when *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 (3.31).

(3.31)
$$\int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m \left(p^0 \left(\frac{(1-p)}{100}\right)^n \int_{s_1}^{s_2}\int_{e_1}^{e_2} \left(\frac{1}{(e-s)}\right)^0 deds + np^1 \left(\frac{(1-p)}{100}\right)^{n-1} \int_{s_1}^{s_2}\int_{e_1}^{e_2} \left(\frac{1}{(e-s)}\right)^1 deds \right.$$
$$\left. + \sum_{r=0}^{n-2} C_n^r p^{n-r} \left(\frac{(1-p)}{100}\right)^r \int_{s_1}^{s_2}\int_{e_1}^{e_2} \left(\frac{1}{(e-s)}\right)^{n-r} deds \right)$$

Simplifying terms which now are now raised to the power of one, or to the power of zero, results in equation (3.32).

(3.32)
$$\int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m \left(\left(\frac{(1-p)}{100}\right)^n \int_{s_1}^{s_2}\int_{e_1}^{e_2} deds + np \left(\frac{(1-p)}{100}\right)^{n-1} \int_{s_1}^{s_2}\int_{e_1}^{e_2} \frac{1}{(e-s)} deds \right.$$
$$\left. + \sum_{r=0}^{n-2} C_n^r p^{n-r} \left(\frac{(1-p)}{100}\right)^r \int_{s_1}^{s_2}\int_{e_1}^{e_2} \left(\frac{1}{(e-s)}\right)^{n-r} deds \right)$$

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 (3.30) results in equation (3.33), which is the equation to be integrated in the case that the hypotheses span no colour term examples at all.

$$(3.33)\quad \int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m \int_{s_1}^{s_2}\int_{e_1}^{e_2} deds$$

Substituting the value of one for *n* in equation (3.30), results in equation(3.34), which is the equation to be integrated in the case that the hypotheses span only a single colour term example.

$$(3.34)\quad \int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m \left(\frac{(1-p)}{100}\int_{s_1}^{s_2}\int_{e_1}^{e_2} deds + p\int_{s_1}^{s_2}\int_{e_1}^{e_2}\frac{1}{(e-s)}deds\right)$$

We may note that equation (3.33) contains an instantiation of the term corresponding to the first integral in (3.32) and (3.34) 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 from 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 (3.32), and omitting those terms which will not appear when *n* is equal to one or zero. Hence, we can now proceed with the integration of this equation, first rewriting the final integral as in (3.35), so as to make the integration more transparent.

$$(3.35)\quad \int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m \left(\left(\frac{(1-p)}{100}\right)^n \int_{s_1}^{s_2}\int_{e_1}^{e_2} deds + np\left(\frac{(1-p)}{100}\right)^{n-1}\int_{s_1}^{s_2}\int_{e_1}^{e_2}\frac{1}{(e-s)}deds + \sum_{r=0}^{n-2} C_n^r p^{n-r}\left(\frac{(1-p)}{100}\right)^r \int_{s_1}^{s_2}\int_{e_1}^{e_2}(e-s)^{r-n}deds\right)$$

When the three integrations with respect to *e* are performed, the result is as given in (3.36).

$$(3.36)\quad \int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m \left(\left(\frac{(1-p)}{100}\right)^n \int_{s_1}^{s_2} [e]_{e_1}^{e_2}\, ds + np\left(\frac{(1-p)}{100}\right)^{n-1} \int_{s_1}^{s_2} [\ln(e-s)]_{e_1}^{e_2}\, ds \right.$$

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

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*.

3.4.2.2 Identifying the Limits on the Integration

The nature of the limits on the integrations will be considered with reference to an example of generalizing from a specific set of colour examples, to a previously unlabeled colour. Figure 3.6 below shows a representation of the conceptual 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*. We are considering the situation in which we wish to calculate the probability that *x* comes within the denotation of the term *yellow*.

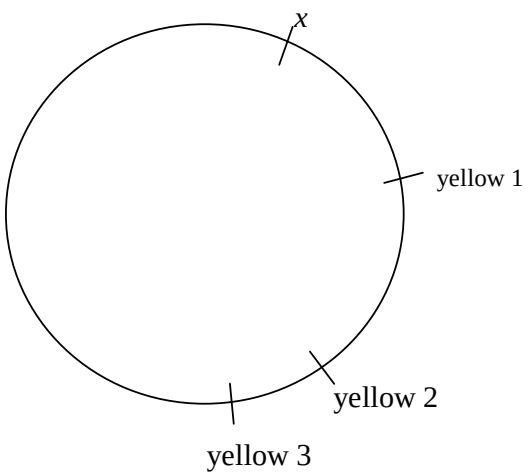


Figure 3.6. The Phenomenological Colour Space with Observed Colour Term Examples.

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 that point 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 3.6 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 which start after *x*, but before *yellow 1*, and which end 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 3.4.2.3 (which starts on page 117). 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 starts of hypotheses from 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 3.6, there are four separate ranges of such hypotheses. These are those hypotheses which both start and end between *x* and *yellow 1*, between *yellow 1* and *yellow 2*, between *yellow 2* and *yellow 3* or between *yellow 3* and *x*. Consider 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 must 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 (3.36) (from page 112) results in equation (3.37).

(3.37)

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

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

However, (3.37) 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 (3.38). 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 3.4.2.4 (which starts on page 121).

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

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

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 3.6 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 (3.36) (on page 112) results in equation (3.39).

(3.39)

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

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 (3.40). The integration over *s* in this case is presented in section 3.4.2.5.

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

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 (3.36), (3.38) and (3.39) with respect to *s*.

3.4.2.3 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 (3.36) (on page 112) are substituted in to give equation (3.41).

(3.41)

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

Rewriting the third integration results in (3.42).

(3.42)

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

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 (3.43). Rewriting the equation in this way assumes that *n* is greater than or equal to three, so it will be necessary to consider another special case of the equation for when *n* is equal to two.

(3.43)

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

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

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

Performing the integrations results in (3.44).

(3.44)

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

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

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

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

When the integrals are expanded, by substituting the values of the limits on the integration for *s*, the resulting equation is given in (3.45). 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.

$$(3.45)\ \int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m \left(\left(\frac{(1-p)}{100}\right)^n (e_2 s_2 - e_1 s_2 - e_2 s_1 + e_1 s_1) + np\left(\frac{(1-p)}{100}\right)^{n-1}\right.$$

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

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

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

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

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 (3.46). This equation applies in all instances where the limits on the 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.

$$(3.46)\ \int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m (e_2 s_2 - e_1 s_2 - e_2 s_1 + e_1 s_1)$$

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 (3.45), and substituting the value one for *n*, results in equation (3.47). This equation applies in all cases where the limits on integration are constants, and the hypotheses contain a single colour example.

$$(3.47)\ \int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m \left(\left(\frac{(1-p)}{100}\right)(e_2 s_2 - e_1 s_2 - e_2 s_1 + e_1 s_1)\right.$$

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

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 (3.45), and set *n* equal to two. This results in equation (3.48), which applies in all cases where the limits on integration are constants, but where the hypotheses contain two colour examples.

$$(3.48)\ \int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m \left(\left(\frac{(1-p)}{100}\right)^2 (e_2 s_2 - e_1 s_2 - e_2 s_1 + e_1 s_1) + 2p\frac{(1-p)}{100}\right.$$

$$\big((e_1 - s_2)\ln(e_1 - s_2) - (e_2 - s_2)\ln(e_2 - s_2) + (e_2 - s_1)\ln(e_2 - s_1) - (e_1 - s_1)\ln(e_1 - s_1)\big)$$

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

Four final equations have now been derived, equations (3.45), (3.46), (3.47) and (3.48). 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 were used in the final implementation of the model.

3.4.2.4 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 (3.38) (from page 116), which results in(3.49).

$$(3.49)\quad \int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^{n} \int_{s_1}^{s_2} s - s_1 ds + np\left(\frac{(1-p)}{100}\right)^{n-1} \int_{s_1}^{s_2} \ln(100) - \ln(s_1 + 100 - s)ds + \sum_{r=0}^{n-2} C_n^r \frac{p^{n-r}}{r-n+1}\left(\frac{(1-p)}{100}\right)^{r} \int_{s_1}^{s_2} \left(\frac{1}{100}\right)^{n-r-1} - \left(\frac{1}{s_1 + 100 - s}\right)^{n-r-1} ds$$

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 (3.50).

$$(3.50)\quad \int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^{n} \int_{s_1}^{s_2} s - s_1 ds + np\left(\frac{(1-p)}{100}\right)^{n-1} \int_{s_1}^{s_2} \ln(100) - \ln(s_1 + 100 - s)ds - \frac{n(n-1)}{2} p^2 \left(\frac{1-p}{100}\right)^{n-2} \int_{s_1}^{s_2} \frac{1}{100} - (s_1 + 100 - s)^{-1} ds + \sum_{r=0}^{n-3} C_n^r \frac{p^{n-r}}{r-n+1}\left(\frac{(1-p)}{100}\right)^{r} \int_{s_1}^{s_2} \left(\frac{1}{100}\right)^{n-r-1} - \left(\frac{1}{s_1 + 100 - s}\right)^{n-r-1} ds$$

When the integrations over *s* are performed the result is equation (3.51).

$$(3.51)\quad \int_{h\in H_i} P(d\mid h)dh = \left(\frac{(1-p)}{100}\right)^n \left[\frac{s^2}{2} - s_1 s\right]_{s_1}^{s_2}$$

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

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

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

Finally, substituting in the limits on *s* produces equation (3.52). 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 where there are only two colour examples, where there is only a single colour example, or where there are no colour examples at all.

$$(3.52)\quad \int_{h\in H_i} P(d\mid h)dh = \left(\frac{(1-p)}{100}\right)^n \left(\frac{s_2^2}{2} - s_1 s_2 + \frac{s_1^2}{2}\right)$$

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

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

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

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

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 is substituted in for *n*, the result is as given in (3.53).

$$(3.53)\quad \int_{h \in H_i} P(d \mid h) dh = \frac{s_2^2}{2} - s_1 s_2 + \frac{s_1^2}{2}$$

However, in the case that equation (3.53) 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 (3.53), using this equation, results in equation (3.54).

$$(3.54)\quad \int_{h \in H_i} P(d \mid h) dh = 5000$$

Equation (3.54) 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[43]. This

[43] Obtaining a probability equal to 5000 may seem an impossibility. However, it should be noted that this is a sum over a range of hypotheses, and not the probability of any one particular hypothesis. Furthermore, at the beginning of section 3.4.2 it was noted that the equation to be integrated contained a constant term, but that this term would cancel out, and hence it has been omitted in later stages of the

equation may seem to be fairly meaningless, given that it applies only in the case that there is not 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[44].

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 (3.52). When these changes are made to the equation, the result is as given in (3.55).

$$(3.55)\quad \int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)\left(\frac{s_2^2}{2} - s_1 s_2 + \frac{s_1^2}{2}\right) + p\left((s_2 - s_1 - 100)\ln(100) + (s_1 + 100 - s_2)\ln(s_1 + 100 - s_2) - s_1 + s_2\right)$$

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 (3.52). Including just these integrals, and setting *n* equal to two, results in equation (3.56).

---

derivations. This term would also modify the obtained value of 5000 if it were reintroduced, but as one value in it, *q*, tends to infinity, we cannot actually put a true correct value on the sum of probabilities over this range of hypotheses.

[44] In such a case, application of the Bayesian inference procedure will reveal that each colour is equally likely to be within the colour word's denotation as outside of it. This result follows from the *a priori* assumption that denotations are equally likely to be of any size, and are equally likely to occur anywhere in the colour space (see section 3.2.5 which starts on page 87).

$$(3.56)\quad \int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^2\left(\frac{s_2^2}{2} - s_1 s_2 + \frac{s_1^2}{2}\right)$$
$$+ 2p\left(\frac{(1-p)}{100}\right)\left((s_2 - s_1 - 100)\ln(100) + (s_1 + 100 - s_2)\ln(s_1 + 100 - s_2) - s_1 + s_2\right)$$
$$- p^2\left(\frac{s_2}{100} + \ln(s_1 - s_2 + 100) - \frac{s_1}{100} - \ln(100)\right)$$

The four final equations derived in this section, (3.52), (3.54), (3.55) and (3.56) 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.

3.4.2.5 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 (3.40) (from page 117 above), substituting in the values for the limits on *e* results in equation (3.57).

$$(3.57)\quad \int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m \int_{s_1}^{s_2} (s_2 - s)ds$$

Integrating over *s* produces equation (3.58).

$$(3.58)\quad P(d \mid H_i) = \left(\frac{(1-p)}{100}\right)^m \left[s_2 s - \frac{s^2}{2}\right]_{s_1}^{s_2}$$

And substituting in the limits on *s* results in equation (3.59). 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*.

$$(3.59)\quad \int_{h\in H_i} P(d \mid h)dh = \left(\frac{(1-p)}{100}\right)^m \left(\frac{s_2^2}{2} - s_1 s_2 + \frac{s_1^2}{2}\right)$$

3.4.3 Applying the Equations

Table 3.3 summarises the conditions under which each of the nine final equations derived in section 3.4.2 are applicable. They 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. (Remember that *x* is the location in the colour space for which we are calculating the probability of whether it can be denoted by the colour term or not.) These values can be substituted into equation (3.20) (from page 105) 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.

Number of colour examples within hypothesis space	**Hypothesis includes *all* the points and *x*, *none* of the points nor *x* or *neither* of these is the case.**	**Equation**
>2	Neither	(3.45)
2	Neither	(3.48)
1	Neither	(3.47)
0	Neither (such a hypothesis will always contain *x*)	(3.60)
>2	All	(3.52)
2	All	(3.56)
1	All	(3.55)
0	All (such a hypothesis will always contain *x*)	(3.54)
0	None	(3.59)

Table 3.3. Summary of the conditions under which each equation applies.

3.5 Computer Implementation

The model was implemented on a PC running the windows operating system, using the C++ programming language. The source code is given in Appendix A, and the same model was used for the purely acquisitional investigations, and the combined

acquisitional and evolutionary investigations presented in Chapter 4. The execution speed when naming colours is cubic on the number of colour examples which a particular agent (artificial person) has remembered for each individual colour, and linear on the number of colour terms which the agent knows. Generally, on a computer with a CPU speed of 1.72GHz, with only a small number of colour examples, the program will be able to name a colour in a fraction of a second, but with greater numbers (for example one hundred examples), it may take several seconds.

3.6 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 (that is *red*, *orange*, *yellow*, *green*, *blue* and *purple*). This experiment was conducted primarily to determine whether the resultant denotations would have prototype properties. Five examples of each of these colour terms were given to the model, and from these it learned the denotations shown in Figure 3.7. This figure gives a graphical representation of the fuzzy set denotation derived for each colour using the Bayesian model.

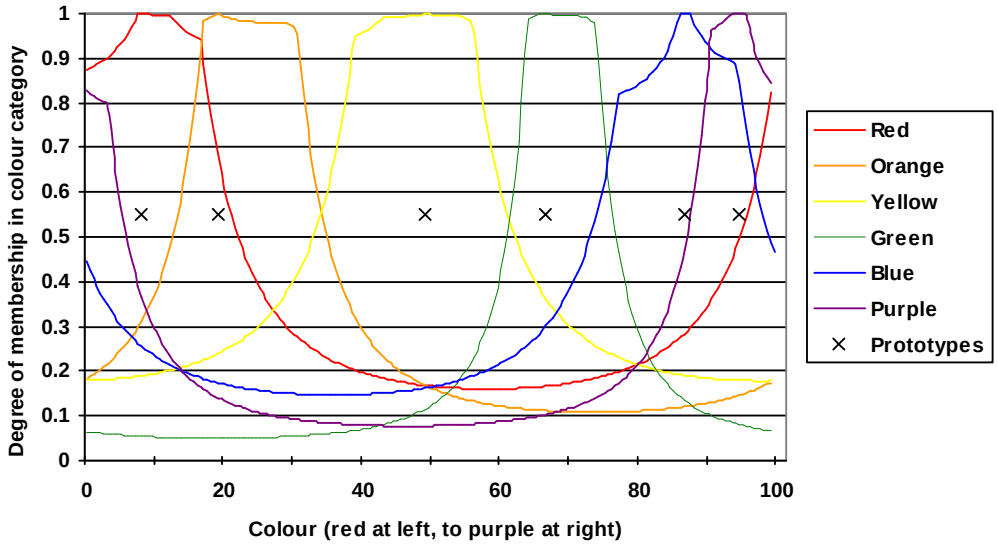


Figure 3.7. The Fuzzy Denotation of English Basic Colour Terms after 5 Examples, with *p*=0.8.

The examples were created by randomly selecting points within ranges of the colour space corresponding to each colour term. The size of the denotation of each colour term was based on the data for English colour terms given in Berlin and Kay (1969). Berlin and Kay showed the extent of each colour term on an array of Munsell colour chips. As the Munsell system is designed to space the chips in phenomenologically even gradations of colour, the number of colour chips within the range of each colour term should be approximately proportional to the size of the part of the conceptual colour space that can be denoted by that colour term. (Munsell chips, and Berlin and Kay's use of them, are discussed above in section 2.1.) The denotations of each colour term were then mapped onto the colour space of the model, so that they filled the whole of the colour space, and yet the size of each was proportional to its extent on Berlin and Kay's array of Munsell chips. The denotation of *red* began at hue 0, followed by the denotations of *orange*, *yellow*, *green* and *blue* in that order, and finally ending with *purple*, the denotation of which finishes at hue 100, which, due to the circular nature of the colour space, is also hue 0. The actual ranges of each

denotation are given below in Table 3.4. In order to simulate inaccuracy in the data, and to allow for the fact that colour term's extensions are not in reality so precisely and consistently delineated, each example was then randomly moved within the range of plus or minus five units[45]. The accuracy parameter, *p* (introduced in section 3.3.1 above), was set to 0.8, so that the model would have an expectation that 80% of examples would be accurate, and that 20% would be positioned at random.

Colour Term	**Range of denotation**	
	Start	End
red	0	13
orange	13	30
yellow	30	56
green	56	80
blue	80	91
purple	91	0

Table 3.4. The Ranges of the Denotations for the English Colour Terms Taught to the Model.

The horizontal axis in Figure 3.7 corresponds to the range of colours in the conceptual colour space, from red to orange, yellow, green, blue, purple and finally back to red. (As the colour space is circular, the left and right edges of the graph represent

[45] We should note that this has introduced a form of noise into the model, but that this noise is in a form that is very different to what the model expects. The model expects noise to be completely random, and below in Chapter 7 an investigation is made of the effect of adding that kind of noise to the model. However, arguably the kind of noise introduced here is more realistic, as it simulates the situation in which the speakers of the language who are the source of the data have all learned slightly different denotations for the colour terms, which is what is normally observed when the colour term denotations known by several speakers of a language are investigated empirically. Probably in reality both kinds of noise are present in input data, as if a hearer mistook the referent of a colour term, we would expect the colour associated with that term to be completely random, as it would be the colour of some object other than the one that the speaker had intended to label using the colour term.

adjacent points in the colour space.) The vertical axis represents the probability with which the learner believes that each colour comes within the denotation of each colour term. Therefore values on this axis range from zero at the bottom of the graph, indicating that a colour is definitely not within the range of a colour term, to one at the top, indicating that it definitely is. These values may alternatively be interpreted simply as the degree of membership of each colour within the categories corresponding to the colour terms. For every colour term, the probability that it could be used to denote each colour was calculated, and so there are no sudden jumps in the degree of membership of neighbouring colours. Hence, when these values are plotted on the graph, they form continuous curves.

From Figure 3.7, we can observe some important properties of the learned denotations. Firstly, we can see that each colour category has a single prototype. In Figure 3.7, the prototypes locations are all marked by an ×. Consider, for example, the curve corresponding to the colour term *yellow*. This curve rises to a single peak near the middle of the graph, where the learner is very sure that that colour can be referred to by the term *yellow*, and this would appear to correspond to the colour which is the prototype of this person's *yellow* category. However, on either side of the prototype, the curve falls away, showing that the probability that a colour is a member of the category *yellow* decreases the further the colour is away from the prototype. There is a range of colours which the person is almost certain come within the denotation of *yellow* (where the *yellow* curve is very high), and there is a section of the colour space which the person thinks is unlikely to come within the denotation of *yellow* (where the curve drops low down).

In between the ranges of colour about which the person is fairly confident of whether they can be denoted by *yellow* or not, are colours which may be considered marginal examples of the colour term, especially where the curve is near the 0.5 level. These points correspond to colours which the person considers are equally likely to come within the denotation of the colour term as to be outside of its scope. Hence we can see that the denotations have another key property of prototype categories besides prototypes, and that is that they have fuzzy boundaries. Moving between these fuzzy edges of the category and the prototype, colours become better exemplars of the category as they get more similar to the prototype.

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 3.7, we can see that there are colours which are considered more likely than not to be within the denotation 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* a learner considers the denotation of all colour terms to be equally likely to be of any possible size, and hence the average size for all possible hypotheses is equal to half the colour space. As the person has observed only five examples of each colour term, the model has been influenced to a very large extent by the *a priori* assumptions made in assigning probabilities to hypotheses.

Next, it was investigated how a person's representation of colour term denotations would change once they had observed more examples of a word's use. Fifteen more examples of each colour term were added to the model, and the resulting denotations are shown in Figure 3.8. The main difference between this graph and Figure 3.7 is that

the model is now much more certain about which colours 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, where they are very flat, because in these areas the model is almost completely certain that the colours come within the denotation of the corresponding colour term. (While on the graph it may look as though the curves are completely flat at these points, and that they have reached all the way to the top of the graph, actually they are just very close to the top. If reference is made to the points from which they were plotted, it can be seen that each curve still rises to a single peak, and the degree of membership decreases very slowly on each side of this point.)

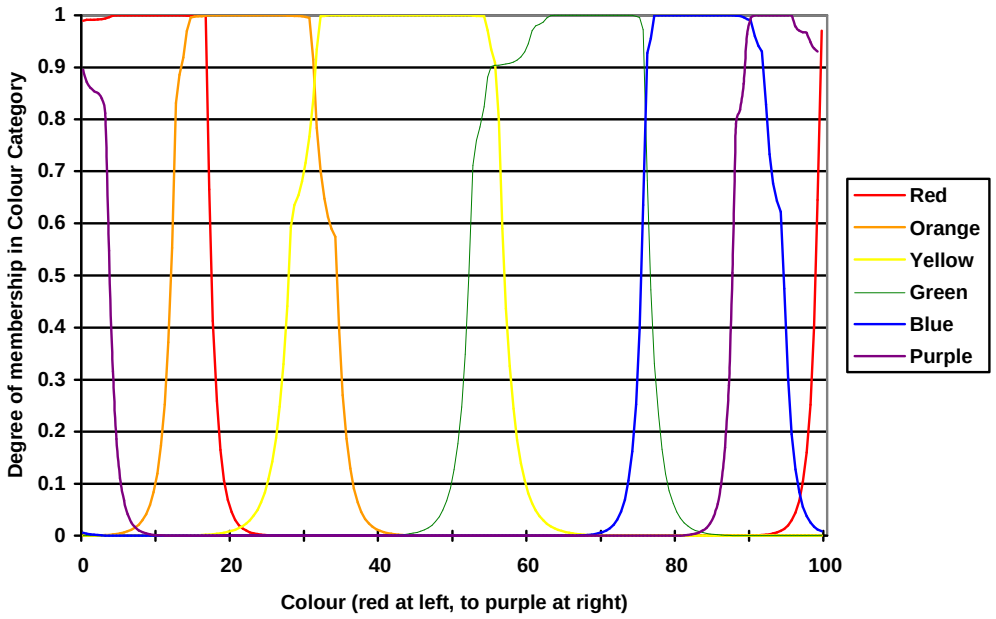


Figure 3.8. The Fuzzy Denotations of English Colour Terms after 20 Examples, with *p*=0.8.

We can also see that the model is almost completely certain that some colours cannot be labelled with particular colour terms. This is the case in areas of the graph where the curves are very close to zero, running along the bottom of the graph. As the learner now had more data available from which to learn, they were able to determine the correct denotation of the colour terms with a greater degree of accuracy, and so

whether a colour is a member of a colour category is only really in doubt for a small range of colours. However each category still retains the overall prototype structure: colours have varying degrees of membership, and each category has a single best example, and some marginal examples.

Further experiments have determined that colour terms can equally well be learned by the model, no matter where in the colour space their denotations are, and that the model is able to learn colour terms covering much larger parts of the colour space than the English terms do. (Languages with fewer colour terms would have such colour terms, as in such languages, each term must denote a larger range of colours.) However, the general findings concerning prototype effects are also applicable to such colour terms. Therefore, it seems that the model is able to learn the basic colour term system of any language, at least in so far as the colour term system can be represented in the one-dimensional colour space.

3.7 Learning with Unreliable Data

Given the success of learning the English colour terms from reliable evidence, it was decided to investigate to what extent the model would be able to learn if it was presented with data which contained a large proportion of erroneous examples. As noted in section 3.2.3, it is important that the model is able to learn when presented with inaccurate data, as it is likely that this is similar to the situation in which children learn the meanings of words.

The examples given to the model were generated in the same way as in section 3.6, but this time only the denotation of *green* was considered[46]. A comparison was made between the denotation learned when the model was presented with only accurate examples, and when it was given an equal number of accurate and random examples. In all cases the accuracy parameter was set to 0.5, so that a learner would have the same expectation of encountering random examples, as of encountering accurate ones.

The learned denotations, together with the exact position of each example, are shown in Figure 3.9. We can see that when the model was presented with just five accurate examples, a denotation is learned which assigns very high degrees of membership to colours which are within the correct denotation of *green*, and low degrees of membership to other colours.

[46] The choice of *green* was completely arbitrary, and these results should be equally applicable to other colour terms.

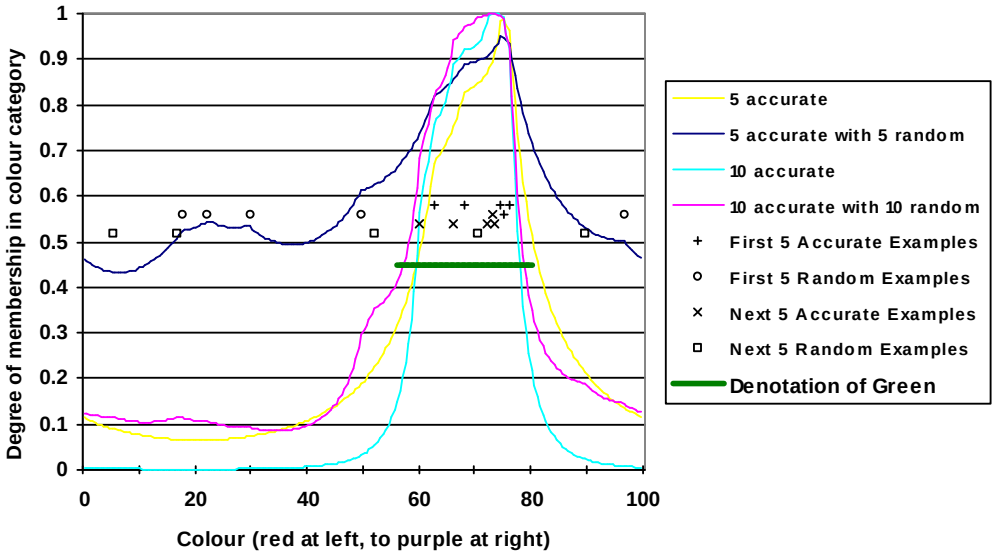


Figure 3.9. The Fuzzy Denotations Learned for *Green*, with *p*=0.5.

When five random examples are present, as well as the five accurate ones, we can see that the model is unsure of the limits of the denotation. While the prototype remains in approximately the right place, the degree of membership of colours declines more gradually moving away from this point. Never does the degree of membership get below 0.4, so in this situation a learner would not be sure that any colour could not be named by the colour term. This problem has arisen because the model has no way of distinguishing between which examples are accurate and which are not, and so it will not be certain which examples really come within the denotation.

However, we can see that the situation changes dramatically if we add a further five accurate examples, even if they are accompanied by a further five random ones. Now the model's performance at determining the denotation correctly is almost as good as if it had only observed the accurate examples. With this greater number of examples, the cluster within the correct denotation of green becomes clearer, and the model is able to decide with quite a high degree of confidence that examples outside of this

range are erroneous. Hence, we can see that just a small increase in the number of examples allowed the model to learn accurately, despite the fact that half of those examples were erroneous.

3.8 Robustness of the Model to Random Noise

Having seen that the system is able to learn when presented with data containing a large proportion of misleading examples, it was decided to investigate more rigorously to what extent the learning process was robust in the face of large quantities of random noise. This was done by adding varying levels of random noise to the input data, and measuring how accurately the model determined the target category in each case.

A target category of size 30 was created for the model to learn (the whole colour space being 100 units wide)[47]. Examples of this category were generated in the same way as for the English colour categories, but varying amounts of completely random examples were added to the training data to simulate noise[48]. The parameter, *p*, was

---

[47] The decision to make the category 30 units wide was arbitrary. However, 30 units probably corresponds roughly to the size of many real colour categories.

[48] It should be noted that *noise* is used differently here to the way in which it is typically used within the machine learning community. I count any example which is random as noise, but some of these will in fact be accurate simply by chance. In contrast, in the machine learning literature, typically only examples which are inaccurate are counted as noise, so in this case that would be examples falling outside of the target category. (As machine learning research is usually performed on empirical data, it would often be impossible to separate examples which are accurate by chance from those which accurate for any other reason.)

adjusted, so that it always accurately reflected the proportion of examples which was accurate, so that the model had advance knowledge of how reliable the data was, though it did not know which particular examples were accurate and which were not.

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

The results of these experiments are given in Figure 3.10, which shows the proportion of colours classified accurately, and left unclassified, when the level of noise in the data varied from 20% to 80%, and the number of accurate training examples observed varied from 5 to 30. The program was run twenty times in each condition, and the values plotted in the graph are averages over all the runs in each condition. These examples would in each case be accompanied by the number of random examples needed to simulate the appropriate level of noise (see Table 3.5). The results were derived from sampling at 100 evenly spaced points in the colour space, and in each case investigating whether that colour was classified by the model as coming within the colour category, outside of the colour category, or whether the model did not classify that colour at all. If we viewed these experiments from the perspective of the standard machine learning test data–training data paradigm, then these 100 points would correspond to a stratified sample of 100 test data items, each of which has to be

classified. The results from which Figure 5 was plotted are shown together with their standard deviations in Table 3.5.

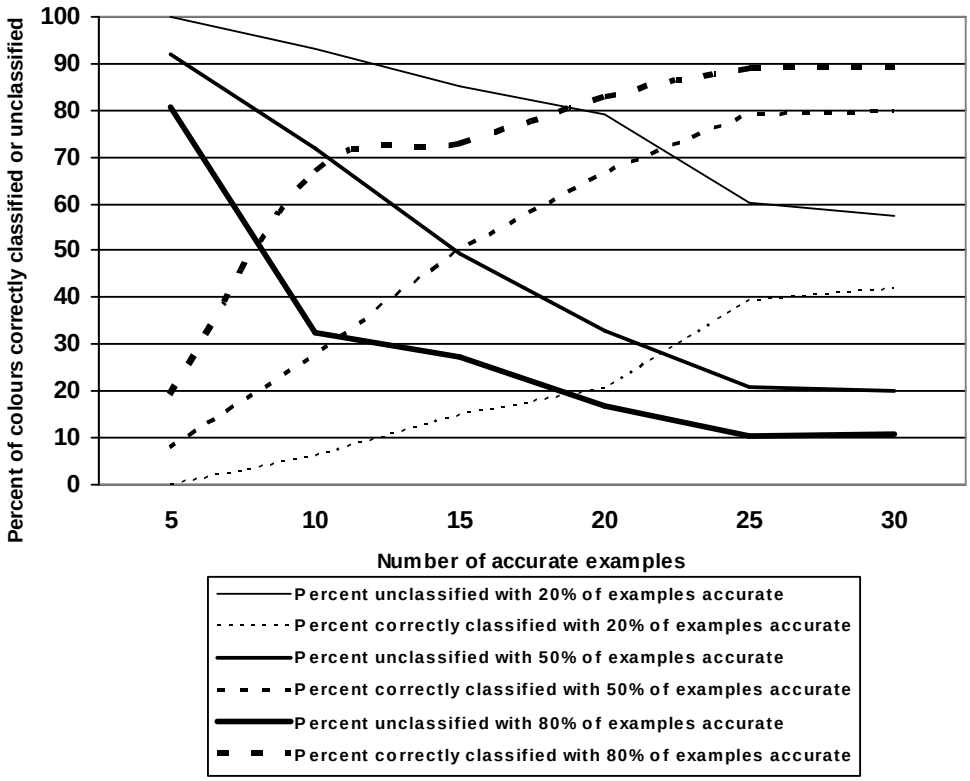


Figure 3.10 Accuracy of Learning with Noisy Data

Percent of examples which were accurate		Number of Accurate Examples	Number of Random Examples	Percent incorrectly included		Percent incorrectly excluded		Percent unclassified		Percent correctly classified	
				Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
	20	5	20	0	0	0	0	99.9	0.654	0.150	0.654
		10	40	0.200	0.872	0.600	2.62	93.3	15.3	5.90	12.8
		15	60	0	0	0	0	85.0	15.0	15.0	15.0
		20	80	0.450	1.43	0	0	79.0	16.8	20.6	16.4
		25	100	0.300	1.31	0	0	60.3	25.1	39.5	25.1
		30	120	0.450	1.43	0.200	0.872	57.6	23.2	41.8	23.0
	50	5	5	0	0	0	0	92.1	8.37	7.90	8.37
		10	10	0.450	1.36	0	0	71.9	17.1	27.7	16.9
		15	15	0.200	0.678	0	0	49.6	21.1	50.3	21.1
		20	20	0	0	0.300	1.31	33.0	18.3	66.7	18.0
		25	25	0	0	0.0500	0.218	20.9	14.4	79.1	14.4
		30	30	0	0	0.0500	0.218	20.0	6.63	80.0	6.60
	80	5	1	0.100	0.436	0	0	80.8	9.76	19.2	9.73
		10	2	0.0500	0.218	0.500	1.36	32.6	14.3	66.9	13.7
		15	4	0.200	0.678	0	0	27.2	8.33	72.6	8.49
		20	5	0.100	0.300	0	0	17.0	4.68	82.9	4.67
		25	6	0.0500	0.218	0.500	1.07	10.6	3.75	88.9	3.51
		30	7	0	0	0	0	10.7	3.02	89.4	3.02

Table 3.5. Means and Standard Deviations Showing Precision and Accuracy of Learning with Noisy Data. (Accuracies are to 3s.f., and results are averages over 20 runs.)

We can see that, as more training examples are observed, a higher proportion of colours is correctly classified. Also, if a higher proportion of training examples are accurate, this leads to more accurate classification. However, even when 80% of the data are random, once 30 accurate training examples have been observed (by which time 120 random training examples would also have been seen), over 40% of test colours are classified accurately. When 50% percent of the data was accurate, the model achieved very good performance with 30 accurate training examples, correctly classifying over 80% of the test colours.

Of course, classifying a high proportion of examples accurately would not be an impressive result if a large proportion of examples were also classified inaccurately. However, we can see from Table 3.5 that there was a very low level of error in

making classifications. The highest proportion of test colours which were ever classified inaccurately, in any condition, was 0.8%. (This was when 80% of examples were random, and only 10 accurate examples had been observed, and 0.8% is an average, based on all 20 runs of the system in this condition.) In most cases, there were even fewer colours classified inaccurately, and in many cases none at all. So we can see that learning can proceed with a very high degree of precision even in the presence of large quantities of noise. When there are very high levels of noise in the training data, or when there is only a small number of training examples available, a large proportion of test colours are left uncategorized. However, even in such conditions, very few examples are classified incorrectly.

It is often reported that children are able to learn accurately in the face of very noisy data (Siskind, 1997), though whether they can really learn with the very high levels of noise used in these experiments is an interesting question. Resolving this issue would involve empirical investigations concerning exactly what input children receive, or what they are able to learn from input which was intentionally made noisy. Both of these possibilities go beyond the scope of this thesis, so they will be left as possible future research goals.

Chapter 4

Comparing Acquisitional and Evolutionary Simulations

So far, all of the experiments conducted with the Bayesian model have investigated whether it can learn real colour term systems of the types which have been observed in the world's languages. However, some authors have suggested that an acquisitional linguistic theory should do more than simply account for how attested languages can be learned. In section 2.2 it was noted that Chomsky (1972, 1984, 1986, 1995) has argued that an acquisitional theory should be largely responsible for explaining, not only how attested languages are learned, but also why we do not see languages of unattested types. This section investigates whether the acquisitional model correctly predicts the existence of only the attested types of colour term systems, or whether the range of attested languages is better explained when an evolutionary dimension is incorporated into the theory, as was proposed by Hurford (1987, 1990).

4.1 Learnable Colour Term Systems

In order to investigate whether the acquisitional model was able to explain the properties of colour term systems, it was investigated whether it could learn colour term systems which had properties inconsistent with the colour term systems of real

languages. If the model were not able to learn such systems, then this would provide an explanation as to why such systems are not seen in real languages, but if it were able to learn them, then it would have failed to have explained why those types of colour term system do not exist[49]. The model was tested with respect to two key properties of colour term systems: prototype properties and partition.

Figure 4.1 shows one colour term system which was learned by the Bayesian model, but which is not of a type reported in empirical studies. The colour terms were learned by first deciding on a section of the colour space which was to correspond to the denotation of each word, and then selecting random example colours from within these parts of the colour space, and presenting them to the model[50]. We can see that the colour term system shown in Figure 4.1 contains two overlapping colour terms, which both denote hues in the red and yellow part of the colour space. The model has observed ten examples of each of these colour terms, and using these it has been able to determine roughly which colours come within each term's denotation. We can see that both of these terms display the prototype properties which are characteristic of basic colour terms. Each term rises to a single point which is the best example of the term, but the degree of membership in the colour term category declines gradually

---

[49] We should note that a theory of colour term acquisition also needs to be able to account for the acquisition of attested colour term systems, but it was shown in Chapter 3 above that it is able to do this, so there is no need to repeat the demonstration here.

[50] In all the examples in this section and in section 4.2, the accuracy parameter, *p*, was always set to 0.5.

moving away from that point. (It may be difficult to determine the exact location of the prototypes from the graph alone. This is because the graph is only drawn with limited accuracy, and so the curves may appear flat on top. However, reference to the values from which the graph was plotted, reveals the exact locations of the prototypes, which have been marked on the graph.)

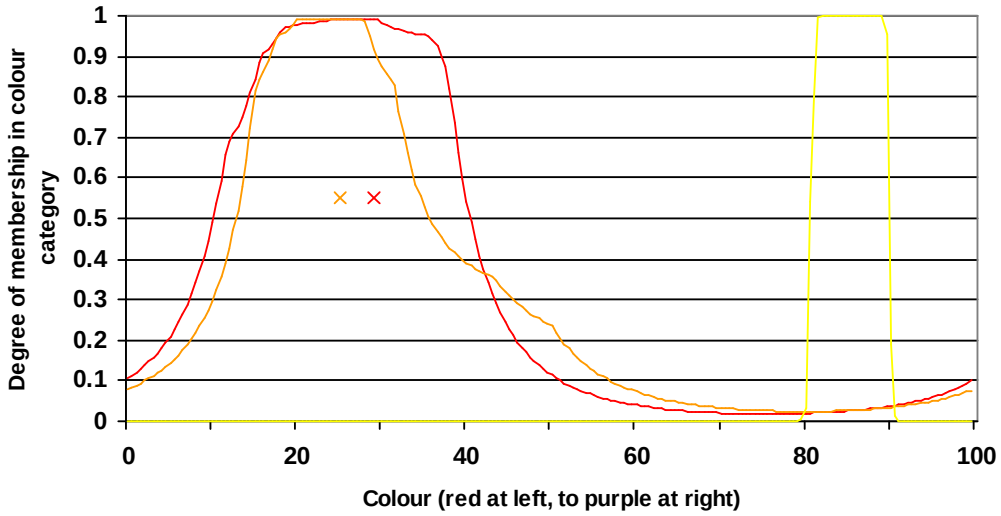


Figure 4.1. A Learnable Colour Term System of a Type which is Unattested Typologically.

(×'s mark prototype locations and *p*=0.5.)

In contrast, the colour term on the right hand side of the graph has quite different characteristics. This colour word was learned from examples generated in the same way as for the other two colour words, but in this case the model has observed 60 examples of colours named by the colour term. There is a range of hues for which the model has a very high degree of certainty that they are members of the colour category, and in this part of the graph the curve is almost completely flat, and right at the top of the graph. However, for almost all other colours, the model is very certain that they cannot be named by the colour term, which is indicated by the curve being very close to the bottom of the graph. There are only a very few colours for which the

degree of membership is at an intermediate value, and so the boundaries of the colour category are demarcated by almost vertical lines.

This colour term does not have prototype properties, as it does not have fuzzy boundaries, and the degree of membership of colours in the category is almost completely constant throughout its denotation. (The degree of membership does in fact rise to a single maximum close to the centre of the colour category, but this cannot be seen on the graph, because both the colour with the greatest degree of membership, and those immediately surrounding it, have almost identical degrees of membership). This colour term clearly does not resemble the colour terms seen in real human languages, which shows that the Bayesian model is able to learn languages with properties which are unattested typologically.

If we look at the colour term system as a whole, we can identify another property of this system which is not in accord with the colour term systems which have been observed in real languages, and that is that the colour terms do not partition the colour space. Rather than having a single word which can be used to name each range of colour, we have two overlapping colour terms, with their prototypes in almost the same part of the colour space, something which is not usually observed in real languages[51]. There is a further inconsistency between this colour term system and those observed in real languages, and that is that there are large gaps between the

[51] As mentioned above, MacLaury (1997a) identified the coextension phenomenon that is seen in some languages, in which two overlapping colour words denote roughly the same range of colours. However, in such cases, each colour term tends to have its prototype in a different part of the colour space, so this phenomenon does not correspond to the case of the overlapping colour terms seen here.

overlapping colour terms and the other term, so that many colours are left without any corresponding linguistic label. In contrast, empirical evidence shows that colour terms almost always partition the colour space, so that for every colour there is a corresponding colour word which may be used to name it[52]. If Figure 4.1 corresponded to a real language, then we would expect to observe a series of colour terms, with little or no gap between them, and only minimal overlapping of neighbouring terms.

4.2 Evolvable Colour Term Systems

The results of section 4.1 clearly show that the acquisitional model alone is insufficient to explain the empirical data concerning colour term universals, and so the program was extended so that it could model not only learning, but also the social processes in which language is used, and through which it is passed on to each new generation of speakers. Rather than just using a single model of acquisition, and presenting it with random examples, multiple copies of the model were created in order to simulate a whole community of people. In all the simulations reported in this

[52] As noted in 2.1, Kay and Maffi (1999) and Levinson (2001) do report the existence of a very small minority of languages in which the colour terms do not appear to partition the colour space. However, such colour term systems are exceptional, so it would seem that we are more in need of an explanation of why almost all languages partition the colour space, rather than an explanation of why a minority do not. Once an explanation of why partition occurs has been developed, we may then be able to explain non-partition as a chance occurrence, especially if the explanation relies on rules which only make statistical rather than absolute predictions.

chapter, ten artificial people were used[53]. These artificial people were then made to talk to each other, and to learn from one-another. This process was simulated over a number of generations, until eventually the simulation was stopped and the properties of the emergent language were examined.

In the initial state of the model, each person had observed a single random colour anywhere in the colour space, together with a colour word which had been used to name it. Initially the colour words known by each person were all different, so that there would be no coherent language in the community. Each person was assigned a random age, varying from zero to the maximum age to which people in the simulation could live. The procedure through which the simulation proceeds is outlined in Figure 4.2. Firstly a speaker and a hearer were chosen at random (the only restriction being that these could not both be the same person). A colour for the speaker to name would then be chosen, also at random, and the speaker would find the word which they thought most likely to be a correct label for the colour[54]. This word, together with the corresponding colour, would then be observed by the hearer, and remembered by

---

[53] Varying the number of people used in the simulations does not appear to have a significant effect on the results, but as the number of people is increased the program tends to run more slowly. Clearly a real language community would contain more than ten people, but increasing the number of people simulated would not seem to be necessary for present purposes.

[54] At first each person only knows one word, so they will have to say that word, but as soon as they have remembered examples of more than one word, they will then choose a word according to the examples they have observed, and the colour which is being named.

him[55] as an example. He would then use this example to help determine the best word to choose when it came to be his turn to be the speaker. This procedure was then repeated many times, to simulate people talking to each other and using colour terms. However, one time in every thousand, instead of the speaker choosing the best word based on the observations they had made, they would be creative instead, and make up a completely new word. This occasional creative behaviour is necessary, because otherwise there would be no way for new words to enter the language, or for the overall number of terms known within the community to increase.

[55] For convenience I refer to all artificial people in the simulation as though they were male, although, as no distinction is made in the model concerning the sex of the artificial people, this decision is completely arbitrary.

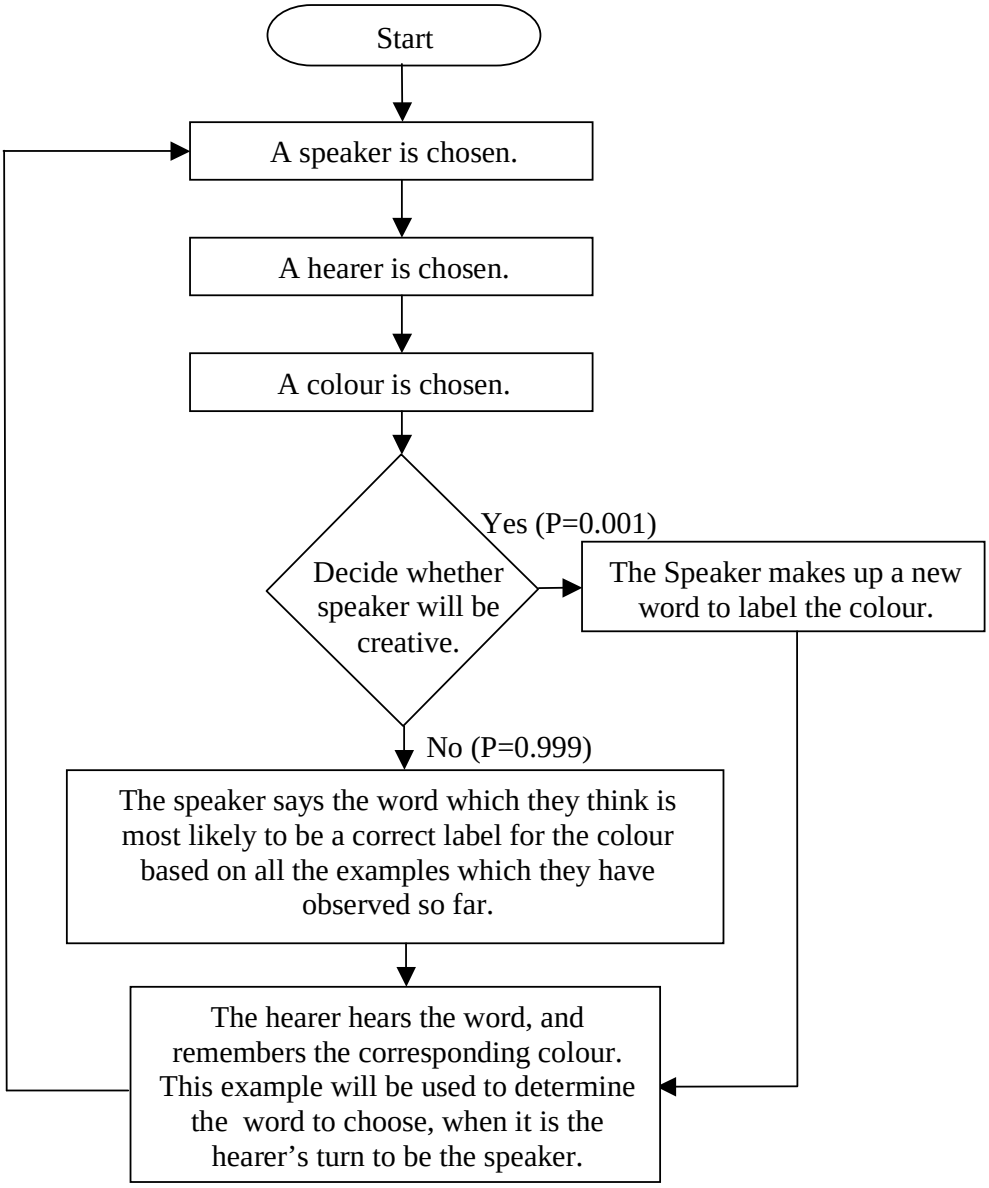


Figure 4.2. Outline of the Evolutionary Algorithm.

(*P* stands for the probability of making each choice.)

A parameter in the model controlled how long each person lived for, measured in terms of how many times a person would speak during their lifetime. The actual life span of each person was varied randomly, by an amount of up to 20% either above or below the chosen average life span. Once a person reached the end of their life span, they would be replaced by a new person with an age of zero, who had not observed

any colour term examples. (If a person should ever be chosen as the speaker before they had observed any colour terms, then the program would just go back and choose another person instead.)

Figure 4.3 shows the result of one experiment, where the simulation was run for a period of time equal to ten average life spans, and where, on average, each person heard 60 examples during their lifetime. This graph shows the colour words learned by one person in this simulation who was near the end of his life span. It shows that a language has evolved which has six colour terms, each of which is focused in a different part of the colour space, and each of which has prototype properties. Most of these terms were those known by the people at the start of the simulations, but the parts of the colour space in which these colour terms have their denotations does not appear to be related to where the initial examples of these terms were located. At the beginning of the simulation, each person only knows one colour term, so they will use that term to name the whole of the colour space. Hence, we would expect initial naming behaviour to be fairly random, until a coherent colour term system becomes established. However, even after this point, semantic drift occurs, so that the denotations of each colour term tend to move in the colour space, and after a number of generations, the range of colours denoted by each term may change completely. Two of the colour terms in this simulation were ones that had been created after the simulation had begun (by speakers making up new colour terms), and which had then become established in the language. These terms may well have displaced established colour terms, which would then have been lost from the language.

The terms roughly partition the colour space, dividing it up so that there are only small overlaps or gaps between the colour terms. All the other people in the simulation who had observed more than 30 examples, had learned similar colour term systems, each containing the same six terms. (Although the location of the category prototypes and boundaries varied somewhat between people, as each would have observed a unique set of examples.) The colour term systems of the youngest members of the community were more variable, as these people would not have observed enough data to determine accurately the correct denotation for all the colour terms, and may not even have observed any examples at all of some terms.

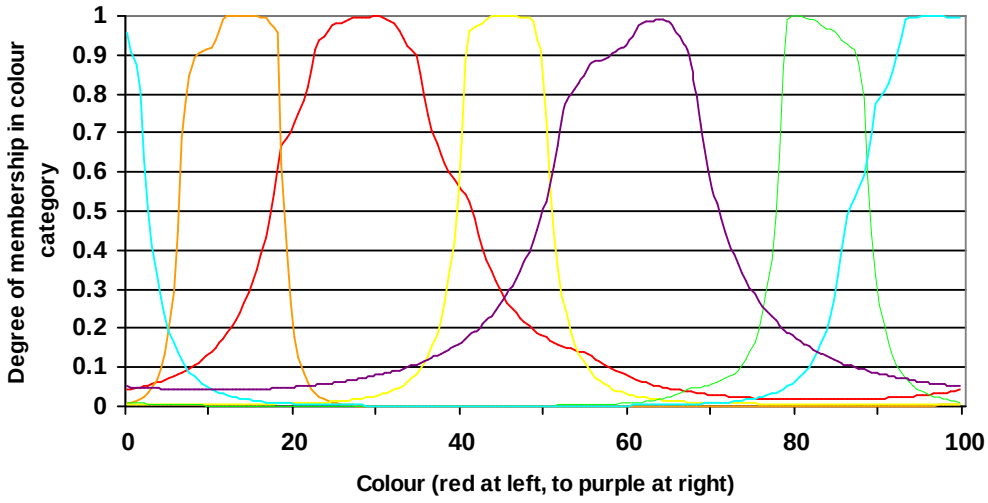


Figure 4.3 A Colour Term System which Emerged in an Evolutionary Simulation. (*p*=0.5)

The simulation has produced a colour term system which appears to have the general properties of colour term systems found in real languages, in that it partitions the colour space, and each term clearly has prototype properties[56]. Repeating the

[56] Clearly the simulations do not explain all the reported properties of colour term systems, most obviously because they make no attempt to explain the typological data of the type reported by Berlin

simulation produced similar results, although there was some variation as to the exact number of colour terms which emerged. We might expect that if people observed more colour term examples during their lifetimes, then towards the end of their lives they would learn the denotations of the words with a very high degree of confidence and precision. This would cause words to lose their prototype properties and become like the rightmost term in Figure 4.1.

However, this does not happen in practice during the evolutionary simulations. When the average number of examples which a person observes during their lifetime is increased, the emerging colour term systems tend to have more words, and so the number of examples of each word observed by each speaker remains more or less constant. Conversely, decreasing the number of examples observed by each speaker tends to produce colour term systems with fewer colour terms, but again people will observe a similar number of examples of each term. The relationship between the number of colour examples people observe during their lifetime, and the number of colour words in their languages, is investigated more thoroughly in section 6.3, but here we should just note that the mechanism through which language is transmitted between generations appears to control the number of colour terms, and hence ensures that all colour terms will always have prototype properties.

---

and Kay (1969). Before learning begins, the colour space in the present model is completely uniform, and so gives no special status to any particular colour, and hence there is no possibility of the model explaining why colour terms tend to be focused on particular locations in the colour space. It will be shown below in Chapter 6 that, when unique hue points are added to the model, as was proposed in section 3.2.4, the model is then able to account for much of the typological data.

4.3 Implications of the Results

The results of the simulations of sections 3.6, 4.1 and 4.2 clearly show that both partition and non-partition colour term systems are learnable by the acquisitional model, as are colour terms with prototype properties and those without. As empirical observations have found that languages do generally partition the colour space, and that basic colour terms do have prototype properties, it would appear that the acquisitional model fails to sufficiently constrain the range of learnable languages.

That, at least, is the view consistent with Chomsky's (Chomsky 1986) focus on language acquisition, and on speakers' individual knowledge of language, as the primary objects of study in linguistics[57]. This is because, if we took as our primary data colour term systems of the type which emerged in the evolutionary simulations, we would reach the conclusion that people are equipped with an innate Language Acquisition Device which forces the learned colour term systems to both partition the colour space, and to have prototype properties, as all the systems which emerged in these simulations had those properties. However, in the case of the simulations reported here, we can see that any such conclusion would be completely incorrect, as there is nothing in the acquisitional model which gives any preference to learning colour term systems which conform to the property of partition, and the model is quite capable of learning colour term denotations which do not have prototype properties. The simulations therefore demonstrate that I-language (see section 2.2), is too narrow a concept to allow us to understand the observed properties of colour term systems.

---

[57] Chomsky's approach to linguistics, and the concepts of language which he defines, were introduced in section 2.2.

The extensions which Hurford (1987, 1990) makes to Chomsky's model of language acquisition are uncontroversial, in that it is clear that we learn language from other people, and so the language which provides the input to our Language Acquisition Devices will be determined by other individual's I-languages, and the social context in which language is used. What is controversial about Hurford's model is whether it is necessary to consider the diachronic perspective when understanding central aspects of synchronic language[58]. In my evolutionary simulations of colour term systems (section 4.2), we saw new properties emerging which were not predictable from the properties of the acquisitional model. This demonstrated that, at least in this situation, the social processes in which language is used are as important as individual psychology for understanding the properties of colour term systems. This is clearly supportive of de Saussure's (de Saussure, 1959/1916) view that language is simultaneously a social and a psychological phenomenon. It seems that we can only understand the synchronic properties of language through considering the diachronic processes of language evolution, although the nature of diachronic change is determined by synchronic processes.

This discussion about the need for the psychological model to be placed within a social simulation in order to explain the data adequately relates to the particular acquisitional model developed in this thesis. However, there are almost certainly significant differences between how the acquisitional model learns colour terms, and how people learn them. This raises the question of how general this result is, or whether with a significantly different model, which potentially could more closely

---

[58] *Synchronic* and *diachronic* were defined in section 2.2.

reflect the actual psychological processed being modelled, both partition and prototypes would emerge, even when no social aspects of language were modelled. Such a model would have to explain these properties as arising from the psychological process through which the input data is mapped to create the learners knowledge of the language. This would involve showing that when a model was presented with input data, it always produced a grammar with the partition and prototype properties. It is common in linguistics to assume that linguistic universals must exist because languages with alternative properties simply are not learnable (for example Nowak, Komarova and Niyogi, 2002). However, as discussed in sections 2.1 and 4.1, there appear to exist a small number of languages for which partition does not apply, so for any purely acquisitional model to provide an adequate account of partition, it would have to achieve partition with almost all, but not quite all input data.

However, non-partition appears to be a relatively stable phenomenon in some languages, and this is presumably because the learners of those languages receive input reflecting this from other speakers of the language. Therefore it would seem that it is unproblematic for input data received by a learner to map to an adult grammar with a non-partition colour term system, and so whether a learner learns a partition or a non-partition language depends on the language spoken by other speakers in their community. Such an explanation therefore necessarily requires a model with a social dimension, as it makes reference to input received from other speakers. In contrast, as far as I am aware, all colour terms in all languages have prototype properties. So it could be that the only type of representations for colour terms that humans are capable of learning is a prototype one. If this is the case, the correct explanation of the prototype properties of colour terms would be purely in terms of a psychological

process, either concerning colour term acquisition, or representation, or both. Therefore it would seem that this chapter has provided stronger evidence that a social explanation of language is needed with regard to the property of partition, as compared to the phenomenon of prototype properties.

I believe that this model also exemplifies the value of the computational evolutionary modelling methodology[59] in helping us to gain a better understanding of language. Surprising new properties emerged in the evolutionary simulations – properties which it would have been difficult to predict simply by extrapolating from the properties of the acquisitional model. This raises some interesting questions regarding other computational models of language acquisition. For example, Ellefson and Christiansen (2000) constructed a recurrent neural network which they used to model the acquisition of syntactic rules concerning question formation. They found that the neural network could learn simple artificial languages, in which the question formation rules were of the type found in real languages, better than it could learn artificial languages in which the question formation rules violated universal constraints on the syntax of question formation. They suggested that this learning bias has caused languages to evolve in such a way that they all conform to what now appears to be a universal rule. However, in the light of the findings of this chapter, it would be interesting to investigate whether Ellefson and Christiansen's model would in fact produce languages with the predicted properties if a community of speakers

---

[59] By *computational evolutionary modelling*, I simply mean any approach that simulates a process of evolution on a computer, including all of the expression-induction models discussed above in section 2.4.

was modelled over several generations, and whether any other unexpected properties would emerge. At present, most acquisitional models take too much time to learn to make such simulations practicable, but as computers become more powerful there will be increasing opportunities to make use of this kind of evolutionary methodology.

In conclusion, this chapter has presented evidence which suggests that colour term systems partition the colour space as a result of diachronic processes. The simulations therefore do not provide any evidence to support the hypothesis that a component of the Language Acquisition Device, or any aspect of the ontogenetic process, prevents us from learning basic colour terms which either overlap, or which leave large ranges of colour without any corresponding colour term[60]. This is because, partition in the simulations clearly does not emerge as a result of the non-learnability of non-partition languages, and so, in the languages emerging in the simulations, partition in the model is clearly not due to learnability constraints imposed by the Language Acquisition Device. It was also proposed that the mechanism that we use to learn colour terms may be equally able to learn colour terms which have prototype properties, as ones which do not. Colour terms in real languages may have prototype properties only because of the social processes which moderate the number of colour terms which emerge in a language. In general, the synchronic properties of a language may best be understood by placing the language in a diachronic context, and so using existing acquisitional models as part of an evolutionary simulation may increase their

---

[60] We should note, however, that neither does the model provide evidence that the Language Acquisition Device does not impose the constraint that acquired languages must be partition languages. It is quite possible that the acquisitional model used in these simulations is not accurate in this respect.

explanatory power, thus demonstrating the importance of the computational evolutionary approach to linguistics.

Chapter 5

Adding Unique Hue Points to the Model

In Chapter 3 and Chapter 4 it was shown that the Bayesian model can account for some of the properties of basic colour terms, but so far the model is completely unable to account for the typological data. In section 3.2.4 it was proposed that the model should incorporate points in the colour space termed unique hue points. These points are especially salient, and appear to be particularly well remembered by people compared to other areas of the colour space (see section 2.3). However, in the version of the acquisitional model used so far, no special status was given to any part of the colour space. This chapter describes the new model which was created to remedy this earlier omission, and investigates what predictions it makes concerning colour term acquisition.

5.1 Specification of the New Model

The previous acquisitional model (described in Chapter 3), treated the colour space as a continuous dimension, but this necessitated the use of calculus for performing summations over ranges of hypotheses. As should be apparent from section 3.4, this necessitated the use of some fairly complex maths, and the maths would have been

more complex if unique hue points had been incorporated into the model. To avoid this problem, the colour space was divided into 40 discrete colours, and so each individual colour could be indexed with a number between 1 and 40. The choice of 40 colours is an arbitrary one, but treating the colour space as a number of sections allowed discrete summation to be used rather than calculus, hence making the maths simpler. This difference also allowed the model to run more quickly[61], and yet it did not appear to greatly affect the results obtained[62]. We may note that, as the number of unit colours in the whole colour space is increased, the model will increasingly closely correspond to the version with a continuous, non-discrete, colour space.

Using this coordinate system, the red unique hue point is at hue 7, yellow at 19, green at 26, and blue at 30, so that the largest distance between adjacent unique hue points is 17 units between blue and red, and the smallest is just 4 units between green and blue. Yellow and green are also considerably closer than are red and yellow. These locations where chosen partly by a process of trial and error, and were adjusted in order to make the results of the model parallel the empirical data as closely as possible. Note, however, that these values are compatible with the less precise specification, concerning the unique hue point locations, made in section 3.2.4. Each

[61] While the time complexity of the first model was cubic on the number of examples of each colour term, the second model was constant with respect to this variable. However, the second model was quadratic with respect to the number of units into which the colour space was divided, a measure not relevant to the first model.

[62] No rigorous investigation was made concerning the differences between the first and the second model, but a number of tests showed that the new model learns colour term denotations which have similar properties to those produced by the first model.

colour is assigned a probability corresponding to how likely a person is to remember an example if it corresponds to that colour. These values will be written R_c, and were set at 0.05 for colours which did not correspond to unique hue points, and at 1 for unique hues, so that it was 20 times as likely that examples of unique hues would be remembered as examples of other colours. These values were also chosen largely by a process of trial and error, and hence are fairly arbitrary. However, they implement the specification made in section 3.2.4, concerning the relative memorability of unique hues and other colours. These changes were made in the hope that they would allow the model to account for typological data. The new model is essentially the same as the original one in other respects, as will become evident from the description below.

The next step towards achieving an implementation, was to specify how values could be calculated for the terms in Bayes' rule. First of all we needed to identify what would correspond to hypotheses for the purpose of learning colour words' denotations. Again a hypothesis will correspond to one possible denotation of a colour word, and so will specify the range of colours that come within the word's extension if that hypothesis is correct. A hypothesis can vary in size from taking up only one unit of the colour space to taking up the whole of the space, and can start and end anywhere in the colour space[63]. It was noted above, in section 3.2.5, that all such hypotheses are considered to be equally likely *a priori*, so all hypotheses will have equal *a priori* probabilities. Hence the term *P(h)* will be the same for all hypotheses.

---

[63] We should note that there will hence be 40 hypotheses of each size, except for the largest hypothesis, which takes up the whole colour space, as there can only be one such hypothesis.

Determining the probability of the data with respect to a hypothesis, *P(d | h)*, is somewhat more difficult, because this probability will vary depending on how accurately the hypothesis predicts the observed examples. If an example is accurate, then it must appear within the range of the hypothesis. If that is all we know about an example, then it is equally likely for that example to have been observed on any of the colours with the range of the hypothesis, assuming that the hypothesis is correct. However, because some examples will be forgotten, the proportion of examples which we would expect to have remembered for each particular colour would be equal to the probability of remembering examples of that colour, divided by the sum of the probabilities of remembering examples on all the colours within the range of the hypothesis, which will be written as R_h[64]. This ratio, which is given in (5.1), would correspond to the probability of an example, when that example was within the range of the hypothesis, and when we knew both that the hypothesis was correct, and that the example was accurate.

(5.1) $\frac{R_c}{R_h}$

Erroneous examples are equally likely to be observed anywhere in the colour space, but because we are more likely to remember them if they occur at unique hue points than elsewhere, the actually probability of an erroneous example being remembered at any particular colour, is equal to the probability of remembering an example if it occurs at that colour, divided by the sum of the probabilities of remembering

[64] Again this assumes that people are equally likely to observe examples of colours anywhere in the colour space. For discussion of the validity of this assumption see section 3.3.1.

examples of all colours throughout the colour space (R_t). This ratio is expressed as (5.2).

(5.2) $\frac{R_c}{R_t}$

(5.1) and (5.2) apply when we know whether an example is accurate or not, but in reality, when a person has remembered an example, they will not be sure whether or not it is accurate. A parameter *p* was added to the model, which corresponds to the probability that each individual example is correct. Now, if we see an example outside of the hypothesis space, we know that it must be inaccurate (we are still assuming that the hypothesis is correct). Because the probability that an example is accurate is *p*, the probability that it is not accurate is *(1-p)*. So the overall probability of an example which comes outside of the range of a hypothesis will be the probability of an inaccurate colour being observed at that point in the colour space multiplied by the probability of an example not being accurate, and so the probability of each example which is outside of the hypothesis is given by (5.3).

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

However, if an example is within the scope of the hypothesis, then we cannot be sure whether it is accurate or not (because it could have come within the range of the hypothesis simply by chance). So, in the case of an example which is within the range of a hypothesis, we have to add its probability assuming that it is accurate, to what its probability would be if it was erroneous, each of which must be weighted by the probability of examples being accurate (*p*), or inaccurate (*1-p*). The resulting overall probability of such examples is given by (5.4).

$$(5.4)\ P(e \mid h) = \frac{pR_c}{R_h} + \frac{(1-p)R_c}{R_t}$$

Equations (5.3) and (5.4) allow us to calculate the probabilities of individual examples with respect to a hypothesis, but usually we will have several examples for a particular colour word, so we need to combine these individual probabilities to obtain an overall probability for all the data. This can be done simply by multiplying together the probabilities of each individual example, *e*, from the set of all examples, *E*, as shown in (5.5). For every example, we must use either equation (5.3) or (5.4) to calculate *P(e | h)*, depending on whether or not the example is within the scope of the hypothesis.

$$(5.5)\ P(d \mid h) = \prod_{e \in E} P(e \mid h)$$

So far we have specified two of the three terms on the right hand side of Bayes' rule, but in order to determine hypotheses *a posteriori* probabilities, we also need to be able to assign a value to the third term, *P(d)*. This is the probability of the data, before we start to consider any particular hypothesis. We can again calculate this probability by multiplying the probability of the data, given each individual hypothesis, by the *a priori* probability of each hypothesis, and then totalling the resulting probabilities for each hypothesis in the set of all possible hypotheses, *H*. This is expressed mathematically in (5.6).

$$(5.6)\ P(d) = \sum_{h \in H} \left[P(h) P(d \mid h) \right]$$

If we substitute this equation into Bayes' rule, as it was given in (3.1), we obtain equation (5.7), which we can simplify by cancelling out the constant terms *P(h)* and $P(h_i)$. (The *h*'s of equation (5.6) now have a subscript, *i*, to distinguish them from the

specific hypothesis under consideration, *h*. However, as the *a priori* probability of all hypotheses is equal, each $P(h_i)$ will be equal to $P(h)$.)

$$(5.7)\quad P(h \mid d) = \frac{P(h)P(d \mid h)}{\sum_{h_i \in H}\left[P(h_i)P(d \mid h_i)\right]} = \frac{P(d \mid h)}{\sum_{h_i \in H} P(d \mid h_i)}$$

Equation (5.7) lets us calculate the probability of an individual hypothesis which corresponds to one possible denotation of the colour word. However, remember that what we really want is to calculate how likely it is that a colour can be denoted by the colour word, given all the examples that have been remembered. We can do this by calculating the *a posteriori* probability of each hypothesis which includes the colour within its range, and adding together all these probabilities. This will give us the overall probability that the colour comes within the denotation of the colour word.

As with the previous model, if we calculate such values for all colours, then we will obtain a fuzzy set, where each colour is assigned a degree of membership in the category corresponding to the colour word, according to the probability with which a person believes that that colour comes within the extension of the colour word. The only difference between the fuzzy sets obtained with this model, and those obtained with the previous one, is that those of the present model assign degrees of membership at 40 discrete points throughout the colour space, while, with the first model, the degree of membership changes continuously throughout the colour space.

Like the original model, this model was also written in C++, and the source code is given in Appendix B. Most of the code is the same, with the main differences being in the parts of the program which actually calculate the probability of colour words denoting particular colours. It generally executes somewhat more quickly than the original model, although, when only a small number of examples of each colour

(about ten) are used, there is no great difference. In all of the simulations using this model, the value of the parameter, *p* (which occurs in equations (5.3) and (5.4)), was set to be 0.5.

5.2 Predictions of the Acquisitional Model

While the aim of this research was to model the evolution of colour words, it is worth investigating exactly what the acquisitional model was able to explain on its own, before it was placed in an evolutionary context. The ability of the model to learn colour term systems was investigated by presenting it with examples corresponding to the colour term system of Urdu[65]. The data was obtained from a colour chart showing the range of colours which each word denotes, which was published in Berlin and Kay (1969). From the chart, the approximate range of colours denoted by each colour term was determined, and these were then mapped onto the colour space of the Bayesian model, taking account of the location of the unique hues in the model, which do not correspond exactly to their locations on Berlin and Kay's chart. Berlin and Kay report that Urdu has 8 basic colour terms, corresponding roughly to the English terms *red*, *yellow*, *green*, *blue*, *purple*, *brown*, *black* and *white*, although there was no term for *grey*, *orange* or *pink*. Because the model is restricted to only considering the dimension of hue, only the red, yellow, green, blue and purple terms are relevant here. It would appear that the Urdu term for yellow has extended to include most of the colours which would be denoted by *orange* in English, although its prototype is in the same part of the colour space as that of English *yellow*.

---

[65] Urdu was chosen simply because data concerning Urdu colour terms was readily available, and because Urdu has a colour term system somewhat different to that of English.

Hues were then chosen at random, and the corresponding Urdu colour term was determined in each case. These hues were then passed as examples to the model, except that only 5% of those examples which were not of unique hues were remembered (as compared to all the examples for unique hues). This process was repeated until the model had remembered 40 examples. This process simulated the acquisition of the colour term system, although in the condition where it was learned from a completely reliable informant, who was always completely consistent in which colour term he used to name each colour[66].

Once the training of the model was complete, the degree of membership in the colour category corresponding to each colour word was determined, and is displayed in Figure 5.1[67]. In the graph, the unique hue points are labelled +, the leftmost one being red, then yellow, green and finally blue. We can see that each of the colour terms which contains a unique hue point has that point as its best example, which is consistent with the empirical results. Furthermore, each colour word has prototype properties, with one particular colour being the best example of it, but colours further

---

[66] We should note that the colour space in the model does not correspond to the Munsell colour space which Berlin and Kay (1969) used in collecting the data. As, at least in this thesis, distances between hues in the Munsell colour space are regarded as being somewhat arbitrary (see section 2.1), the mapping of colour terms between the Munsell colour space and the one used by the model is fairly approximate. A colour term which denotes a large part of the Munsell colour space might denote a much smaller part of the colour space in the model, or vice versa, and there is no single correct way to map from one colour space to another.

[67] The raw data for this experiment, and for all the other experiments performed with the model using a discrete colour space, is given in Appendix C.

away from that colour being progressively worse examples of the colour word. This experiment has therefore demonstrated that the model is able to learn the colour term system of a real language. Further experimentation has shown that colour terms denoting both larger and smaller regions of the colour space can be learned by the model, and that all these terms will normally have prototype properties. Furthermore, there is a very strong tendency for the prototypes of colour terms to occur at unique hues.

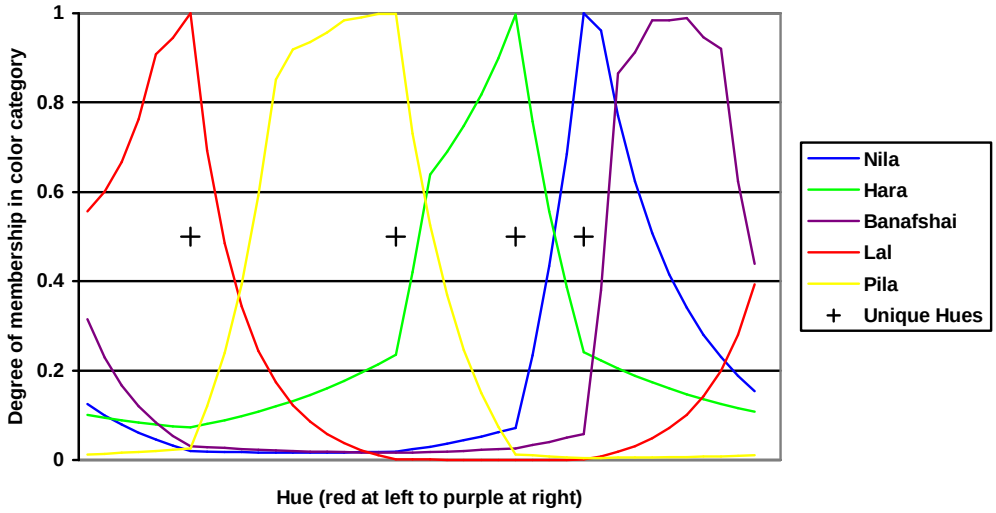


Figure 5.1. Learned Denotations for Urdu Colour Terms.

(*p*=0.5, the colour space is of size 40, and the unique hues are at 7, 19, 26 and 40.)

The model still does not explain the typological patterns seen in colour term systems cross-linguistically, because, as was shown in section 4.1 using the previous version of the model, colour systems which were of types which had not been attested cross-linguistically could equally well be learned. With the new model, it is probably in general easier to learn attested colour term systems than non-attested ones, because, for example, green and blue are closer together than are blue and red, so attested terms, such as green-blue composites, can be learned more easily than unattested

ones, such as blue-red ones. However, a number of informal experiments have shown that languages of unattested types can still be learned. Hence, we cannot use this model to explain colour term typology as the result of a purely psychological process. The next stage of the research involved incorporating the acquisitional model into a social context, in order to produce an expression-induction model, and so investigate whether we could account for colour term typology as the product of an evolutionary process.

Chapter 6

Simulating Colour Term Evolution

As the acquisitional model of Chapter 5 is, on its own, insufficient to account for all the typological data concerning colour terms, it was decided to test Berlin and Kay's (1969) original hypothesis that the typological patterns observed in colour term systems across languages are the result of an evolutionary process. This was done by incorporating the Bayesian model into an evolutionary expression-induction model, which simulated the historical change of colour term systems over several generations.

6.1 The Evolutionary Model

The evolutionary model was based on the one described in section 4.2. It also contained ten copies of the acquisitional model, each of which corresponded to a simulated person. As noted above, this community is of course unrealistically small, but, as with the previous model, the results obtained do not appear to differ radically with larger communities[68]. Therefore, simulations using ten artificial people were

[68] This conclusion was obtained based on the results of several experimental simulations using various numbers of artificial people, but the effect of the number of artificial people on the results was not investigated rigorously.

considered to be adequate, especially as this allowed the program to run much faster than when more artificial people were used. As with the simulations reported in section 4.2, in the initial state of the simulation, each of the artificial people had observed one example of a colour word, which would have been used to name a randomly chosen hue[69]. The colour words were randomly created by concatenating any three letters together, and so usually each person would know a different colour word.

Ages were assigned to each person in the same way as in the previous evolutionary simulation (section 4.2). To obtain the results reported in this chapter, the average lifespan was variously set at any of the values 18, 20, 22, 24, 25, 27, 30, 35, 40, 50, 60, 70, 80, 90, 100, 110 or 120, with 25 separate simulations being made in each condition. (Recall from section 4.2, that lifespan corresponds to the number of colour terms that each artificial person will observe on average during their lifetime.) The simulation then proceeded in the same way as the previous ones, with the exception that when examples were passed to the hearers, if they did not correspond to a unique hue then there would only be a 5% chance of them being remembered (compared to a 100% chance of unique hues being remembered). The algorithm is summarised in Figure 4.2, and the rate of creativity was kept the same as in the previous evolutionary simulation. Hence, on average, a new colour word would be introduced once in every 1000 times that any artificial person spoke. All the simulations reported in this chapter

---

[69] Because the model has a tendency not to remember examples of colours which are not unique hues, several examples may need to be given to the model before it remembers one of them. There is hence a greater probability for the initial example to correspond to a unique hue rather than another colour.

were run for a period of time equal to twenty average life-spans of the people in the simulation. As there were 17 different lifespans used, and 25 simulations were performed under each condition, there would be a total of 425 separate simulations. The results reported in this chapter are based on the languages spoken by the people at the end of these simulations. In all cases, the value of the parameter, *p*, was set at 0.5.

6.2 Emergent Colour Term Systems

A general picture of the kind of colour term system typically emerging in the simulations can be obtained by examining Figure 6.1. This shows the colour term systems of four people from the end of one evolutionary simulation, in which the average lifespan of the artificial people was set at 100. As is the case with all the results reported in this chapter, only colour terms for which the person has observed at least four examples were included. This was because one of the necessary criteria for a colour term to be considered basic is that it must be salient for a speaker, and it seems reasonable to propose that if a person has observed only one or two examples of a colour word, then that word would be less salient for that speaker than one for which they had observed more examples. We can see that, with respect to this criterion, speaker (a) and (b) both know four basic colour terms, with their prototypes at the red, yellow, green and blue unique hue points (which are marked by crosses). Speaker (c) knows these same four terms, plus two extra terms, which both denote purple hues, while (d) does not know the green term, and so has only three basic colour terms.

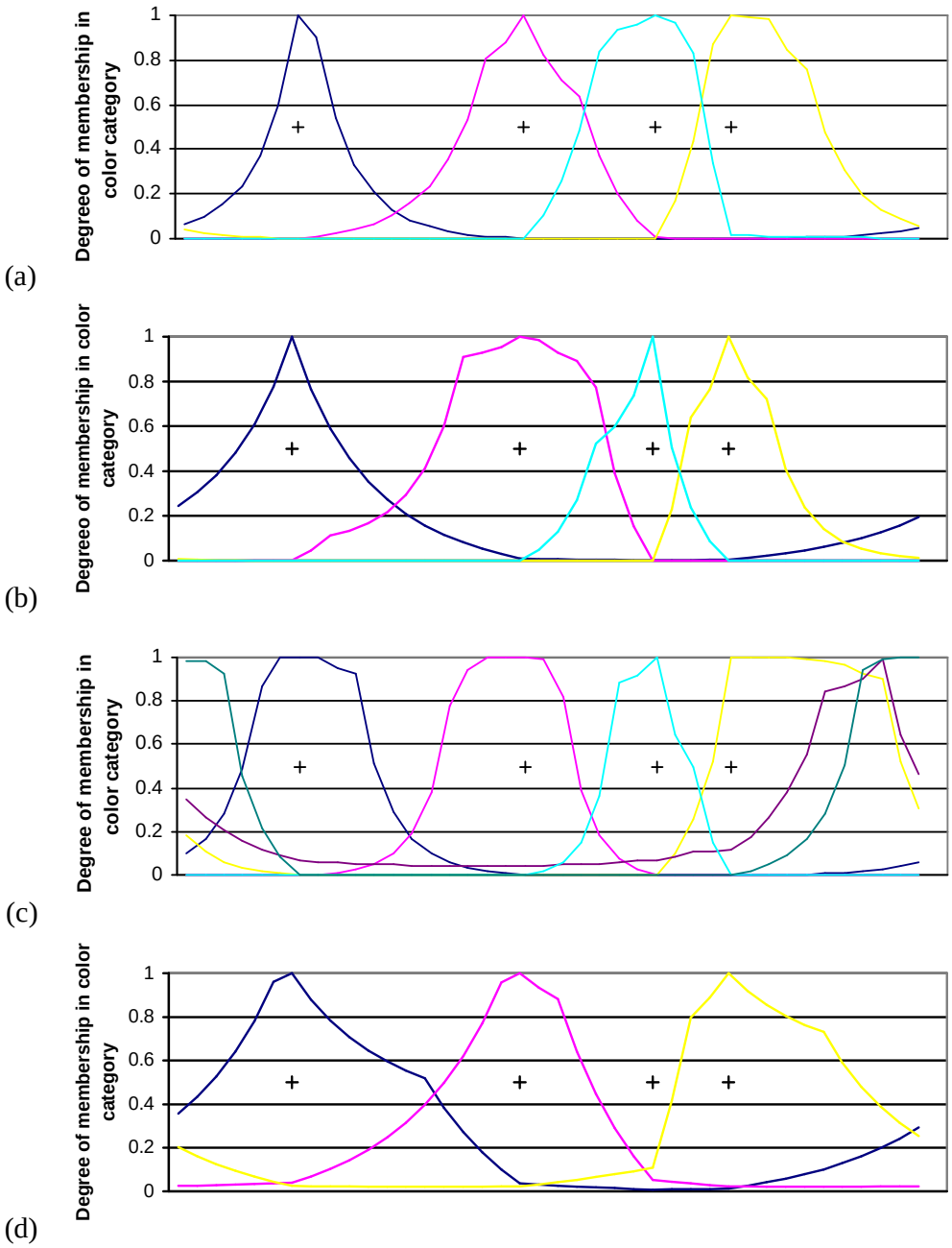


Figure 6.1. The Basic Colour Term Systems of Four Artificial People from the same Simulation.

(+'s mark unique hue locations, *p*=0.5, and people are creative one time in a thousand.

Only terms for which people have remembered at least four examples are included.)

As the average lifespan of the artificial people was 100, they would, on average, remember 100 examples during their lifetimes. (a) was aged 78, (b) 72, (c) 113, and (d) 38. This helps to explain why (c) knows more basic colour terms than any of the

other people, because he has had more chance to observe examples, and so will have seen enough examples of each term for it to attain basic status. (d), being aged only 37, has seen many fewer examples, and so has not observed enough examples of the green term for it to be considered basic. In fact (a) and (b) have both observed examples of each of the purple terms, but not enough for either of these terms to be considered basic for those speakers. All of (a), (b) and (c) have also seen at least one example of an orange term, though this term was not basic for any of the people in the simulation. This particular simulation, and, within it, these four artificial people, were chosen fairly arbitrarily, simply in order to illustrate one particular language community. The other people in the simulation had similar colour term systems, except that the youngest people had seen very few colour examples, and consequently did not know as many colour terms, and nor had they learned the denotations of all the colour terms accurately. Full details of this community appear in Appendix C.

We can see that a colour term system has emerged which is in general shared by most members of the community, but in which there is considerable variation between individual people. The variation concerns which colour terms each person knows, which they consider to be basic, and the exact denotation which they have learned for each colour term. Sources such as MacLaury (1997a) report that these are all phenomena that are prevalent in real language communities, and so I would see it as a strength of the model that there are inconsistencies between the individual I-languages learned by each person. In reality, probably no two people speak exactly the same language, and this is equally true in the simulation.

6.3 Number of Basic Colour Terms

Clearly, the number of basic colour terms which exist varies considerably between different languages. Empirical evidence appears to show a correlation between the type of society in which a language is spoken, and the number of colour terms in the language (Berlin and Kay, 1969; MacLaury, 1997a). Languages spoken by tribal people, with relatively low levels of technology, tend to have fewer basic colour terms than languages spoken in highly industrialized societies. One of the crucial differences between these environments would seem to be the range of coloured objects, and the uses to which colour is put. In tribal societies, there will typically be a limited range of coloured objects available, and those objects will tend to simply have their natural colours, or the colours of the limited range of dyes which are available. In contrast, in highly industrialized societies, we can make many objects in any colour we want, and so there is much more opportunity to use colour to identify particular objects. We even use colour as a form of language in itself, for example in colour coding of electrical wires, traffic lights, etc. These factors suggest that people in industrialized countries might use colour words much more frequently than people living in societies with lower levels of technological development, and it is possible that this is the reason for the variation seen in the number of colour words in the languages of these societies. It was investigated whether a relationship holds in the simulations, between how often people use colour words during their lifetimes, and the number of basic colour words emerging in the languages which they speak. The results in this and the following section are based on all 425 runs of the simulation, in which the average lifespans of the artificial people was varied from 18 to 120, as described in section 6.1.

In general, when trying to classify the languages spoken in each simulation, only people whose age was greater than or equal to half the average lifespan were included. (These people are subsequently referred to as ‘adults’.) This is because younger people might not have observed enough examples to have determined with a reasonable degree of accuracy the language spoken in the community. In practice when field linguists collect mappings showing the denotations of basic colour terms, they would also exclude young children from their studies, instead relying on adults, who would be expected to have completed the process of language acquisition, and to have relatively stable language competencies.

For each such person alive at the end of each simulation, the number of colour terms of which they had observed at least four examples was determined, and these values were used to calculate the mean number of colour words spoken by people under each condition of the simulation, where the average number of colour examples observed during a person’s lifetime was varied from 18 to 120. These means are plotted in Figure 6.2, and are given together with the corresponding standard deviations in Table 6.1. We can see that there is a clear positive correlation between how often people use colour terms during their lifetimes, and the number of colour terms in their language, so this could provide an explanation of why languages of industrialized societies tend to have more colour words than those of less industrialized societies.

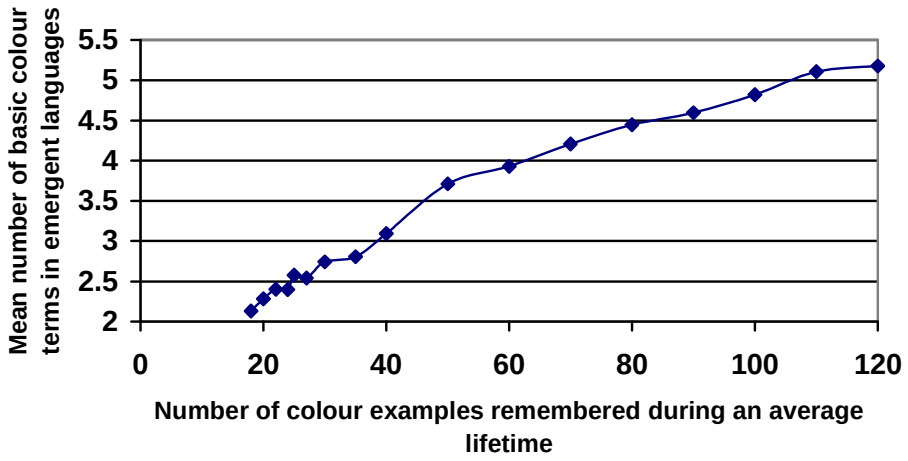


Figure 6.2. Relationship of Number of Frequency of use of Colour Terms to Number of Basic Colour Terms in Emergent Languages.

Number of colour examples remembered on average during lifetime	Mean number of basic colour terms in emergent languages	Standard Deviation of number of basic colour terms in emergent languages
18	2.13	0.373
20	2.28	0.452
22	2.40	0.492
24	2.39	0.510
25	2.57	0.496
27	2.53	0.515
30	2.74	0.493
35	2.80	0.618
40	3.09	0.538
50	3.71	0.584
60	3.92	0.419
70	4.20	0.528
80	4.44	0.626
90	4.59	0.712
100	4.82	0.809
110	5.10	0.896
120	5.17	0.452

Table 6.1. Means and Standard Deviations of the mean number of Basic Colour Words in Emergent Languages.

The above result is probably unsurprising, but we should note that nothing was built into the model to force people who use colour words more often to have more colour words in their languages. We could conceive of a situation in which people who used colour words more frequently simply used the same number of words, but used each

one of them more often. This would give each person the opportunity to observe more examples of each word. However, the above graph shows that this does not appear to be what happens in the simulations. We should note that the artificial people do not receive any reward for communicating, nor in fact any feedback about whether communication is successful, so they cannot have added more colour words to their languages because this helps them communicate more effectively. It is not clear exactly why this relationship between frequency of use of colour words and number of basic colour terms holds, but it does suggest that, in general, we may have a lot of words in a particular domain simply because we talk about that domain often, rather than because those words are actually useful.

6.4 Typological Analyses

The previous section looked at the results of the simulations in a very general way, because it took account only of the number of colour words emerging, but not of what ranges of the colour space each one denoted. In order to investigate whether the typological patterns identified by Kay and Maffi (1999) are replicated in the simulations, it was necessary to first classify each language in terms of which kinds of basic colour terms it contained. In order to make this process consistent and objective, a number of rules for classifying colour terms were developed.

Firstly, for each person, every colour was considered in turn, and calculations were made to determine how confident the person was that that colour could be named by each colour term which the person knew. The colour term which received the highest confidence score was taken to be the colour term which the person would use to name that colour. The results of this process, for one run of the simulation, in which people observed 60 examples on average during their lifetimes, is shown in Figure 6.3, where

each colour term is given an arbitrary label, A, B, C or D, and each row represents one person. (The choice of this particular community to serve as an example was made completely arbitrarily.) Each column corresponds to a colour, with hue 1 at the left, and hue 40 at the right. The boxed columns correspond to the unique hues, red, yellow, green and blue.

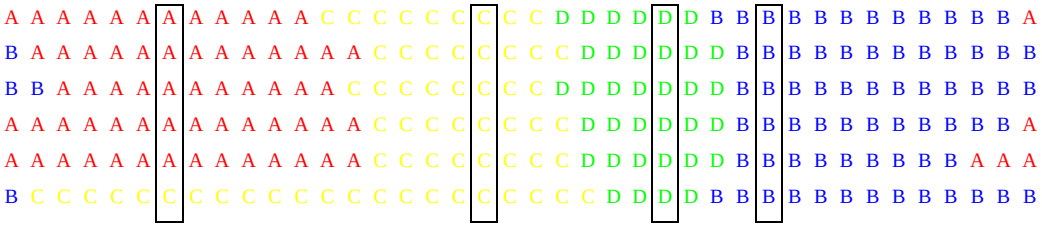


Figure 6.3. Denotations of Basic Colour Terms for all Adults in a Community.

(Adults are people whose age is over half the average lifespan of people in the community.)

The full range of each colour term was then considered to be the smallest range of colours which included all the hues which would be named by the colour term, if the person always used the term which they were most confident was correct. (This could potentially include some hues which would be named by a different colour term, because there might be another term which had greater confidence values for a smaller range of colours within the range of hues named by the first term.) Colour terms were classified as *red, yellow, green* or *blue* if they included the corresponding unique hue point, and no other unique hue point. If they did not include any unique hue points, then they were classified as *orange*, *lime, turquoise* or *purple*, depending on whether they were between the red and yellow unique hue points, the yellow and green ones, the green and blue ones, or the blue and red ones respectively. If a colour term included more than one unique hue point, it would be classified as a composite of those unique hue points, for example *red-yellow* or *yellow-green-blue*.

The next stage of the analysis consisted of determining which colour terms the language spoken by each community as a whole could be said to contain. This process was not entirely straightforward, because not all terms would necessarily be considered to be basic for all speakers, nor would each term necessarily be given the same classification for each speaker. (We should note that this is entirely consistent with empirical data, because mappings of colour terms on charts collected by field linguists for different speakers of the same language typically show considerable discrepancy, and often one informant will not report all the colour terms used by other speakers (MacLaury, 1997a).)

If we examine Figure 6.3, we can see that the classification of all the terms would be the same for the first five speakers, in that A contains the red unique hue point and so would be classified as *red*, and B, C, and D would likewise be classified as *blue*, *yellow* and *green* respectively. However the sixth speaker, who corresponds to the bottom line of the chart, has not seen enough examples of term A for it to be considered basic, and consequently he names the red unique hue, and the surrounding colours, with term C, which he also uses to name unique yellow. C was therefore classified as a *red-yellow* term for this speaker.

In order to arrive at a consistent classification for each community, a number of rules were derived for cases in which speakers disagreed on the denotation of a colour term. Firstly, a colour term was considered to be basic in a language only if it was known by at least half the adults. This criterion was justified because one of Berlin and Kay's (Berlin and Kay, 1969) original criteria for a colour term to be considered basic was that it should be known by all members of a community, so it would seem likely that if this term was not salient enough to be considered basic by more than half the

speakers, then this criterion might well not be satisfied. Colour terms which did not satisfy this criterion were excluded from the analysis.

The second rule was that, if not all speakers agreed on the classification of a term, in terms of which unique hues it denoted, then that classification which was supported by the greatest number of people would be chosen. If this rule cannot be applied, because two or more possible classifications are supported by equal numbers of people, then, if one of the terms contained fewer unique hue points, it would be chosen over a term which contained more unique hue points. This would introduce a conservative bias into the classification system, so that if there was doubt as to the exact range of colours which a term denoted, a smaller range would, in general, be chosen over a larger one. If the application of all these principles failed to produce a unique classification for each term in a language, then the whole language would be excluded from the analysis. In addition, if a speaker did not know any colour terms for which they had observed at least four examples, or if they knew only one such term, they would not be considered during the analysis[70]. After the application of all

---

[70] Speakers knowing only a single term would use that term to name all parts of the colour space, including even those hues which they believed were very unlikely to come within the term's denotation, and it was for this reason that such speakers were excluded from the analysis. If colour terms were really used in this way, they would appear to be communicatively useless. It would seem more likely that real people would simply not use any word to name hues which did not have at least a moderately high degree of membership in the denotation of any colour term. However, taking account of this would have complicated the analysis, as it would have necessitated adding a parameter concerning the degree of membership in a colour term's denotation at which to demarcate its boundary, and so no such addition was made to the analysis system.

these criteria, a unique classification was obtained for the languages emerging in 420 of the 425 runs of the simulation.

The number of terms which were classified as being of each type in all the emergent languages is listed in Table 6.2. For the terms which contain a unique hue point, these data were converted to percentages, and are shown in Figure 6.4. Both Table 6.2 and Figure 6.4 also contain equivalent data from the World Colour Survey (WCS), as reported in Kay and Maffi (1999). Kay and Maffi did not take account of colour terms which did not contain a unique hue point in making their classifications. Hence, the relevant data on these terms is absent from their paper, and consequently does not appear here. However, Kay and Maffi did take account of colour terms which are either achromatic, or distinguished from other colours on the basis of some dimension other than hue. Such colour terms have simply been excluded from the analysis. Composite terms which included white or black and one or more unique hue would be treated as if they only contained the unique hue. For example, a colour term denoting warm colours (white, red and yellow), would be treated the same as a colour term denoting only red and yellow. This is necessary in order to reconcile the results of empirical surveys with those of the model, in which the colour space has been reduced to a one dimensional hue circle.

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

Table 6.2. Frequencies of Colour Terms of each type in the Simulations and the World Colour Survey.[71]

Kay and Maffi (1999) did not arrive at an unambiguous classification for 25 of the 110 World Colour Survey languages, as those languages appeared to be in transition between two of the types attested in their evolutionary sequence, so these languages were not included in determining the total numbers of each type of colour term attested. There were also six languages which did not appear to fit into any clear evolutionary sequence, so these too were not counted when compiling the data shown

[71] This table could also potentially include Red-Yellow-Green-Blue composites, which included every unique hue. However, no such terms were found in the World Colour Survey or in the results of the simulations. We should remember though that terms which were used to name every hue were excluded from the analysis of the results of the simulations, so this could have eliminated some such terms. Red-Yellow-Green-Blue composites emerging in the simulations could only have been included in the analysis if the person using the term used another term to name at least one part of the color space in between unique hues.

in Table 6.2 and Figure 6.4. The total counts for each colour term were derived simply by counting them once for every language in the World Colour Survey which Kay and Maffi reported contained those colour terms as basic colour terms.

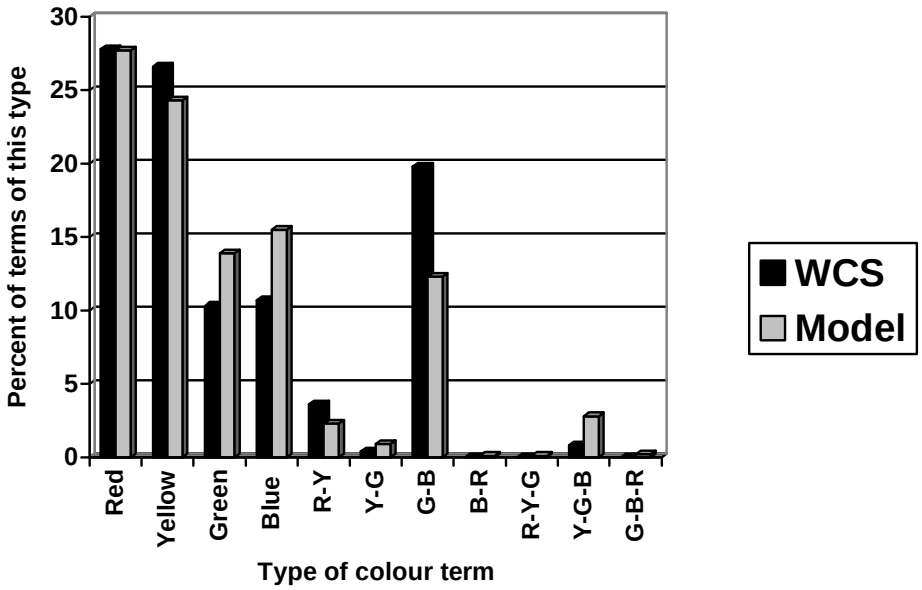


Figure 6.4. Percentage of Colour Terms of each type in the Simulations and the World Colour Survey.

Figure 6.4 clearly shows a close relation between the frequency of each term in the World Colour Survey and in the simulations. The key differences between the empirical data and the results of the simulations are that the simulations produce somewhat too many Yellow-Green-Blue composites, and too few Green-Blue ones. As was mentioned above, there is also empirical data concerning the relative frequencies of basic colour terms which do not contain unique hues. It has been reported (for example by MacLaury, 1997a), that both purple and orange derived terms are frequently reported, and that purple terms occur more frequently than do orange ones. This finding is also supported by the results of the simulations given in Table 6.2, in which it can be seen that more emergent languages contained purple terms than orange terms.

As well as accounting for the expected colour terms, the simulations also produce a small number of terms of types which have not been attested empirically. There are 1 Blue-Red composite, 1 Red-Yellow-Green composite, 3 Green-Blue-Red composites, and 4 Lime terms. The presence of a small number of previously unattested colour terms should not be surprising. As linguists have examined the colour terms of more and more languages, colour terms of types which were not found in Berlin and Kay's original survey (Berlin and Kay, 1969) have been discovered. The evolutionary model reported in this chapter does not place absolute restrictions on the types of colour terms which can evolve, but simply introduces biases so that some kinds of colour term emerge much more frequently than others. The simulations produced 425 languages, while the World Colour Survey has sampled only 110, so it would seem likely that it would have missed some types of colour terms which might exist in some human language. Even if the colour term system of every language in the world were examined, some colour terms which might potentially evolve may not be seen, simply due to historical accident, which has resulted in no language which exists at this point in time having evolved in such a way as to include that colour term. Hence this kind of result is unproblematic because, as Poortinga and Van de Vijver (1997) argued, it appears that 'constraints on colour categories are probabilistic rather than deterministic' (p205).

The only colour terms which it would seem might be problematic are the Green-Blue-Red composite and the Lime term, because these terms both occur in several simulated languages. It is not clear if it is best to explain these terms simply as diverging randomly from the data of the World Colour Survey, or whether the evolutionary model is at fault in over-predicting the occurrence of these terms. There is also the possibility that such colour terms do exist in real languages, but that they

have either been given an alternative analysis by the linguists making the investigation, or have been classified as non-basic, and hence excluded from the analysis. The process by which Kay and Maffi's (Kay and Maffi, 1999) classifications were produced was necessarily somewhat impressionistic and subjective, and it is not clear how much attention is paid to existing theories when eliciting and analyzing field data. When I have discussed this issue with field linguists, they have sometimes suggested that it is likely that existing theories influence how field data is analysed, and what data is elicited. Furthermore, when other linguists come to reinterpret that field data, there are further opportunities for the raw data to get distorted. It would certainly seem that in a language such as English there is a lime colour term, namely 'lime' or 'chartreuse', but it seems clear that this term should not be considered a basic colour term with respect to the criteria of Berlin and Kay (1969), so perhaps the lime terms emerging in the simulation should likewise have been excluded from the analysis.

Kay, Regier, Cook and O'Leary (n.d.) have acknowledged the need for more rigorous methodology, and report that they are undertaking a project to comprehensively investigate World Colour Survey data using statistical tests, but results are not yet available. MacLaury (1997a) analysed some of his data using statistical tests, but it would seem that the way in which he sampled the data may mean that few interesting conclusions can be drawn from the results. If we wish to compare two languages, then we would normally draw a number of samples from each language, probably taking each sample from a different speaker. We could then test for a difference between the set of data for the first language, and that for the second. However, if we wished to make a generalisation about languages in general, then we would take one sample for, for example, each of two or more constructions in each language, and then we could

look for differences between types of language, or correlations between particular features. The samples taken for each language could be based on data from individual speakers, or on an analysis of the language as a whole (as in the case of the results for colour term systems given above). However, MacLaury appears to sample data by taking samples from several languages, while also including more than one sample from each language (MacLaury, 1997a, 1995, p241). This is because, for at least some of the languages, he takes one sample from each of several speakers. It would seem that the results of statistical tests which analyse such data will be of little interest, because they can neither be used to make claims about languages in general, or about the particular languages concerned[72].

---

[72] This argument might be made clearer by considering an extreme case. Suppose that we wanted to compare languages of type A to type B, and we had available data from 5 languages of each type. If we took one sample from each language, then we would likely find that the results of a statistical test were not significant, which would not be surprising, because a cross-linguistic pattern might not show up in such a small sample of languages. If, however, for 4 of the languages of each type we had only 1 speaker available, but for the other 2 languages we had 1000 speakers, we might be tempted to include one sample for each speaker. We would now be much more likely to obtain a significant result for the statistical test (assuming that there is in fact a significant difference between the two types of languages in the feature for which we are testing), because there is now much more data available. However, I think that it will be clear that any result from this test should not be taken to be applicable to languages in general, because almost all of the data came from only two languages. Sampling in this way is simply a more extreme case of the way in which MacLaury (1997a) obtained samples for his tests. Of course, the sampling problem is a general problem in typology, because it is always difficult to argue that two languages are independent. A correlation between features seen in two languages could be due to contact between those languages, or to their historical relatedness, in which case we should take the

So far, the similarities and differences between the empirical data and the results of the simulations of this chapter have been compared only impressionistically, but it would clearly be preferable to make a more rigorous statistical analysis of the performance of the model. Hence, for each type of colour term (again excluding those which do not contain unique hues), the number of languages in which it occurs in the World Colour Survey is plotted in Figure 6.5 against the number of languages in which it occurred in the results of the simulations. There appears to be a clear positive correlation between these two variables, as is highlighted by the trend line. Pearson's product moment coefficient was calculated for these two variables, and it was found that there is a correlation of 0.959 between the results of the World Colour Survey and the simulations. This correlation is highly significant ($P \ll 0.01$)[73].

---

two languages as providing only a single example supporting the generalization that those features tend to co-occur universally.

[73] We should, however, use caution in interpreting this significance value, because several different values of the parameters of the model were adjusted in order to achieve a close fit between the predictions of the model and the empirical data. Hence we should really correct this value to account for the multiple comparisons. However, it would seem that this significance value is not overly important, as it is clear from inspection that the correspondences between the frequencies of each type of term in the simulations and the empirical data are not simply due to chance, and so it seems inconceivable that these results could constitute a type II error.

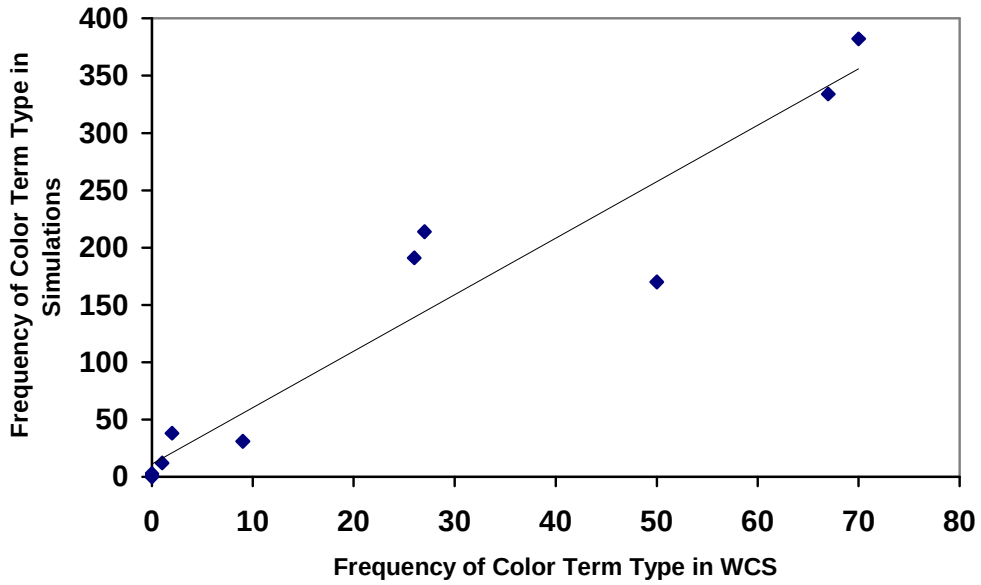


Figure 6.5. Scatter Graph Showing Relationship between the Frequencies of Colour Terms in the World Colour Survey and the Simulations.

Focusing simply on the frequencies of each type of colour term has enabled a clear and objective comparison to be made between the results of the simulations and the World Colour Survey, but we should remember that the evolutionary sequences which have been proposed (Berlin and Kay, 1969; Kay, Berlin, Maffi and Merrifield, 1997; MacLaury, 1997a; Kay and Maffi, 1999), have been based on analysis of colour term systems as a whole. Therefore, it seems desirable to examine what kind of overall colour term systems existed in each language emerging in the simulations.

Table 6.3 lists the 353 most common colour term systems found in the simulations. When the systems have been classified in terms of which colour terms they contain, and which they do not contain, only the eleven types of colour term system given in Table 6.3 occur five or more times. These results contain examples of all of the types

of system which Kay and Maffi (1999) placed in their evolutionary sequences[74], plus three other types of system. Two of these extra systems, Red, Yellow, Blue, and Red, Green-Blue, do not have a term which can consistently be used to name one of the unique hues, and so the overall classification for the systems do not include terms which can name either green in the first case, or yellow in the second. This situation has arisen because different people would name those hues with different terms, but the best classification that could be arrived at for those terms did not include the unique hue. The third unexpected type of system is the Red, Red-Yellow, Green-Blue system. This system contains two terms, both of which can be used to name the red unique hue. MacLaury (1997a) notes that many languages appear to have more than one colour term which can name particular hues, although usually one term is dominant and the other is a less salient term. Hence systems of this type do not seem to be problematic as far as the validity of the evolutionary model is concerned, and

---

[74] Remember that Kay and Maffi (1999) did not take account of colour terms which did not contain a unique hue when making their classifications. We should expect that purple and orange terms will be observed to exist together with terms containing a unique hue, and that we should see them more often in systems which contain more colour terms, than in those systems which contain fewer colour terms. Therefore, if a system corresponded to one on the evolutionary trajectory, except that it contained a purple or orange term (or both), it was regarded as coming within the evolutionary sequence. This is supported by empirical evidence, as MacLaury (1997a) notes that sometimes we will observe purple terms in systems in which there is still a green-blue composite term, and we do in fact see four purple, red, yellow, green-blue systems. There was also one purple, red, yellow-green-blue system. This was classified as being consistent with the evolutionary trajectories, even though systems with yellow-green-blue composites probably do not usually contain purple basic colour terms.

their existence might even strengthen the claim that it accurately parallels the evolutionary processes affecting real languages.

Basic Colour Terms in Language	Number of Languages with this Type of Colour Term System in Simulations
Red, Yellow, Green, Blue	112
Red, Yellow, Green-Blue	110
Purple, Red, Yellow, Green, Blue	44
Red, Yellow-Green-Blue	30
Red-Yellow, Green-Blue	22
Orange, Purple, Red, Yellow, Green, Blue	7
Red, Yellow, Blue	7
Red, Green-Blue	6
Orange, Red, Yellow, Green, Blue	5
Red, Blue, Yellow-Green	5
Red, Red-Yellow, Green-Blue	5

Table 6.3. The Most Common Colour Term Systems Emerging in the Simulations.

The simulations also produced 67 systems of types which occurred four times or less. Most of these systems diverged from those in Kay and Maffi's evolutionary trajectory (Kay and Maffi, 1999), because they either contained more than one term which could name a unique hue, or because no term was classified as denoting one of the unique hues. As noted above, five of the less common systems conformed to Kay and Maffi's evolutionary trajectories (the four purple, red, yellow, green-blue systems, and the one purple, red, yellow-green-blue system). Nine systems also contained colour terms which are of unattested types. All of the languages, including those already counted in Table 6.3, are summarized in Table 6.4 in terms of these criteria. The full classification, including the frequency of each type of colour term system, appears in appendix C.

Classification of Colour Term System	Number of Systems
Conforms to Evolutionary Sequence	340
Contain Unattested Colour Term	9
No term consistently names one or more unique hues	35
More than one term can name one or more unique hues, or there is more than one purple term	37

Table 6.4. Classification of How the Languages in the Simulations Diverge from the Attested Evolutionary Sequences[75].

What is clear from these results, is that there is a small set of colour term systems which occur very frequently, and that the colour term systems of the vast majority of languages can be classified as belonging to one of these types. However, there is also a significant subset of languages which diverge from the classification in some way or another, in that they have extra terms which could be classified as basic, they do not give a consistent name to one part of the colour space, or the exact combination of basic terms does not conform to the expected pattern. However, this finding is consistent with the empirical findings of Kay and Maffi (1999) and MacLaury (1997a), who note that many colour term systems are not easily classifiable in terms of one widely attested type. Kay and Maffi considered six of the 110 World Colour Survey languages to be exceptional and resistant to classification in terms of their evolutionary trajectories[76].

---

[75] There was one Orange, Lime, Purple, Purple, Red, Yellow, Green-Blue system. This system contains an extra purple term, and also a lime term, so it was classified both as having an extra term and an unattested term, and is counted twice in Table 6.4.

[76] Comparing the number of systems which are exceptional in the World Colour Survey and in the simulations is problematic, because the analysis of World Colour Survey data was somewhat subjective, and so the number of languages which was classified as being exceptional would depend on exactly how the analyses were made. In contrast, the analysis of the results of the simulations was

6.5 Investigating the Effect of Unique Hue Points

Section 6.4 discussed the results of the simulations largely in terms of the typological classification of Kay and Maffi (1999), but it did not look closely at the effect of the unique hue points on the emergent colour terms, or at the prototype structures of the colour terms. A question can be asked about exactly what effect the unique hue points have on the resulting languages. Even if the unique hue points were not added during the simulations, we would still expect to find, for example, more green-blue composites than blue-red ones, simply because the green and blue unique hues are closer together in the colour space. In order to investigate the effect of simulating unique hue points, the simulations were repeated without the unique points being made any more salient than any other colours. In all other respects, the simulations were identical to those performed previously, and again 425 separate simulations were performed, with the same life expectancies as before[77]. This section describes the results of these simulations, and compares them to those of the previous section, and

---

made by a computer program which applied precise rules. While this would have resulted in a completely objective analysis, the choice of the rules themselves was fairly arbitrary, and consequently so was the proportion of systems which was classified as being exceptional. Therefore it is difficult to determine whether there is a greater number of exceptional systems in the results of the simulations, or in the World Colour Survey.

[77] Note, however, that age and lifespan are always measured in terms of the number of examples remembered by agents, so that the same number of examples will, on average, be remembered by agents regardless of whether unique hue points are simulated or not. This is even though the simulation of unique hue points necessitates that agents do not remember a certain proportion of the examples generated.

also investigates just what effect the unique hue points have on colour terms containing them.

Figure 6.6 shows the number of colour terms with their prototypes at each of the 40 hues, both for simulations with unique hues and those without. All basic colour terms known by adults are included[78]. Each person is treated individually here, and so a colour term is counted once for every person who knows it. This shows very clearly that the simulation of unique hue points has concentrated most of the colour term prototypes on just four hues. In contrast, in the absence of simulated unique hue points, the prototypes of colour terms are distributed evenly across the space of colour terms. Hence the model provides a possible explanation of the observation that the prototypes of most basic colour terms in most languages appear to be clustered in a few small regions of the colour space.

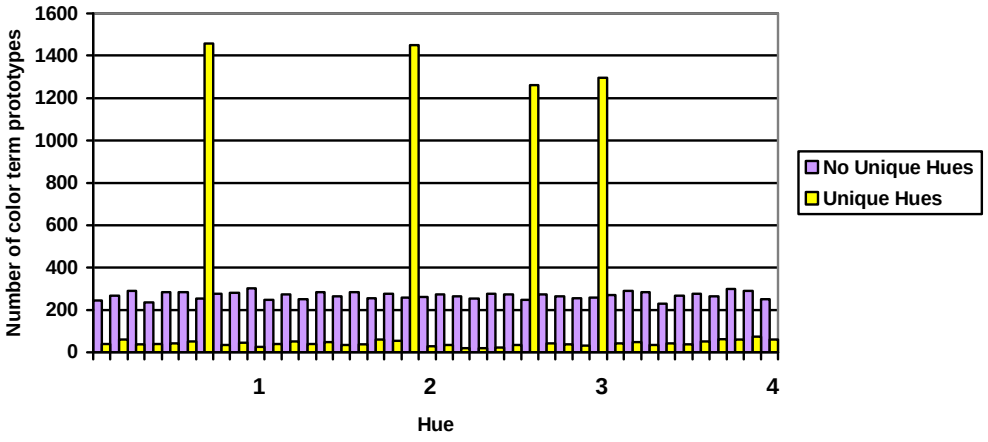


Figure 6.6. Locations of Colour Term Prototypes.

[78] Remember that an 'adult' is a person whose age is over half the average lifespan within the simulation, and a term is considered 'basic' if a person has remembered at least four examples of it.

However, it is possible that the evolutionary model is too constraining in the predictions it makes about where colour term foci will emerge. Figure 6.6 can be compared to Figure 6.7, which is an equivalent graph, except that it uses empirical data from the World Colour Survey. Figure 6.7 also shows four clear peaks, corresponding to locations in the colour space in which most of the colour terms have their prototypes. These are not in the same places as in Figure 6.6, because the unique hues are placed at different points in the colour space in the model compared to where they are in the Munsell system. However, if we ignore this point, it is still clear that there are significant differences between the two graphs. In Figure 6.6 the peaks are each on a single hue, while in Figure 6.7 they are more gradual, and spread out onto neighbouring hues. It is necessary, therefore, to provide some explanation of the differences between these two graphs.

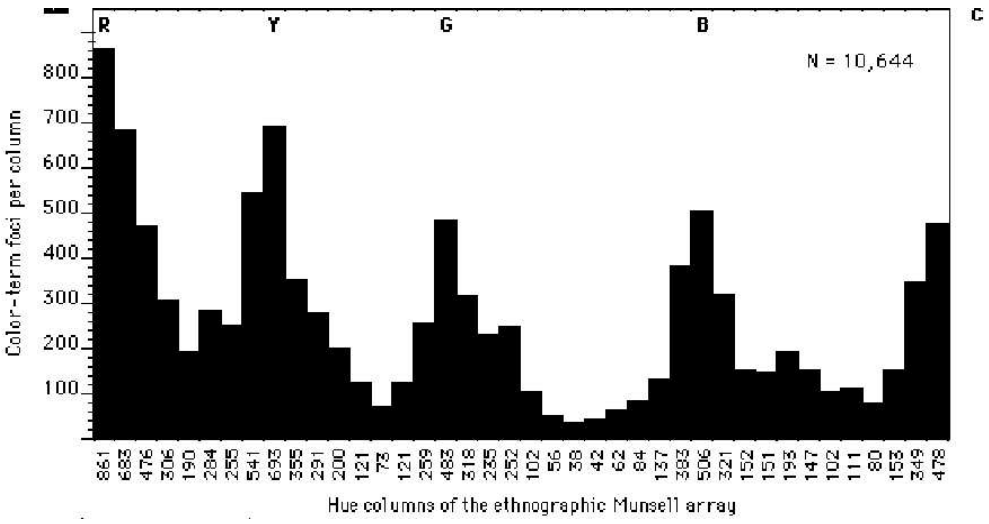


Figure 6.7. Frequency Distribution of 10,644 WCS Colour-term Foci across the Hue Columns of the Ethnographic Munsell Array. (Reproduced from MacLaury, 1997b, p 202.)

Perhaps part of the difference is due to different criteria being used for choosing which terms to include. MacLaury (1997b) included colour terms regardless of whether they were basic or non-basic, and whether they named ‘hue, brightness,

saturation or another quality of the light sense' (p202). MacLaury noted that 'the consequent noise decreases the percentage of terms that will be focused in reference to hues.' (p202), so this could explain why there is a greater number of prototypes on chips which do not correspond to unique hues, but it does not explain why those prototypes which are not on chips corresponding to unique hues are nonetheless mainly clustered close to those chips. A possible explanation for that, is that varying light conditions might lead to inconsistent recognition of unique hues, and we might also expect that different people's perceptual processes are not necessarily entirely consistent, even when they are presented with exactly the same stimuli. Either of these possibilities could lead to small, effectively random, displacements of where in the colour space informants located unique hues, and this would be expected to result in gradual peaks of the kind seen in Figure 6.7. The computer model did not simulate varying light conditions, and all the simulated people were identical in the way in which they conceptualized colour, so we would not expect to observe gradual peaks in the results of the simulations.

An alternative explanation of the differences between Figure 6.6 and Figure 6.7 is simply that the simulated unique hue points in the computer model were made too strong, and hence were too constraining. If this is the correct explanation of the differences between the two graphs, the problem could be solved by making the foci weaker. One other possibility is that unique hues do not give an especially salient status only to such narrow ranges of colour. Perhaps the effect of unique hues is spread out more gradually over a range of neighbouring chips, with the focus being strongest on the centremost one. There are clearly a number of plausible explanations of the differences between the results of the World Colour Survey and the simulations

using unique hues, but at present there does not seem to be any clear way to determine which, if any, is the correct one.

Another finding which seems to be implicit in the results of the World Colour Survey is that the unique hues tend to be evenly distributed between colour terms. For example, we would not generally expect to find a colour term containing two unique hues (for example, a red-yellow term) in a language which also had a colour term containing no unique hues (for example a purple term). We would normally expect to see a purple term only if a language had separate red and yellow terms. MacLaury (1997a) does however report that there is a considerable number of exceptions to this generalization, as purple and orange terms are sometimes seen even when there are also composite terms containing two unique hues[79]. Here I refer to colour term systems as *balanced* systems if the difference between the number of unique hue points in the term with the most and the term with the fewest is one or zero. Other colour term systems are called *unbalanced* systems. So, for example, a system with red, yellow, green, blue and purple terms would be balanced, because the unique hues are distributed as evenly as is possible in a five term system, four terms having one each, and the purple term having none. The issue which I wanted to investigate was whether simulating unique hue points would lead to an increase in the number of balanced systems.

---

[79] However, MacLaury does not give a precise figure concerning in exactly what proportion of languages such systems are found.

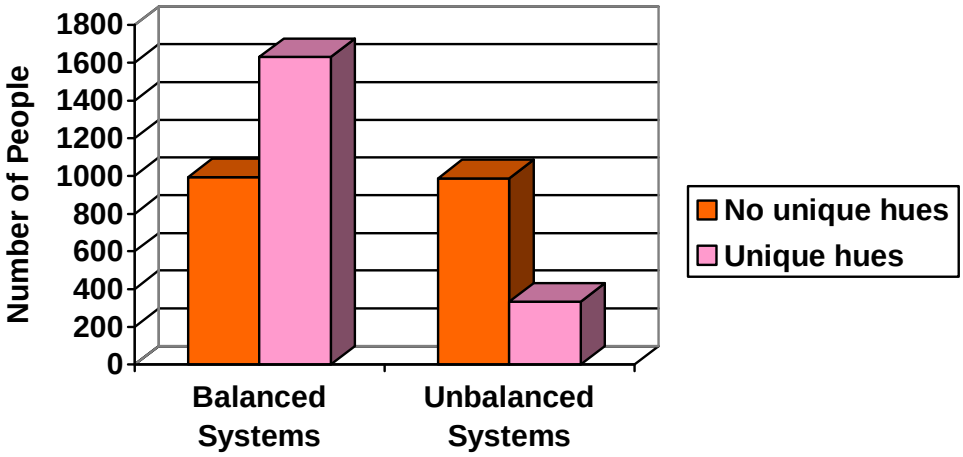


Figure 6.8. The Distribution of Unique Hue Points between Colour Terms.

Figure 6.8 shows the number of balanced and unbalanced systems in the simulations, both when unique hue points were made especially salient, and when they were not treated differently to other colours. For the simulations in which unique hues were not especially salient, the classification of systems as balanced or unbalanced was made based on the same locations for the unique hues as when the unique hues were especially salient. (People were again treated individually here, as in the analysis from which Figure 6.6 was produced, so the counts concern the number of individual people in the simulations with balanced colour terms.) We can see that, even when unique hue points were not simulated, half of all colour term systems were balanced. This result is to be expected, because we would generally expect the denotations of colour terms to be of roughly equal size, and so they will tend to spread out evenly throughout the colour space, so that there is a considerable chance that each will contain similar numbers of unique hues. However, because of the uneven distribution of unique hues, we should also expect to see a large number of unbalanced systems, and this is confirmed by Figure 6.8.

It was hard to predict the effect of adding simulated unique hue points to the model. However, Figure 6.8 reveals that it led to a big increase in the proportion of balanced systems. This is probably because a large proportion of examples which are remembered are now at unique hues, so that if a colour term does not contain a unique hue point, or if it includes fewer unique hue points than other colour terms, then the artificial people in the simulations may not observe enough examples of it for it to survive transmission between generations. Regardless of whether this explanation is correct or not, Figure 6.8 clearly shows that the simulation of unique hue points has had a major effect, beyond simply determining prototype locations.

Given the above results, we would probably expect that the simulations without unique hue points would diverge considerably from those with them, in terms of the frequencies with which colour terms of each type emerged. Figure 6.9 plots the percentage of terms of each type emerging in the simulations without unique hues, alongside the equivalent results for simulations with unique hue points and empirical data from the World Colour Survey, which was previously given in Figure 6.4. (All of this data ignores derived colour terms, though the frequencies of these, and of all the other types of colour terms, are given in Table 6.5.) We can see that, the removal of unique hues from the model has had little effect on the proportion of colour terms of each type which emerge. Simply because of the varying distances between unique hue points used when analyzing the data, some types of colour term are much more frequent than others. For example, there are far more green-blue terms than blue-red ones, because the green and blue unique hues are closer together than the blue and red ones.

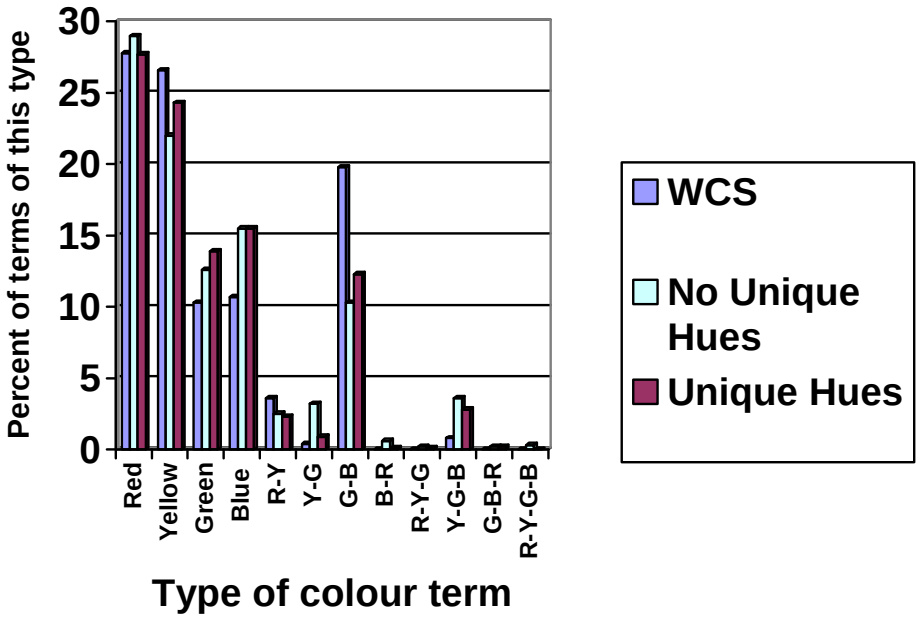


Figure 6.9. Percentage of Colour Terms of each Type.

A greater effect of unique hues can be seen when we look at the frequencies of the derived terms shown in Table 6.5, and compare them to the equivalent results when unique hue points were simulated, given in Table 6.2. The most obvious finding is that when the simulation of unique hues is removed, there are a far greater proportion of derived terms than in the previous simulations. This is presumably because in simulations with unique hue points, colour terms would first become established on these most salient points, and so derived terms tended to be seen only in languages with greater numbers of colour terms. Table 6.5 also shows that the simulations without unique hue points contain large numbers of derived colour terms of unattested types (lime and turquoise terms, of which there is a total 134 occurrences), while in the simulations with unique hues, there were only four such terms.

Type of Colour Term	Number of Terms in Simulations with no Unique Hues
Orange	374
Lime	118
Turquoise	16
Purple	644
Red	366
Yellow	278
Green	159
Blue	196
Red-Yellow	32
Yellow-Green	40
Green-Blue	130
Blue-Red	7
Red-Yellow-Green	3
Yellow-Green-Blue	45
Green-Blue-Red	2
Red-Yellow-Green-Blue	4

Table 6.5. The Frequency of Each Type of Colour Term in Simulations without Unique Hues.

Looking just at the proportion of colour terms of each type gives the impression that the unique hue points have had relatively little effect on the simulations, but when we look at overall colour term systems, we can see that this is not really the case. Table 6.6 lists the types of colour term system which emerged most often in the simulations without unique hue points (the full analysis of all the systems can be found in Appendix C). We can see firstly that many of these languages are inconsistent with the empirical data, because they contain several purple or orange terms (each of which usually denotes a separate part of the colour space). This result could have been predicted, given the extremely large number of derived terms in the simulations as a whole. However, we can also see that many of the languages lack any term which consistently names one or more of the unique hues. For example, the six purple, red, yellow-green systems are missing a term which consistently names the blue unique hue. (In five languages, no colour term could be found which consistently named any of the unique hues at all.) This is in contrast to the results in which unique hues were

simulated, where only 35 languages were missing terms for a unique hue (Table 6.4 on page 192). It seems that the unique hues had the effect of ensuring that, in the majority of languages, most speakers would know a colour term for each of the unique hues. When unique hues were not simulated, the whole colour space effectively became uniform, so colour terms would be just as likely to have unique hues at their boundaries as near their prototypes. This would lead to inconsistency concerning which colour terms would name each unique hue. This is because, if speakers disagreed slightly about a colour term's boundary, this could result in that term being given a different classification for each speaker.

Basic Colour Terms in Language	Number of Languages with this Type of Colour Term System	Consistent with Evolutionary Trajectories
Purple, Red, Yellow, Green-Blue	15	yes
Red, Yellow, Green-Blue	13	yes
Purple, Red-Yellow, Green-Blue	8	yes
Red, Yellow-Green-Blue	8	yes
Red-Yellow, Green-Blue	8	yes
Red, Blue, Yellow-Green	7	yes
Orange, Orange, Orange, Purple, Purple, Purple, Purple, Red, Yellow, Green. Blue	6	no
Orange, Orange, Lime, Purple, Purple, Purple, Purple, Red, Yellow, Green	6	no
Orange, Orange, Purple, Purple, Red, Yellow, Green, Blue	6	no
Orange, Purple, Red, Yellow, Green-Blue	6	yes
Purple, Red, Yellow-Green	6	no
Purple, Red, Yellow-Green-Blue	6	yes
Red, Yellow	6	no
No consistent colour terms	5	no
Orange, Red, Green-Blue	5	no

Table 6.6. The Most Common Types of Emergent Colour Term Systems in Simulations without Unique Hues.

Classification of Colour Term System	Number of Systems
Conforms to Evolutionary Sequence[80]	87
Contain Unattested Colour Term	122
No term consistently names one or more unique hues	148
More than one term can name one or more unique hues, or there is more than one of one type of derived terms	230
Total number of languages classified	415

Table 6.7. Classification of Emergent Languages.[81]

Table 6.7 classifies all the languages in the simulations without unique hues, and shows that only a fairly small proportion of them is consistent with the evolutionary trajectories. Many languages were problematic, because they contained lime or turquoise terms, or they contained composite terms of unattested types (for example, seven languages contained blue-red terms). A bigger problem was the lack of terms consistently naming unique hues, and the presence of extra terms, either extra orange or purple terms, as noted above, or simply multiple terms for one or more unique hues (for example, some languages had two yellow terms). Some common types of system were also extremely rare in these simulations. For example, there was only one red,

[80] Systems are included here so long as all their terms are of attested types, there is a term corresponding to each unique hue, and there are not multiple purple or orange terms. Some of these systems do not strictly belong on the evolutionary trajectories. For example, one system included here has red-yellow, green and blue terms, although we would not normally expect to find red-yellow composites with separate green and blue terms. Systems were classified in this way primarily so that Table 6.7 would be comparable with Table 6.4 (on page 192), which did not contain any such problematic cases. Also, many of the systems in Table 6.7 contained purple or orange terms (or both) as well as yellow-green-blue or red-yellow composites, and such systems are unattested, or at least extremely uncommon.

[81] Many systems are counted here more than once, because they diverge from the evolutionary trajectories in more than one of the ways listed in the table.

yellow, green, blue system, while in the simulations with unique hues this was the most common system, occurring 112 times (Table 6.3 on page 191). The general conclusion that we can draw from this data is that the addition of unique hue points to the model has had a significant effect on the emergent languages, in that it has ensured the emergence of terms naming each unique hue, and has prevented the emergence of unattested types of colour term. It has also prevented the emergence of such large numbers of derived terms, of both attested and unattested types.

Chapter 7

Adding Random Noise to the Evolutionary Model

Chapter 6 showed that the evolutionary model could account for much of the typological data, but there is one key aspect of the evolutionary simulations which appears to be unrealistic, and that is that the data from which the artificial people learn was completely free from noise. As was noted in section 3.2.3, inferring the intended referent of a word used by another person would seem to be a somewhat difficult task, and so it seems unlikely that this is always accomplished without error. Hence not all the data from which children learn colour words will be accurate, due to errors made by the children when observing other speakers. The original acquisitional model, which had a continuous colour space, and was described in Chapter 3, was designed to be able to cope with erroneous data. It was shown above (in section 3.8), that the model was able to learn even when as much as 80% of the data presented to it was random noise. We could expect that the new acquisitional model, with a discrete colour space, would also have this property. However, in the evolutionary simulations described in Chapter 6, no random noise or erroneous data was added to the data from which the artificial people learned.

The research described in this chapter was conducted to investigate whether coherent colour term vocabularies would emerge in the presence of large quantities of random noise, and, if so, whether the colour term systems would still reflect the typological patterns described by Kay and Maffi (1999). The same model was used as in Chapter 6, and unique hue points were simulated in the same way. However, 50% of the time, instead of the data from which an artificial person learned being produced by another artificial person, a completely random colour was paired with the colour word produced by the speaker.

The simulation was run 170 times[82], 10 times in each of 17 conditions, and again the number of accurate colour examples which each artificial person observed during their lifetime on average was varied between 18, 20, 22, 24, 25, 27, 30, 35, 40, 50, 60, 70, 80, 90 , 100, 110 and 120. However, in each case there would be, on average, one random example for each accurate one, constituting a level of random noise equal to 50%. Figure 7.1 shows some of the results of these simulations, together with those in which there was no random noise (repeated from Figure 6.2). The average number of basic colour terms emerging in each condition was measured as before (see section 6.3), and these averages were plotted on the graph together with the equivalent results from the simulations without random noise.

---

[82] The decisions to set the noise level at 50%, and to perform a total of 170 simulations were made fairly arbitrarily. The aim was simply to generate a level of noise that might well be expected to have a major impact on the results, and to perform enough simulations for the results to be reliable.

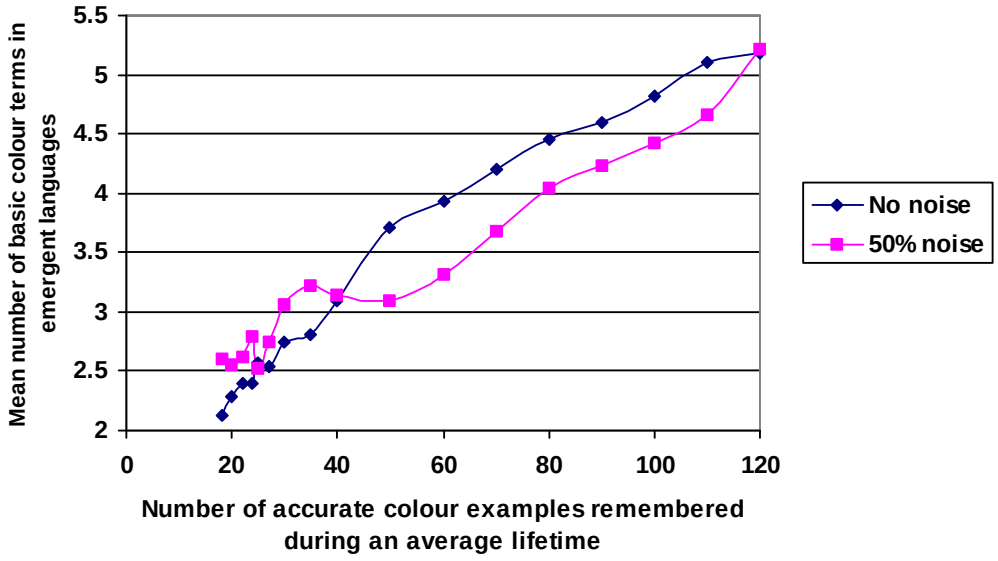


Figure 7.1. The Number of Basic Color Terms in Emergent Languages.

We can see that the number of colour words emerging in the languages is still, on average, roughly proportional to the average number of colour examples observed by the artificial people. Perhaps surprisingly, the number of words emerging seems to be dependent solely on the number of *accurate* examples of colour words which people observe during their lifetimes. Even though in the condition with 50% noise, twice as many examples were observed by each person as the people who observed the same number of accurate examples but no random noise, the number of colour terms emerging seems to be essentially the same in each condition. (The small differences between the no noise and 50% noise conditions can be attributed simply to random variation.) This result was obtained despite the models being identical in other respects, as each time the acquisition mechanism had an expectation that half of the examples would be random (p=0.5). We should remember here that this is despite the fact that the artificial people received no feedback for successful communication, so we cannot attribute this effect to them learning fewer colour words in order that they can learn each one better, and so communicate more accurately.

It seems that in the condition with 50% percent random noise, the simulations have performed in almost exactly the same way as when that noise was not present. This result seems somewhat counter intuitive, because no parameter was changed between the simulations which would have given the model any indication that there were varying amounts of random noise. (In both cases the parameter *p* was set at 0.5, so that 50% random noise would have been expected in *both* sets of simulations), and there was no indication given to any of the artificial people which would have allowed them to distinguish accurate from random examples.

The most important consideration, however, was whether the simulations would still reproduce the typological patterns, even when there was so much random noise. Figure 7.2 compares the proportions of basic colour terms which were classified as red, yellow, green, or blue, or as composites of these terms, in each condition of having no noise, or 50% noise, to the proportions of terms which were classed as being of each of these types in languages on the evolutionary trajectories in Kay and Maffi (1999).

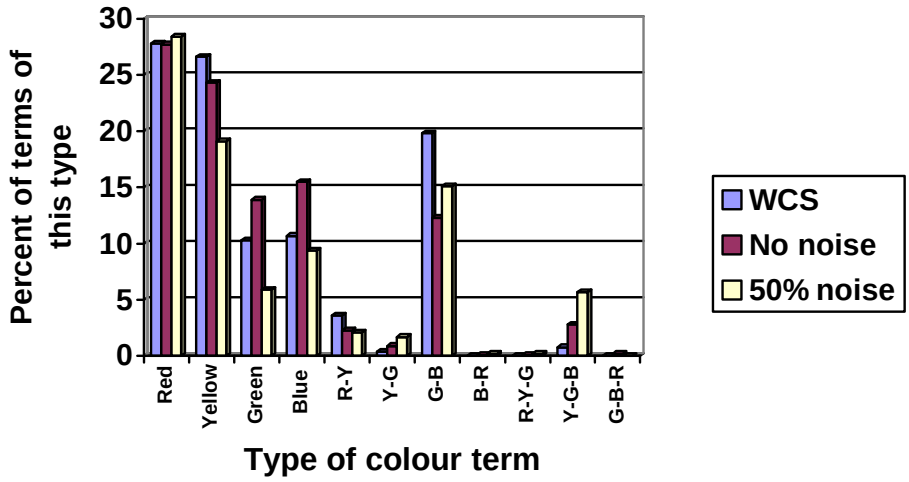


Figure 7.2. The Frequencies of Each Type of Colour Term.

We can see that the typological patterns in the relative frequencies of each type of colour term are still roughly reproduced in each condition. The only major differences between the condition with no noise, and that with noise, is that there are fewer green, blue and yellow terms when there is a high level of noise, and a greater proportion of yellow-green-blue terms. These differences might be due simply to random variation. However, if this result is not simply due to random variation, then it might be possible to make the results with 50% noise, more closely reflect those with no noise by altering the parameters controlling the location of the unique hue points. In any event, it is clear that in some ways adding noise has resulted in the simulations more closely reflecting the empirical data, especially in that the proportion of blue terms is now almost the same as that found in the World Colour Survey. (Although the simulations with noise in some ways mirror the World Colour Survey less closely than those without noise, for example, in that there are far too many yellow-green-blue terms emerging in the simulations with noise.)

We should also consider colour terms which do not contain a unique hue point. Remember that, in the noiseless condition, 76.9% of terms not containing a unique hue point were purple, while 19.2% of such terms were orange (Table 6.2). With 50% noise, these figures were 60.6% for purple, and 26.8% for orange. Hence in both conditions the empirical finding that orange is less common than purple was supported by the simulations.

The corresponding figures for lime and turquoise terms were 3.8% and 0% with no noise, and 9.9% and 0.3% with 50% noise. These results are consistent with the empirical data, in that in general basic lime and turquoise terms are not found in natural languages. Possibly the occurrence of basic lime terms is somewhat more frequent than should be expected, although it is possible that sometimes such terms are simply ignored in linguistic analyses, as there is a theoretically motivated expectation that they will not be basic, and so this is the classification which field linguists may be inclined to make.

The crucial conclusion to be drawn from these results, is that the addition of large quantities of random noise to the simulation, which ought to have made it more realistic, has not prevented the model from accounting for the empirical data. In fact, it has not radically affected the results, as compared to the noiseless condition, at all. This would seem a very desirable property for a model of language evolution, as it would be a very poor model of language which was unable to account for empirical data when attempts were made to reproduce conditions similar to those in which real language evolution takes place. Generally simulations using expression-induction models are carried out without the simulation of any noise at all, and it may be that the behaviour of many such models would change radically if a high level of noise

were added to the simulation[83]. Hence, the results of this chapter provide further support for the claim that the model simulates colour term evolution in a realistic way.

[83] One notable exception to this generalization is the model of Steels and Kaplan (1998). They showed that coherent languages could emerge among populations of artificial people in the presence of moderate amounts of random noise, but that when there were very high levels of noise, coherent languages would not emerge. The high level of noise in Steels and Kaplan's experiments was 70%, compared to the 50% level of noise used in the simulations of this chapter. However, there were considerable differences between Steels and Kaplan's model, and the model of colour term evolution, in terms of the types of meanings communicated, and the interactions between agents. Steels and Kaplan also simulated noise which disrupted the transmission of word forms as well as their meanings, while, in this chapter, word forms are always transmitted intact, and only the transmission of meanings is subject to noise. Hence, it is not clear that a meaningful comparison (in terms of resistance to noise) can be made between Steels and Kaplan's system, and the simulations of this chapter. However, Steels and Kaplan showed that, even if the level of noise was set at a level that would prevent coherent languages from emerging, if a coherent language already existed, then it could be maintained, despite the presence of the noise. This demonstrated that established languages were more resistant to noise than developing ones.

Chapter 8

Implications and Future Directions

The results presented above, especially those given in section 6.4, demonstrate that the typological patterns observable in basic colour term systems can be explained if it is assumed that the unique hues are not evenly spaced in the conceptual colour space, and that people remember the unique hues better than other colours. That unique hues are better remembered and more salient than other colours, is well supported by results obtained in psychological experiments, as is the hypothesis that the unique hues are not evenly spaced in the conceptual colour space. The unequal spacing of unique hues is exemplified by most colour order systems, such as the Munsell system (Cleland, 1937), as these systems do not place the unique hues at equidistant locations around the colour space[84]. However, establishing that the unique hues are not

[84] The natural colour system (Hård, Sivik and Tonnquist, 1996) is a notable exception to this generalisation. It does place the unique hues at evenly spaced points around the colour space, so that red and green are at opposite points on the hue circle, as are blue and yellow. However, it does not make any claim that this corresponds to a conceptual colour space.

conceptually equidistant, does not necessarily mean that the particular unique hues locations proposed in this thesis are correct.

The primary evidence for the locations chosen for the unique hues is the data contained in the World Colour Survey, but a question should be asked concerning just how closely that data constrains the locations. The set of parameter settings used was arrived at by a process of trial and error, in which hundreds of simulations were performed for each setting, and the resulting languages compared to the World Colour Survey. Based on each set of results, the parameter settings were adjusted to try to make the results more closely reflect the empirical data. However, it was not possible to exhaustively test every possible set of parameters, and so it could be possible that there exists a quite different set of locations for unique hues that would account equally well for the typological patterns.

A further problem arises in that the modelling of the colour space was not completely accurate, especially as the dimensions of saturation and lightness were neglected. We could expect that extending the model to incorporate these dimensions would affect the evolution all colour terms, and not just the emergence of black, white and brown terms, which cannot be learned by the existing model. If the model were changed, so that it used a full three dimensional colour space, the current settings might result in quite different typological patterns being apparent in the emergent languages. When using such a model, we might only be able to reproduce the patterns observed in the World Colour Survey data by changing the locations of the unique hues. (Conceivably, it might not be possible to replicate the patterns at all when using such a model, which would raise serious questions concerning the correctness of the present model). Any change in the way in which the model learns, or in how the

unique hues are modelled, could well affect the results of the simulations in unpredictable ways. Hence it is difficult to predict what results such a model would produce, without actually implementing the model.

It is important to note that the kind of data which the results of the simulations should be compared with, is data which gives an accurate picture of how languages in general tend to evolve. This raises the issue of how accurate a picture the results of the World Colour Survey which were used in this thesis give of the ways in which languages tend to evolve. It is not clear that the languages sampled for the World Colour Survey do in reality accurately reflect the frequency with which colour terms and colour term systems of each type tend to emerge. Because many of the languages in the World Colour Survey were probably closely related (either genetically, in the sense of being descended from a common ancestor, or simply because there was contact between speakers of the languages), some of the similarities between languages might have been due to their relatedness. This might lead to some types of colour term appearing to be much more common than they would have been if all the languages in the survey were unrelated. (For example, if a type of colour term only existed in a few related languages, but the World Colour Survey included data from several of those languages, it would appear to be common, when in fact it was not.) Factors such as these could lead to considerable discrepancies in the patterns apparent in the typological evidence, which would influence our evaluation of how accurately the model replicated the empirical data. Hence, it seems that we should be cautious about making any assumption that the World Colour Survey provides a completely accurate picture of the quantitative facts concerning colour terms. However, the World Colour Survey does contain data from a very diverse range of languages, and it is certainly by far the most comprehensive data source available for studies of colour

typology. Hence, it would seem that the World Colour Survey almost certainly gives a good indication of the approximate frequencies of at least the most common types of colour term, even if factors such as biased sampling and language relatedness distort the results to some extent.

Another potential objection to the results of the World Colour Survey might be that they rely on the Munsell colour system, and that if another colour order system had been used, quite different results might have been obtained. Clearly, if we regard the spacing between colours in colour order systems as being to some extent arbitrary (a position which was argued for in section 2.1), the number of colour chips denoted by each colour term will be equally arbitrary. However, the typological analyses of section 6.4 did not take account of the size of the denotations of any colour terms, but simply classified colour terms based on the locations of the unique hues relative to the terms' denotations. As which colour chips in the Munsell colour system correspond to the unique hues is not in doubt, it is unproblematic to equate these colour chips with the unique hue points in the model. Hence, so long as the analysis of the data is based on unique hue locations, and no account is taken of the actual sizes or locations of the boundaries of colour term denotations, the results should not be affected by the colour order system used in obtaining the empirical data.

Given all of the above issues, it seems that we should be cautious about what claims are made concerning the unique hue locations. Probably, they should be interpreted simply as being indicative of the relative distances between unique hues in the conceptual colour space. It seems likely that in reality the relative distances between unique hues are ordered from green and blue being closest, followed by green and yellow, then yellow and red, and with blue and red being the furthest apart. However,

making a more precise claim about the locations of unique hues would be unjustified, and it is not possible to be certain even about this ordering of relative distances.

A more cautious interpretation of the results would be to conclude that the models have shown how learning biases can affect how languages evolve, and that such learning biases could result in the emergence of a range of languages which collectively mirror typological patterns, but not to make any more specific claims. The models showed how human languages could be understood as a product of innate psychological properties acting in combination with cultural pressures, and that typological patterns may only be explainable as the result of evolutionary processes occurring over several generations of speakers. This is clearly an interesting result, because it is at odds with some previous analyses. For example, it has often been suggested that implicational hierarchies exist because they reflect restrictions on the range of languages which people are able to learn, due to restrictions imposed by Universal Grammar (see, for example, Travis,1989). It has also been proposed that implicational hierarchies are due to functional pressures arising from speakers' communicative needs (see Hawkins, 1988, for discussion). The colour term models correspond to neither of these possibilities, but instead explain an implicational hierarchy as the product of both acquisitional and functional pressures.

Other researchers have also tried to relate colour term typology to human physiology. One such approach was Kay and McDaniel (1978), outlined above, but this was not the first such study. Ratliffe (1976) proposed a psychophysiological basis for universal colour terms, noting the three pairs of opponent colours, black and white, red and green, and yellow and blue. They argued that the black-white opponency was the strongest, hence explaining why two colour term languages divide the colour

space up roughly into dark and light colours. Furthermore, they argued that there are neurophysiological factors which cause our discrimination of blue to be particularly weak, and of red to be particularly strong. This could explain why red terms emerge before green and yellow ones, which in turn emerge before blue ones. However, Ratliffe failed to explain why we see only a limited number of types of colours term which do not contain unique hues (for example, why we do not see light green or turquoise basic colour terms), although they did note that brown, pink, purple and orange are all adjacent to red, hence suggesting that the existence of these terms may be related to the special status of red. Ratliffe's proposal has many similarities to my theory, but it leaves unspecified some intermediate steps necessary to explain how neurophysiological effects, giving special status to unique hues, come to affect colour vocabulary.

While the evolutionary model attempts to explain colour term systems in terms of universal properties of the human visual system, attempts have been made to explain colour term typology in different ways. Foley (1997) presented the relativist position, which claims that consideration of wider cultural practices is necessary to gain an understanding of colour terms. Foley stated that 'culture must be the crucial autonomous intermediary between any innate and hence universal neurological perception of colour stimuli and the cognitive understanding of these.' (p160). As noted above, relativists such as Saunders (1992) have stressed that colour terms are not easily isolatable pieces of reality, but are 'culturally constructed' and are linked to cultures' 'meaningful practices' (Foley, 1997, p161). The evolutionary model of this thesis shows how colour vocabularies could emerge as a result of interactions between speakers, but it does so without modelling any wider cultural meanings of colour

terms, suggesting that considering such factors are not necessary for explaining colour term typology.

It has also been proposed that the presence of particular colour stimuli in the environment might have an effect on colour terminology. For example the universal salience of red could be attributed to the universal red colour of blood, something which is constant across cultures. Certainly such factors might well influence colour term systems, but, because the evolutionary model was able to explain typological patterns without reference to such environmental factors, it suggests that they may not play an important role in shaping colour term systems. While such cultural and environmental factors could well affect the range of colour term systems attested cross-linguistically, the present model provides no evidence to support the view that they do.

MacLaury (1997a, 1999) has tried to account for colour term typology using a theory called *vantage theory*, which he has developed based on colour term data, but which he has proposed can be applied as a general theory of category formation. Like the evolutionary colour term model, MacLaury's theory rests on the premise that 'green and blue appear more similar to each other than do red and yellow, yellow and green, or red and blue.' (MacLaury, 1997a, p87). However, MacLaury did not provide a computer implementation of his model, and so it is difficult to be sure that the typological patterns are derivable from the axioms of his theory, and that the theory could not equally well be interpreted in another way to account for a different set of data.

MacLaury's model is based on the premise that colour categories are formed on an analogy with space and motion. His theory relies on the existence of six 'elemental

sensations', corresponding to the four unique hues, plus black and white. He proposed that colour categories are then formed by 'analogy to the fixed and mobile coordinates by which people make sense of their positions in space-time' (MacLaury, 1997a, p380). He claimed that people form colour categories by focussing on one of the elemental sensations, and adjusting the degree to which they see it as similar or different from other colours. MacLaury developed a formal notation for describing colour categories, and he suggested that this system constrains the forms of emergent categories.

As an example, the simplest kind of category corresponds to a word such as English *red*, which is defined using three 'coordinates': a 'fixed' image of elemental red, and 'mobile' emphases on similarity and difference. The category is then constructed by relating these coordinates. Firstly, elemental red and the similarity coordinate are related, so that reds other than elemental red are still seen as *red*, because of the emphasis on similarity. We can also 'zoom in' so that similarity becomes a fixed coordinate, which is then related to the difference coordinate. This second relation curtails the extent of the categories range, preventing it from extending indefinitely, because colours further from the similarity coordinate are now increasingly seen as not belonging to the category.

The theory draws on concepts such as the figure-ground distinction which are well established in cognitive science (Ungerer and Schmid, 1996). However, terms such as 'coordinate' are not used in any sense with which I am familiar, and the justifications for the design of the theory are unclear. While vantage theory is a cognitive theory, and so presumably describes processes which are hypothesized to occur in people's brains, there is no attempt to specify exactly what the cognitive processes are at a

computational level, but instead the theory is described in terms of much more general and vague concepts.

The categorization of red using vantage theory, described above, may seem somewhat intuitive, but I find it hard to find much justification for the description of some other categories such as Hungarian *vörös*, which MacLaury classifies as a 'recessive vantage'. Both this term and *piros* name red in Hungarian, but *vörös* generally names a narrower range of colours, and has a darker red as its prototype. The category underlying *piros* is constructed in the same way as English *red*, but *vörös* is constructed differently, though using the same coordinates. Firstly, elemental red is related to the difference coordinate, and then, in the second step, the difference coordinate is related to the similarity one. The difference coordinate is always further from elemental red than the similarity coordinate. So in this case, the second relation between coordinates involves a move back towards elemental red. MacLaury argues that because this category includes the difference coordinate twice, it narrows the range of *vörös*, because there is now more emphasis on difference than on similarity. It is, however, unclear exactly why we should expect categories to be formed in this way, and I am inclined to question whether the mapping from category structure to the denotational range of the colour term can really be completely explained simply in terms of the coordinates, and the relations between them, proposed by MacLaury. Because of the complexity of MacLaury's theory, it is very difficult to evaluate it objectively, but it seems that we should only accept his theory if the relevant data cannot be accounted for by a theory which makes fewer assumptions.

Categories which have overlapping denotations, such as Hungarian *piros* and *vörös*, where one category appears to be dominant and the other recessive, play a prominent

role in MacLaury's arguments in support of vantage theory. (These terms constitute an example of the phenomenon which MacLaury calls *coextension*, which was mentioned above in section 2.1.) However, this general kind of pattern was seen in languages emerging in the evolutionary simulations, simply because the boundaries of each colour term are not always clear cut, and sometimes a speaker will hear more examples of one term as opposed to another. Hence terms which could be analysed as being coextensive in MacLaury's terminology, appear and disappear during the simulations simply as chance phenomena. This suggests that MacLaury *may* be over-explaining his data, by proposing specific mechanisms to explain what is in reality random variation. As far as I am aware, no other colour term researchers consider coextension to be a systematic occurrence. (Kay, Berlin, Maffi and Merrifield (1997, p35) say they will discuss 'the prevalence in the data (or lack thereof) of the phenomenon of coextension' in a forthcoming book.) Hence, it may not be necessary to account for the kind of co-extensive phenomena which MacLaury has exemplified using terms such as Hungarian *piros* and *vörös*.

Above it was noted that some researchers, such as Taylor (1989), have claimed that colour terms (and other words with prototype properties) have underlying prototype structures (Taylor, 1989). However, this is problematic because the prototype approach seems to be insufficient to account for certain properties of colour terms. Roberson et al (2000, p 395) have also argued that 'Berinmo [colour] categories have not formed around prototypes as, for the most part, there is little agreement about best examples.', and have instead suggested that 'color categories are developed from demarcation at boundaries'. Lammens (1994) created a computer model which was able to represent the meanings of colour terms using prototype representations. This required that the prototype structure be made completely explicit, hence making it

easier to evaluate whether colour terms are actually represented in this way. Lammens' model defined colour categories by specifying the location of the category's prototype in the colour space, and the size of the category. The degree of membership of a colour in the category depended on its distance from the prototype, and the size of the category was determined by a numeric parameter specifying how rapidly the degree of membership in the category decreased as the distance from the focus increased.

Lammens proposed that his model could form the basis of an account of how the meanings of colour terms are learned. However, it seems that in this respect there are some problems with his model. These problems have implications regarding whether people represent colour categories using prototype structures. Lammens' model does not specify how a language learner can establish the foci of colour categories. Instead it was assumed that these must be fixed, presumably innately (p. 143), and learning would then consist just of determining the extent of the colour category. Hence, during learning, only the parameter controlling the size of the category would be adjusted, and the observed data could neither affect the location of the category's prototype, nor the category's shape. This seems somewhat problematic because, as mentioned above, 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 the most problematic aspects of the prototype approach to linguistic categories. It seems that prototype theory cannot account for categories where peripheral members may be further from the prototype in one direction than in others, and that it usually relies on the existence of pre-linguistic natural prototypes. Given these problems, and that the models presented in this thesis

learned categories with prototype properties, despite those categories not being based on prototype structures, it seems best to assume that prototype effects are a by-product of learning mechanisms, and are not due to underlying prototype representations.

Levinson (2001) made an important observation concerning the theories of Kay and Maffi (1999) and Berlin and Kay (1969), who proposed that the typological patterns in colour term systems are due to an evolutionary progression in which languages gradually add basic colour terms, but never lose them. This was based partly on age-stratal data, which in some cases showed that older speakers of a language used fewer colour terms than younger ones[85], and is partly supported by historical textual data, or

---

[85] We should note that this is the opposite situation to that occurring in the model, in which younger speakers tended to know fewer terms than older ones (see section 6.2). This was because younger speakers had not always had sufficient time to learn all the colour terms in the language spoken in their community. We would not expect to find this effect in empirical investigations, because normally only adult informants would be used, or at least informants who were old enough to have learned all the basic colour terms in their language. However, the model cannot explain how we can get a situation in which older speakers know fewer terms than younger ones. It would seem that this situation occurs because people tend not to add new colour terms to their own I-language after they reach a certain age. Hence, if new colour terms enter the language, they will only be acquired by younger people, which would result in younger people knowing more colour terms than older people. This situation could be replicated in the evolutionary models, if the artificial people were modified so that, after a certain age, they no longer remembered new examples, but simply spoke based on the language they had learned up until that point. If the lifespan of agents (measured in terms of how often colour terms are used during a lifetime), was then increased over the course of a simulation run, we would expect that younger speakers (excluding the very youngest speakers who would not have had sufficient time to learn all the colour words), would know more colour terms than older ones. However, such a situation has not yet been simulated, and so it remains a topic for further research.

reconstruction of earlier states of the languages. However, in the majority of cases, the only evidence we have concerning the colour term systems of languages is data obtained from informants concerning the present form of the language. Levinson noted that it is a more usual practice in typology simply to form an implicational hierarchy, as in Figure 2.1 (on page 16 above), rather than to argue that languages progress from one type to another in a predictable order.

While I am not aware of any evidence showing that a language has ever lost a basic colour term, it does not seem clear what form that evidence would take[86]. We can often infer that colour terms are relatively recent additions to a language if they are also the name for some object with the same colour (for example English *orange*), or if they have been borrowed from another language (for example Japanese *buruu*, derived from English *blue*). However, we clearly could not use such evidence if a basic colour term was lost from a language; it would either just cease to exist, or, perhaps more likely, would be retained as a non-basic colour term, with more limited application. So it does not seem that there is clear evidence showing that languages do not lose basic colour terms, just an absence of evidence showing that this has ever occurred.

Above (in section 6.3 and Chapter 7) it was shown that the evolutionary model demonstrated a positive correlation between the number of colour words in a language, and how often people used colour words, which was equated with the level

---

[86] This is true at least for the majority of languages, for which there is not a long written record. For the handful of languages for which we have written records covering a long period of time, we could expect such evidence to be available.

of technological development of those people's societies. Hence, at a point in history when the level of technological development of a society was rising, we would expect that it would be much more likely for that society's language to gain basic colour terms than for it to lose them. As the communities in which most, if not all, languages are spoken, are probably at present rapidly increasing in their level of technological development, we would expect to find evidence of a rise in the number of basic colour terms, and this might appear to be a universal tendency.

If, however, the level of technological development of a community were to remain constant, then we would expect the number of basic colour terms to also remain fairly constant. However, as shown by Table 6.1 (on page 177), the correlation between the number of basic colour terms, and how often these are used, is not perfect, and so the number of basic colour terms in a language is not completely predictable. Hence, in situations in which there was no change in the frequency with which colour terms were used, we might expect to observe some random drift in colour term systems, with basic colour terms occasionally being lost or new ones being gained. This hypothesis is supported by experiments conducted using the evolutionary model. If the frequency with which colour terms are used by people in the community is kept constant, then, once an initial colour term system has been established, occasionally basic colour terms are lost or gained, but the average number of colour terms present during a long range of time in the simulation, remains fairly constant.

Hence we might conclude that Berlin and Kay's (1969) hypothesis of a unidirectional movement towards languages with increasing numbers of basic colour terms is a product of the situation which communities throughout the world are in today, rather than a fundamental property of human languages. At points in human history, it

would seem likely that many communities remained at more or less constant levels of technological development for long periods, and at such times they would be as likely to lose basic colour terms as to gain them.

It might seem obvious that the meanings of colour terms can only be acquired through observing examples of colours which they can denote, and that in the absence of such data it would not be possible to learn their meanings. However, Landau and Gleitman (1985) studied the acquisition of colour terms by a blind child, and found that her knowledge of colour words was in many respects similar to that of sighted children of the same age. For example, she knew 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 she herself could not identify. This clearly demonstrates that many aspects of the meaning of colour terms can be acquired even if no examples of their denotations are available. The child studied by Landau and Gleitman must have learned what she did about colour terms simply from the context in which the words were used, though clearly the most central aspect of the meaning of colour terms, that is, exactly which colours they denote, could not be acquired in such circumstances.

The evidence from Landau and Gleitman's study does, however, suggest that contextual and morpho-syntactic cues may be an additional source of evidence used by children when determining the meaning of colour terms. Such cues were not used by the colour models described above, but could potentially be incorporated into new versions of the models. This could be of particular benefit in addressing some of the issues concerning colour term acquisition not addressed by the present models, including how children come to identify what kind of property colour terms denote. Landau and Gleitman reported that this seems to be a difficult task for children,

noting that colour terms seem to be acquired later than adjectives denoting most other kinds of property, and that, at a relatively late stage of acquisition, children appear not to understand that colour words denote colour. This was not a problem for the computer models, as they were presented only with data concerning hue, and could not consider any possibilities other than that the words they were learning denoted a range of hues. However, in the real world, determining what kind of property a word denotes would seem to be a very difficult task (probably more difficult than learning the specifics of the denotations once the problem of the relevant domain has been solved). If children could use cues from linguistic context to identify the set of colour words, then the task of identifying their domain of reference would seem to be much easier, as once we had identified that one of those words denoted colour, it would be a relatively easy step to proceed to the conclusion that all of them do. We should note that Redington, Chater and Finch (1998) created a program which was largely able to single out the set of English basic colour terms, based simply on their locations in sentences relative to other words[87], and so it would seem likely that children use such a mechanism to identify the set of colour words, before they begin to learn their meanings.

While this thesis has shown that a Bayesian acquisitional model can account well for the data concerning colour term typology, the evidence concerning the correctness of the acquisitional model is somewhat indirect. It is supported because it explains the phenomena of prototype properties, and the typological properties of colour term systems, and because it is similar to other cognitive models that have been well

---

[87] Redington et al's program is discussed in greater detail in Chapter 9.

supported by empirical data. It would, however, be desirable to try to investigate more directly how closely the behaviour of the acquisitional model mirrors humans' behaviour. One way to do this would be to perform psycholinguistic experiments, in which participants were taught artificial colour terms. At the beginning of such experiments, subjects would be told that they were going to be taught colour terms in a language that was completely unrelated to English. They would then be shown several different colour chips, and for each one would be told the word for it in the artificial language. The number of colour chips that participants would be shown for each word would be varied, as would the hues of the colour chips that were shown. The participants would then be shown colour chips which they had not seen before, and asked to name each one with the appropriate word from the language that they had just been taught. The validity of the model could then be judged by seeing if it would have used the same colour term to name each of the colour chips, had it been presented with the same training data as the human participants.

By choosing appropriate training and test colour chips, it would be possible to investigate specific predictions of the Bayesian model, such as that if a person observes a large number of examples chips for a colour term within a particular range, then they will be less likely to use that term to name colours slightly outside that range, than if they had not seen so many examples, or if the examples had been spread over a larger range. The more data of this type that is collected, the better we will be able to determine whether the Bayesian model is accurate. However, a problem with such a technique is that there might be considerable interference from English colour terms (or the colour terms of any other languages that the participants knew), but this could be compensated for by using a range of participants, some of whom did not speak English, or any other language with a colour term system similar to that of

English. Ultimately this kind of experiment ought to be able to resolve the issue of exactly what mechanism children do use to learn colour term denotations.

Above it was argued that partition is not a universal property of colour term systems, and that it most likely that children do not apply this principle when they learn languages. Therefore partition was not incorporated into the acquisitional model as an *a priori* principle. However, almost all colour term systems do partition the colour space, so it is worth considering how a different version of the acquisitional model could be created that would make use of the principle of partition. The simplest way in which the model could be extended to make use of the partition principle, would be to continue to learn each term separately, but to use the positive examples concerning the denotations of the other terms as a form of negative evidence. If the partition principle applies, then the range of colours denoted by the terms cannot overlap, therefore evidence that a particular hue is denoted by one colour term, is also evidence that it is not denoted by any other colour term. It would be relatively straightforward to extend the present model of colour terms to take into account such evidence.

However, a fuller implementation of the partition principle would be to try to learn the denotations for all the colour terms together, and to find the partition of the colour space as a whole that best fitted the data overall. If there are *N* basic colour terms in a language, then *N* boundaries will divide the colour space into the required *N* categories, so that their is one colour category per colour term. If we link one of the colour terms to each of the boundaries, we can then take the denotation of each colour term to be all those hues starting from the location of the boundary, up to the location of the next boundary. We can then generate the full range of possible colour term

systems by allowing each of the boundaries to be at any point in the colour space. Because this will allow the boundaries to appear in any order, it also allows for all the possible ordering of the colour terms. This therefore can define the space of all possible hypotheses concerning the colour term denotations, and we can proceed using a method very similar to that used for the present model. A hypothesis will now specify the denotations for all the colour terms in the system, and hence will predict that accurate examples will all come within the denotation of the appropriate colour term. Any other examples must be treated as erroneous. However, if a continuous colour space were to be used with such a model, the derivation of the integrals required for use in hypothesis averaging would become more complex, and would be increasingly so the more colour terms were in the language. Similarly, if a discrete colour space were used, the time complexity of the problem would increase with the number of colour terms, so that for larger numbers of colour terms, it might not be possible to calculate a solution in a reasonable amount of time.

As noted above, an extension to the model which would clearly be very desirable, would be to extend the colour space used so that it corresponded to the full three-dimensional colour space, based on the three dimensions of hue, saturation and lightness. This would allow the denotations of colour words such as *brown, black, pink, grey* and *white* to be learned. There is no reason, in principle, why this could not be done, but it would make the acquisitional model more complex. In the present model, colour term denotations are represented simply as linear sections of the colour space, but, if a full three dimensional colour space were used, then colour term denotations would correspond to three dimensional volumes of the colour space. This would raise additional problems, because we would now have to specify a probability distribution over the possible shapes of the denotations, not just over their size and

location, as with the present model. This problem would probably not be insurmountable if some simplifying assumptions were made, such as proposing that people *a priori* assume that colour term denotations are roughly round. If possible colour term denotations were then modelled with a distribution over, say, ellipsoids, we could then expect good results to be obtained, so long as each colour term's denotation could be approximated reasonably accurately with an ellipsoid. From a Bayesian perspective, it would seem that people must in reality make some such assumption if learning is to be possible at all, as otherwise a very small but oddly shaped denotation could be found which would account well for any set of observed data.

We should note that, in the present model, colour term denotations must be contiguous, but that is simply an assumption, similar to the proposed restriction on shapes in a three dimensional version of the model. Gärdenfors (2000) has independently proposed that people have a tendency to prefer to partition conceptual spaces in sensible ways, using generally concave shapes, and avoiding demarcating regions with complex boundaries. A problem with extending the model so that it worked in three dimensions might be that it would almost certainly increase the time complexity of the simulations. This might limit the results that could be obtained with the model, but empirical investigations would be needed to determine how significant a problem this would be.

Lammens (1994) and Belpaeme (2002) both used three dimensional colour spaces in their computer models, but at least one previous model has used a one dimensional colour space. The representation used by Kay and McDaniel (1978), who proposed that colour term denotations corresponded to fuzzy sets derived from fuzzy logic

operations on the outputs of opponent process cells, was also one dimensional. Its representations were of almost the same form as those used by the models of this thesis. Cells opposing red and green or blue and yellow would be used to derive membership in fuzzy sets for each point along the hue dimension, and it was proposed that similar cells opposing black and white could be used to derive membership in fuzzy sets along the lightness dimension. They proposed that these two dimensions could be combined to account for colour terms which must be defined both in terms of lightness and hue (such as *pink*, which can be analysed as a light red), but no specific fuzzy set operation was provided for deriving fuzzy set representations for these terms (Kay and McDaniel, 1978, p637, footnote). Also, the dimensions of lightness and saturation are also relevant to the definition of terms containing a unique hue point, such as *green*, as they are needed to distinguish these terms from white, black and grey, but Kay and McDaniel did not address this issue. Hence representations of colour terms were obtained partly by extrapolating from the fuzzy logic model in much the same way as it might be proposed that the present model could account for such terms if it were extended to include the lightness and saturation dimensions. However, neither model in itself truly models the three dimensional colour space.

One solution to the problem of how to account for colour term typology in a model which can account for all the colour terms, and not just those terms defined by hue, would be to modify the model of Belpaeme (2002). If Belpaeme's model were modified, so that it gave a special status to the unique hues, in a similar way to the model described in this thesis, then it might well be able to account for the typological data. Conducting such experiments would also help to determine to what extent the

results obtained here rely on the specifics of the model, and to what extent they might be replicable by other models which learn in somewhat different ways.

While the work of this thesis has attempted to explain typological data in one domain, that of colour term systems, the general methodology used, namely expression-induction modelling, should be applicable to explaining other aspects of linguistic typology. Some of these have already been mentioned above, for example de Boer's (1999) work on vowel typology. However, typological patterns have been identified in many other semantic domains besides colour, and in most cases no attempt has yet been made to explain these patterns using the expression-induction methodology. One domain that has recently attracted a lot of interest is that of spatial relations (Levinson, 2003a, b). All languages have ways of expressing spatial relations between objects, such as that one object is on top of, or above, another. However, the exact distinctions which are lexicalised vary between languages, so that while English makes a distinction between something being *on* something else and something being *over* something else, Japanese conflates this distinction, and uses the same word, *ue*, to express both relations. Which relations are conflated in particular languages is however far from random, as only some relations are ever conflated, and if it is known that certain relations are conflated, then it is sometimes possible to predict which other ones will also be merged.

It would seem likely that this data concerning spatial relations might be explainable using an expression-induction model. First it would be necessary to formalise a representation for the position of, and relation between, objects in space. Then an acquisitional model could be created that would try to infer the meaning of words based on a number of examples of specific relations that could be expressed with the

word. (These relations would be represented using the spatial formalism.). An expression-induction model could then be constructed, in which the simulated people would express spatial relations using words they had learned by observing the utterances of other artificial people. If such simulations were performed many times, and the words expressing spatial relations which emerged in each simulation were analysed, we might well see a replication of the typological patterns reported in the literature. The emerging languages would probably be shaped primarily by the acquisition mechanism used, but the natural properties of space that would be encoded in the formal representation of spatial relations could also potentially influence the results. This is just one example of the many areas in which expression-induction modelling might be applied in the future.

Generally, then, it can be concluded that the models of this thesis have provided plausible accounts, both of how colour term denotations can be acquired, and of why we see typological patterns in colour term systems cross-linguistically. The prototype properties of colour term systems are emergent properties of the Bayesian learning mechanism used, while the typological patterns are emergent properties of the cultural evolution of colour term systems over time. This generally supports Berlin and Kay's (1969) hypothesis of an evolutionary trajectory, but the models do not support the claim that as languages evolve they only gain colour terms and never lose them. The evolutionary model ties together observations concerning colour psychophysics, neurophysiology and cross-linguistic typology, with a degree of explicitness that has not been achieved by any other theory. The major weaknesses of the model are that it is restricted to modelling only the hue space, and that it is not possible to be sure whether the parameters in the model are accurate, or whether similar results could be obtained with a significantly different model. The next chapter moves on to try to

explain linguistic data in another domain which has been of great interest to theoretically oriented linguists. The specific question addressed concerns the learnability of the dative alternation in English, and it is shown how Bayesian inference can be used to solve this problem as well.

Chapter 9

Bayesian Acquisition of Syntax

This chapter describes work which has investigated how children learn their first language and, in particular, the syntactic system of that language, based on observations of the speech of other people. It conceives of the problem in the following way: when exposed to utterances in that language, how is it possible to infer the grammatical system which produced those utterances. Further, the learner is assumed not to know the meanings of the words, have access to prosodic cues to structure, or to receive feedback about which sentences are not grammatical.

Currently the major paradigm within which language acquisition is explained is probably the parameter setting framework (Chomsky, 1995; Belletti and Rizzi, 2002; Chomsky, 2002). Within this framework, it is proposed that knowledge of language is largely specified innately, and learning consists of identifying word tokens and setting a limited number of parameters according to the syntactic structures to which the child is exposed. Chomsky argues that this position is necessary because ‘even the most superficial look reveals the chasm that separates the knowledge of the language user from the data of experience.’ (Chomsky, 1995, p. 5).

Gold (1967) investigated this problem more formally, and proved that, without negative evidence (explicit information about which sentences are ungrammatical),

languages are not ‘learnable in the limit’, unless the class of languages which the learner may consider, is restricted *a priori*, for example by innate knowledge. Below I will discuss an alternative result by Feldman, Gips, Horning and Reder (1969) which suggests that Gold’s result is not relevant to the circumstances under which children learn languages.

Redington, Chater and Finch (1998) investigated to what extent syntactic categories could be inferred, based on distributions alone, without knowing *a priori* what syntactic categories existed in the language. They formed vectors by taking the two preceding and two following context words for each occurrence of each target word in a large corpus of transcribed speech, and recorded how often each context word occurred in each position. Only the 150 most frequent words were used as context, and so this resulted in 600 dimensional vectors for each word (there being one entry for each of the 150 context words in each of four positions). Clustering those words whose vectors were most similar, in terms of Spearman’s rank correlation, resulted in clusters which corresponded to appropriate word classes for most of the 1,000 target words. While this system was good, in that it could be applied to naturally occurring speech, it was necessary to decide at what level of dissimilarity to form separate classes, and so it does not completely solve the problem of recovering the syntactic classes used by the original speakers.

Elman (1993) demonstrated that not only word classes, but also syntactic patterns in which words belonging to those classes appeared, could be learned without much innate syntactic knowledge, at least for simple languages. He trained a recurrent neural network to predict the following word in artificially generated sentences conforming to a simple syntactic system containing 23 words, and syntactic features

such as number agreement and recursion in relative clauses. Once trained on 50,000 sentences in this simple language, the network performed at near optimum accuracy at predicting the subsequent word at any stage in a sentence, showing that the network had internalized the structural constraints implicit in the data. Lewis and Elman (2001) have extended this work by training a network to learn an artificial language based on utterances occurring in the CHILDES corpora (MacWhinney, 2000). They concluded that 'the stochastic information in data uncontroversially available to children is sufficient to allow for learning.' (p. 360).

While Redington et al (1998), and Elman (1993) and Lewis and Elman (2001) demonstrate that much of syntactic structure can be learned by making statistical inferences based on the distributions of words, Pinker (1989) suggests that some aspects of syntax cannot be learned in this way. He proposes that, in order to determine verbs' subcategorizations in the absence of negative evidence, children must rely on complex innate rules combined with knowledge of the verbs' semantic representations.

Verbs such as *give* can appear in both the prepositional dative construction, as in (9.1a) below, and the double object dative construction (9.1b), but there is a class of verbs such as *donate* which can only appear in the prepositional dative construction, (9.1c and 9.1d). However Gropen et al (1989) observe that, based on the alternation between (9.la) and (9.lb), children sometimes generalize this alternation to verbs such as *donate*, and so produce ungrammatical sentences such as (9.1d). They also demonstrated that when presented with novel, nonce, verbs in the prepositional dative construction, children will productively use them in the double object dative construction in appropriate contexts. However, ultimately children do learn which

verbs cannot occur in the double object dative construction, and so we need a theory which can explain why children first make such generalizations, and then subsequently learn the correct subcategorizations.

(9.1) a. John gave a painting to the museum.
b. John gave the museum a painting.
c. John donated a painting to the museum.
d. *John donated the museum a painting.

While the main point of Pinker (1989) is that syntax cannot be learned from distributions alone, he acknowledges that the fact that certain syntactic structures do not occur could be used as indirect negative evidence that these structures were ungrammatical. However, he notes that children can neither consider that all sentences which they have not heard are not grammatical, and nor do they rule out all verb argument structure combinations which they have not heard. He says that it is necessary to identify 'under exactly what circumstances does a child conclude that a nonwitnessed sentence is ungrammatical?' (p.14). The computational model presented in this chapter is able to do just this, and so predict that a verb such as *donate* cannot occur in the double object dative construction, while at the same time predicting that a novel verb encountered only in the prepositional dative construction will follow the regular pattern and also appear in the double object dative construction.

9.1 Bayesian Grammatical Inference

Most work in syntactic theory assumes that grammars are not statistical, that is that they specify allowable structures, but do not contain information about how frequently particular words and constructions occur. However, if grammars were statistical, it appears that it would be much easier to account for how they were learned. In fact Hendriks (2000) has gone so far as to argue that the logical problem of

language acquisition does not apply to human language learners at all, because the arguments supporting the existence of the problem have been based on the premise that people learn language logically. Hendricks has noted that most aspects of human behaviour are not particularly logical, and so there is no reason to suppose that the language acquisition mechanism is either. Feldman et al (1969) proved that as long as grammars were statistical, and so utterances were produced with frequencies corresponding to the grammar, then languages are learnable. They note that proofs that language is not learnable rely on the possibility of an unrepresentative distribution of examples being presented to the learner. While under Feldman et al's learning scheme it is not possible to be certain when a correct grammar has been learned, as more data is observed, it becomes more and more likely that the correct grammar will be identified.

Feldman et al's proof uses Bayes' theorem, which relates the probability of a hypothesis given observed data to the *a priori* probability of the hypothesis and the probability of the data given the hypothesis. For a fixed set of data, the best hypothesis is that, for which the product of the *a priori* probability of the hypothesis and the probability of the data given the hypothesis, is greatest. Feldman et al relate the probability of a grammar (seen as a hypothesis about language) to its complexity – more complex grammars are less probable *a priori*. As grammars are statistical, it is also possible to calculate the probability of the data given a grammar. This leads to an evaluation criterion for grammars where the complexity of a grammar is weighed off against how much data it has to account for, and how well it fits that data. A more complex grammar can be justified if it accounts for regularities in the data, but otherwise a simpler grammar will be preferred.

Feldman et al's evaluation measure for grammars can be seen as a form of *minimum description length*. Minimum description length is a general purpose evaluation measure for determining how well any theory accounts for some observed data. When applied to the acquisition of syntax, the 'theory' will be a particular grammar, and the 'data', example sentences of the language being learned. The basic principle of minimum description length is that the best theory is the one which gives the simplest explanation of the data, where 'simplest' means 'specifiable using the least amount of information'. This kind of evaluation measure may seem to be unrelated to Feldman et al's Bayesian approach, but we can use information theory (Shannon, 1948) to relate quantity of information to probability. Shannon showed that the amount of information conveyed by an event (or a symbol in a grammar, or a word in a sentence) is equal to the negative logarithm of its probability[88]. It is this equivalence between probability and information that allows us to link minimum description length and Bayesian inference. Rissanen and Ristad (1994) explained that 'It is important to remember that probabilities and code lengths are interchangeable, and so the MDL framework is technically equivalent to the Bayesian framework.' (p165).

While minimum description length can be seen as a form of Bayesian inference, it is often much easier to devise coding schemes for grammars and sentences, which we can evaluate based on the amount of information they contain, than to directly specify prior probability distributions over the set of all possible grammars. Feldman et al's

[88] We should note that *information* is used here in a specific technical sense, which was defined by Shannon (1948). The units in which the quantity of information will be measured depend on the base to which logarithms are taken, but it is conventional to use base two, in which case the units will be bits, and that is the approach taken in this chapter.

learning system gave the best evaluation to grammars with the highest *a posteriori* probabilities, and the *a posteriori* probabilities were calculated by multiplying the *a priori* probability of the grammar by the probability of the data in terms of the grammar. However, when this is restated in terms of minimum description length, the best evaluation is given to the grammar which results in the shortest overall description length, and that is found by adding the description length of the grammar itself to that of the data when it is encoded in terms of the grammar[89].

Coding a grammar involves defining the symbols of which it is comprised, and the probability of each, and then coding it symbol by symbol. Shannon's information theory can be used to calculate the coding length of each symbol (based on its probability), and by adding together the coding lengths of each occurrence of each symbol used in the grammar, we can determine the overall coding length for the grammar. Coding schemes can also be devised for other kinds of elements that we might find in a grammar, but which would not normally be thought of as 'symbols', such as integers (Rissanen, 1983).

The next step is to calculate a coding length for the data in terms of the grammar. A grammar will place restrictions on the possible sentences in a language (or, if it is a

---

[89] This can be shown with the following simple proof, which starts from Bayes' rule. In this proof the data, *d*, is constant, but we are considering alternative hypotheses, *h*.

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

$$\therefore P(h \mid d) \propto P(h)P(d \mid h)$$

$$\therefore -\log P(h \mid d) \propto -\log P(h) - \log P(d \mid h)$$

probabilistic grammar, it may simply specify a probability distribution over all possible sentences). We can see a grammar as specifying a set of options available to a speaker. When a speaker produces a sentence, they will have to choose a series of these options, and, using the minimum description length viewpoint, we can see this as specifying a series of symbols. A grammar will generally only allow a limited number of choices to be made at any point in a sentence, and it is such constraints which result in short description lengths for the data.

In general, the more constraining a grammar is, the shorter the description length it will assign to the data, because less options will be available. However more constraining grammars will generally have to be more complex, so the minimum description length principle will tend to trade off complexity of grammar and degree of fit to data to arrive at a reasonable compromise. In general it would be possible to have a grammar with a very short description length if that grammar did not describe any regularities in the data, but then the data would have a very long description length, because the grammar would not constrain the choices available when coding data, and so such hypotheses would have bad overall evaluations. The other extreme situation is having a complex grammar which specifies the observed data exactly. There would now be no choices to make in specifying the data component, and so it would have a description length of zero, but the grammar would be very complex, because it would have to specify all the regularities in the data. Again this would usually result in a very long overall description length. The shortest description length, and hence best evaluation, would normally be for a grammar which came in between these two extremes.

The invention of minimum description length is sometimes credited to Rissanen (1978), but the general principle goes back at least to Solomonoff (1964a, b) and Kolmogorov's (1965) work is closely related, though he did not present it as a theory of inductive inference. Wallace and Boulton (1968) developed and implemented an evaluation metric known as minimum message length, which is also based on the same general principle, but which can clearly be distinguished from the usual implementations of the minimum description length principle. Baxter (1996), Oliver and Hand (1996) and Wallace and Dowe (1999) discuss some of the properties of minimum message which make it distinct from minimum description length. Wallace and Boulton (1968) were probably the first people ever to implement a learning system using a minimum description length type of inference mechanism, as previous work had been purely theoretical.

An important aspect of Solomonoff's (1964a, b) papers was that he proposed an inference mechanism which would consider all possible grammars simultaneously. These multiple grammars could be used in a weighted sum to provide a more accurate measure of the best overall grammar, and to generalise more accurately when classifying new data. This contrasts with minimum description length as it was described above, as there it was suggested that the aim is simply to find the single grammar with the best evaluation (and hence the highest *a posteriori* probability). However, in section 3.3.1 it was noted that the colour terms model considers all hypotheses simultaneously, and so the classification of each hue is based on a weighted summation of the probability of each individual hypothesis which included that hue. (This is termed hypothesis averaging.) Solomonoff's proposal concerned equivalent procedures for minimum description length, so it would clearly be desirable to implement them, as this should lead to more accurate learning. However,

this is problematic, because when using minimum description length we often have an infinite number of possible grammars, and even if not, there are usually so many that it would not be possible to consider all of them. Hence, in practice, minimum description length approaches are usually limited to searching for the single grammar with the best evaluation. This problem did not occur for Solomonoff, because his papers were theoretical, and he did not implement any learning algorithms on a computer. It might be possible to address this problem by considering a random (or perhaps stratified) sample of possible grammars, rather than all possible grammars, in an approach similar to the Monte Carlo method (Andrieu et al, 2003). This kind of approach was implemented by Fitzgibbon, Dowe and Allison (2002), but I am not aware of any work applying this method as part of a psychological model.

The minimum description length principle allows us to construct an evaluation measure that, given two or more grammars and a corpus of data, allows us to determine which is the best. However, there remains a problem because, as discussed above, there will usually be such a large number of possible grammars that we cannot consider all of them in turn, and it is certainly not possible that a child learning a language could. Hence this raises the problem of how we arrive at the candidate grammars which are to be evaluated. In the next section, I describe computational models which are able to learn grammars despite this problem. They do this by starting with a standard initial grammar, and then making small iterative changes which gradually lead towards the correct grammar. This avoids the need to consider every single possible grammar, and so allows grammars to be learned within a reasonable amount of time.

We should note that minimum description length is typically used as part of a machine learning algorithm by researchers who have little interest in the mechanisms which humans use to learn. However, Chater (1999) argued that minimum description length may form the basis of a fundamental cognitive principle, and that it could have applications in many cognitive domains other than language, including in perception, memory, reasoning under uncertainty, in judgments of similarity, and in learning from experience in general. He hypothesized that 'the search for simplicity is a fundamental cognitive principle' (p298), and that minimum description length, or a related measure, might be the mechanism which is used in a wide variety of cognitive tasks. There have been several cognitive models which have used minimum description length besides Chater's own work. For example, Fass and Feldman (2002) used minimum description length to model how people learned categories from examples. Such work clearly demonstrates that other researchers also consider that minimum description length is a principle that humans may use in learning. Further applications of minimum description length in psychology, this time to modelling language acquisition, are reviewed in the next section.

9.2 Computational Models of Syntactic Acquisition

There is a considerable number of computational models which are related to the one which is the topic of this chapter, either because, like it, they have used minimum description length, or they have incorporated some other bias towards simplicity. Firstly, however, it seems worth mentioning one model which did not incorporate a simplicity bias. Carroll and Charniak (1992) reported a method which was used to induce dependency grammars from corpora. Their system tried to find the grammar which best fitted the corpus, but unfortunately this tended to result in grammars containing both large numbers of rules, and very complex rules. Such results led them

to suggest that it might be necessary to incorporate a preference for simpler grammars into their system, though they were able to improve the performance of the system considerably by placing restrictions on the permissible grammars. However, making such restrictions would seem to give some indication of the correct grammar *a priori*, and so reduce the degree to which the grammars were actually learned as opposed to being pre-specified. Hence their system could be seen as providing an empirical demonstration of the problems involved in learning grammars in the absence of a simplicity principle.

Probably the first proposal to use a simplicity metric as part of a system for learning grammars was made by Solomonoff (1960). He noted that if grammars were to be learned without using negative evidence then 'the problem of finding a grammar that is consistent with a given fixed body of text is complicated by the fact that there are always an infinite number of such grammars.' (p191). However, he went on to propose a solution to this problem which, although he did not specify it formally, was essentially the basis of the minimum description length principle. He stated that 'It is possible, however, to define a "simplest" grammar from among all possible consistent grammars. Another important condition is that the languages defined should contain as "few" sentences as possible, in addition to the fixed body of text. The meaning of "few" must be suitably defined, since most languages of interest contain an infinite number of sentences.' (p191). Solomonoff did not implement such a system, but the model presented in this chapter learns using an implementation of the general principles which he originally proposed.

Langley (1995) created a grammar learning system which learned context free phrase structure grammars using a preference for simpler grammars as a guiding principle.

Langley's system was based on earlier work by Wolff (1987, 1991), and began learning with an initial grammar which did not capture any generalizations, but which essentially just listed every training sentence in its rules. Heuristics were then used to find candidate new rules, which could be created either by adding new categories, which would be substituted for recurring sequences in other rules, or by merging two categories into a single one. After finding a set of candidate merges, the system then chooses the one which results in the simplest grammar, where simplicity is measured by counting the number of symbols on the right hand sides of the grammar rules. Searches proceeded in this way, until no further improvements in the grammar could be found. Throughout the learning process, the grammars were always able to parse all of the training corpus, because the changes made to the grammars could only increase their generality or leave it unchanged, and could never make the grammars more restrictive. However, there was a problem with Langley's (1995) approach, because, for any language, it would always be possible to create a very simple grammar which failed to capture any regularities in the training data. In order to prevent grammars such as these being learned, Langley's system relied on examples of sentences which were not grammatical. Grammars which were able to parse these negative examples would not be permitted. However, this is problematic from a linguistic point of view, because it is usually assumed that such negative evidence is not available to children when they learn their languages.

In a new version of Langley's learning system, Langley and Stromsten (2000) avoided the need to use negative evidence, by changing the measure of simplicity used in evaluating grammars to a minimum description length one. The evaluation measure now took account of how well the grammar fitted the data, as well as how complex it was. They noted that this would continue to direct the learning system

away from ‘large grammars with overly specific rules’ (p 223), but would also avoid ‘very small, overly general grammars because they can describe too many unobserved strings’ (p 223). Given this modification, the system performed well without using negative evidence when learning grammars for simple languages of between 8 and 24 words.

Stolcke (1994) created a grammar learning system very similar to Langley and Stromsten’s (2000) one. His system was somewhat different, however, because the grammars learned were statistical, and his new training data could be incorporated during learning. Langley’s systems required that all the data was available before learning began. In Stolcke’s system, decisions as to what to merge, and when to stop merging, were made so as to maximise the Bayesian posterior probability of the grammar given the data. However, this Bayesian evaluation measure was effectively a form of minimum description length. Using this method, it was possible to learn a grammar for simple subsets of English containing 8-12 words, including one of the languages learned by Langley’s (1995) system. Further experiments applied the system to real language examples, but with only limited success (though the language was, in any case, restricted to a very limited domain, hence making the learning task easier). Chen (1995) applied a similar approach to relatively large corpora of real English, but his system could not learn context free languages, but only the simpler regular languages. Chen showed that his system was better than some others at predicting statistical regularities in the corpus, which was the aim of his system. However, it clearly did not come close to the goal of learning a grammar capable of capturing the overall structure of the language, and Chen did not give examples of specific structures learned, so his work probably has little relevance to theoretical linguistics.

Grünwald (1994) also attempted to learn phrase structure grammars from large amounts of unrestricted text, more specifically from the Brown corpus (Francis and Kucera, 1979). However, he reported that this approach did not lead to very good results, so the results he presented related only to a restricted version of his model. This restricted version did not attempt to learn grammars, but only to put words into classes. Initially each word was placed in a separate class, and bigram statistics were then calculated, based on the frequency of each class following each other class in the Brown corpus. The description length, both for the set of classes, and for the corpus when the bigram statistics were used to help predict regularities, was calculated. Merges of all pairs of classes were then considered, and, for the grammars resulting from each merge, the description length was calculated. Best-first search was used, so whichever merge resulted in the greatest decrease in description length was made, and the set of all possible new merges was then calculated again, and the process repeated. Learning was stopped when no merge could be found which would reduce the overall description length. The result of this learning process was a set of classes which generally appeared to correspond well to the kind of classes proposed by syntacticians, including adjective classes, classes for various subtypes of verb, and one including just the two nominative pronouns *he* and *she*.

Grünwald's results were similar to those of Redington et al (1998) discussed above, though there is an important difference in that Redington et al's program produced hierarchical clusters, but did not actually define distinct classes. Hence, for this reason, we might consider Grünwald's system to be preferable, although this is debatable, because it is not clear that distinct classes such as those found by Grünwald's system do actually exist. It is possible that syntactic categories are really more *fuzzy*, with membership being gradable (Taylor, 1989). Furthermore, it seems

that, for categories such as verbs, there is clear evidence for hierarchical categorisation, as we can identify many subtypes of verb, such as intransitive, transitive and ditransitive verbs. However, regardless of which, if any, of these possibilities is correct, the results of Grünwald's program are clearly interesting from a theoretical linguistic viewpoint, as they demonstrate clearly that syntactic classes can be learned simply on the basis of distributions.

There has also been a considerable amount of work which has used minimum description length to learn some aspects of language structure, but with the principle aims of the technique being concerned with applications in language technology, rather than advancing our understanding of language from a theoretical perspective or explaining how children learn languages. Such work includes Osborne (1999), Dowman (2000) and Starkie (2001). Although the aim of such work is not to advance linguistic theory, that does not mean that is of no interest to researchers in theoretical linguistics. As should be apparent from the above review of literature, developments in language technology can have applications in linguistic theory. For example, Solomonoff's (1960) paper was primarily concerned with machine translation, but his work has now been applied as the basis of a number of cognitive models, including the one presented in this chapter.

There also has been a considerable amount of work which has applied minimum description length to areas of linguistics other than syntax. Firstly Ellison (1992) and Rissanen and Ristad (1994), looked at phonology. Ellison showed how the distinction between consonant and vowel phonemes could be learned based on distributional evidence, and how vowel harmony systems could be learned. Rissanen and Ristad showed how minimum description length could help with modelling the acquisition of

metrical systems within a parameter setting framework. The acquisition of morphology using minimum description length was modelled by Brent (1993) and Goldsmith (2001), who both created systems which could analyse texts and discover the morphological structures of the words they contained. Brent and Cartwright (1997) and de Marcken (1996) showed how minimum description length could be used to deduce the correct segmentation of a stream of continuous speech (or written text) into linguistic units, such as words. This is a problem which children must address, because human speech does not leave gaps between words, and nor is there any other clear demarcation of word boundaries. A similar approach by Venkataraman (2001) could also be seen as a form of minimum description length, though it is described in terms of probabilities and not coding lengths, and it does not distinguish between separate theory and data components. All of this work provides evidence supporting the hypothesis that children use minimum description length to learn language. While it represents a broad spectrum of different approaches, any work showing that minimum description length could be used to learn one aspect of language, would seem to be supportive of the proposition that it is used to learn syntax, and hence of the model of syntactic acquisition described below.

9.2.1 Description of Model

Dowman's (1998) model[90] learned grammars for simple subsets of several languages, including the English data given in Table 9.1, which corresponds to the grammar

[90] Dowman (1998) is a publication of my MA (Honours) dissertation, and so does not form part of the work submitted for the PhD. However, the application of the model to learning verb subcategorizations, and the rest of this chapter, is all new material undertaken for the PhD.

given in Table 9.2. (In the grammar, V_t is an abbreviation for 'mono-transitive verb', and V_s for 'verb taking a sentential complement'.) This was the same model that was used to learn the dative alternation in research for this thesis. The only *a priori* knowledge of the structure of the corpus which was available to the model, was implicit in the grammatical formalism with which grammars were specified. This formalism restricted the model to using binary branching or non-branching phrase structure rules, introducing each word with a non-branching rule, and using no more than eight non-terminal symbols. The non-terminal symbols were all equivalent arbitrary symbols, except that each grammar would contain one special symbol, *S,* with which each top down derivation would begin.

John hit Mary	Ethel thinks John ran
Mary hit Ethel	John thinks Ethel ran
Ethel ran	Mary ran
John ran	Ethel hit Mary
Mary ran	Mary thinks John hit Ethel
Ethel hit John	John screamed
Noam hit John	Noam hopes John screamed
Ethel screamed	Mary hopes Ethel hit John
Mary kicked Ethel	Noam kicked Mary
John hopes Ethel thinks Mary hit Ethel	

Table 9.1. Data for English.

$S \rightarrow NP\ VP$	$V_s \rightarrow$ thinks
$VP \rightarrow$ ran	$V_s \rightarrow$ hopes
$VP \rightarrow$ screamed	$NP \rightarrow$ John
$VP \rightarrow V_t\ NP$	$NP \rightarrow$ Ethel
$VP \rightarrow V_s\ S$	$NP \rightarrow$ Mary
$V_t \rightarrow$ hit	$NP \rightarrow$ Noam
$V_t \rightarrow$ kicked	

Table 9.2 Grammar Describing English Data.

The frequency, and hence probability, with which each symbol (including words) appeared in the grammar was specified, and so the amount of information required to specify each symbol in a grammar could be calculated. (Shannon's (1948)

information theory defines the quantity of information conveyed by an event as the negative logarithm of its probability. It is conventional to take logarithms to base two, so that the units of information will be bits, which is the approach taken here.) A specification of a grammar would consist of a list of groups of three symbols, one for a rule's left hand side, and two for its right hand side (a special null symbol being incorporated for use in non-branching rules). For example, to specify the rule in (9.1), firstly the symbol *VP* would be encoded, followed immediately by the *V* symbol and then the *S* symbol. To encode (9.2), we would first specify the *VP* symbol, then the *screamed* symbol, then the null symbol. If there were ten rules in the grammar, then there would be a total of 30 symbols to be specified. If 5 of these were *VP*, 3 were *V*, 2 were *S*, only one was screamed, and there were 8 null symbols, then the coding length of these two rules (*I*) would be given by (9.3). (9.3) is simply a sum of the negative logarithms of the probabilities of each symbol needed to specify the two rules (each of which is included in the same order as it appears in the rules).

(9.1) VP → V S

(9.2) VP → screamed

(9.3) $I = -\log_2 5/30 -\log_2 3/30 -\log_2 2/30 -\log_2 5/30 -\log_2 1/30 -\log_2 8/30$

As the grammar was statistical, it was also necessary to record how often each rule was used in parsing the corpus. It was assumed that a fixed amount of information could be used to specify these probabilities, and so 5 bits of information was added to the evaluation of the grammar per rule. (The assumption of 5 bits of information is fairly arbitrary, but sufficient for the purposes described here.) The total cost of the grammar was the amount of information needed to specify each symbol in the grammar, and each rule's frequency.

Given such grammars, the data was then parsed left to right, bottom up, with only the first parse found for each sentence being considered, and an ordered list of rules needed to derive the sentence obtained. This list allows us to make a probabilistic encoding of the data in terms of the grammar. Given the probabilities of the rules, and always knowing the current non-terminal symbol being expanded (starting with *S*, and always expanding the left most unexpanded non-terminal), it is only necessary to specify which of the possible expansions of that symbol to make at each stage. Hence, if a grammar accounts well for regularities in the data, little information will be required to specify the data. If a symbol can only be expanded by a single rule (such as *S* in the grammar above), then no information is necessary to specify that that rule is used.

For example, given the grammar of Table 9.2, we could encode the sentence *Mary ran* using the first rule, followed by the third of those rules which expand *NP*, and finally the first rule which expands *VP*. As we start with *S*, the first rule we choose must expand this symbol. As there is only one rule with *S* on its left hand side, we must use that rule. As this is the only rule which can be used, the probability of choosing it is one. Now we have the derivation *NP VP* to expand (as we have substituted for *S* what was on the right hand side of the rule expanding *S*). Firstly we must choose which of the rules with *NP* on their left hand sides to use. If these rules occur ten times in the parses of the data, but the third of them is only used once, then it will have a probability of 1/10. Application of this rule results in the derivation *Mary VP*. *Mary* does not need further expansion, so we now move to expanding the *VP* symbol. We need to use the first *VP* rule, which is used, say, 4 times, while overall the *VP* rules are used with a total frequency of, say, 20. Hence the probability of using this rule at this point in the derivation is *4/20*. We can calculate the total coding length

of this sentence (*I*) by taking the sum of the negative logarithms of the probabilities of using each of the rules needed to code it, as shown in (9.4). (Note that the coding cost of using the first rule, which has a probability of one, evaluates to zero.)

(9.4) $I = -\log_2 1 - \log_2 1/10 - \log_2 4/20$

By summing the amount of information needed to specify the grammar rules, the frequencies of those rules, and the data given that grammar, we obtain an evaluation for each grammar, with lower evaluations corresponding to better grammars. However, in order to complete the model of acquisition, it is necessary to describe the search mechanism that was used for generating and testing grammars.

The model started learning with a simple grammar of the form given in Table 9.3, with a rule introducing each word. This grammar is very simple, hence having a good evaluation itself, but it does not describe any regularities in the data, and so has a very bad evaluation in that respect, resulting in a poor overall evaluation.

S → X S	S → X
X → John	X → Ethel
X → Mary	X → Noam
X → ran	X → screamed
X → hit	X → kicked
X → thinks	X → hopes

Table 9.3. Form of Initial Grammars.

The model would begin learning by making one of five changes to the grammar. The change to be made would be selected at random from one of the possibilities listed below, each of which is followed by the probability of that change being made. (The particular probabilities used were chosen by a process of trial and error, and hence are fairly arbitrary.)

- Adding a new rule (which would be the same as an old rule, but with one of the symbols in it changed at random). (1/6.)

- Deleting a randomly chosen rule. (1/6.)

- Changing one of the symbols in one of the rules. (17/48.)

- Changing the order of the rules[91]. (7/48.)

- Adding a pair of new rules. These new rules would be based on two existing rules, in which a non-terminal system occurring on the right hand side of the first also occurred on the left hand side of the second. The new rules would be created by changing that symbol to another non-terminal system, and then adding the two new rules to the grammar. The original rules would be left unchanged. (For example if there were rules X → Y and Y → Z, then two new rules might be added, X → A and A → Z) (1/6.)

These changes were chosen largely because they are simple, but still enable any grammar specifiable within the grammar formalism to be reached, when a number of the rules are applied in the right combination. For example, starting with the grammar given in Table 9.3, the nouns could be placed in a separate class if the changes listed in Table 9.4 were made. The changes would probably have to be made in this order, as, after every change, the grammar must still be able to parse the whole corpus, so

---

[91] This could potentially affect the resulting evaluation, as the evaluations were always based on the first parse found by the parsing mechanism, which always tried to apply earlier rules before considering later ones.

we could not, for example, delete the original rule introducing *Ethel*, before adding the new one. (This probably would not in itself improve the evaluation of either the grammar, or the data specified in terms of the grammar, but such a change might happen by chance, and could eventually lead to a better overall grammar.)

Add	$Y \rightarrow$ Ethel
Add	$Y \rightarrow$ Noam
Add	$S \rightarrow Y\ S$
Add	$S \rightarrow Y$
Delete	$X \rightarrow$ Ethel
Delete	$X \rightarrow$ Noam

Table 9.4. Examples of Changes that Could be Made to an Initial Grammar.

The last grammar edit rule in the list of rules given above, was introduced to help in cases where a single symbol is used in several rules, but where using two separate symbols would enable a better evaluation to be obtained. (For example, if both nouns and verbs were introduced with the same symbol, then this rule might help to produce a grammar in which nouns and verbs were both introduced by a different symbol.) Dowman (1998) used a slightly more complex system than this, but further investigations have revealed that this learning system works well. It was applied to the same data set used by Dowman (1998) (Table 9.1), and reproduced the results obtained with the more complex system which were reported by Dowman (1998). Hence, I will not repeat here the other rules used by Dowman (1998), as it was this simpler system which was used for deriving the new results presented in this chapter. However, the rest of the description of the model applies to both Dowman's (1998) system, and the modified system developed for explaining the dative alternation.

After each change the evaluation of the new grammar with respect to the data would be calculated. If the change improved the evaluation of the grammar then it would be kept, but if the new grammar was unable to parse the data, it would be rejected. If the

change made the evaluation of the grammar worse, then the probability that it would be kept would be inversely proportional to the amount by which it made the evaluation worse. Throughout learning, the probability that changes resulting in worse evaluations would be accepted, was gradually reduced. This is an implementation of annealing search (Aarts and Korst, 1989), which enables the system to learn despite finding locally optimal grammars in the search space. The program learned in two stages, in the first only taking account of the evaluation of the data in terms of the grammar (making it easier to find the grammatical constructions which best fitted the data), and in the second taking account of the overall evaluation (and so removing any parts of the grammar which could not be justified given the data). After a fixed number of changes had been considered (less than 18,000 in the case of the above data) learning would finish with the current grammar, no improvements usually having been found for a long time. For efficiency reasons, there were also limits placed on how deeply the parser could search for correct parses, and on the maximum number of rules which the grammar could contain at any stage of the search. Because the search strategy is stochastic, it is not guaranteed to always find the optimal grammar every time. Hence the learning mechanism would run the search several times, and select the grammar with the best overall evaluation. This discussion of the model should specify it in enough detail, both to enable its reproduction, and for a full understanding of how it works to be obtained, but it is, none the less, clearly very brief. A more detailed description can be found in Dowman (1998).

9.2.2 Results

When used to learn from the English data in Table 9.1, the system learned a grammar which corresponded exactly to that in Table 9.2 in structure. (As linguistic categories are not known *a priori*, the system simply used a different arbitrary symbol to

represent each learned category.) Table 9.5 shows that this grammar was preferred because, while the grammar itself is more complex than the initial one, and so receives a worse evaluation, it captures regularities in the data, and so improves the evaluation of the data with respect to the grammar by a greater amount. (We should note that the overall evaluation is equal to the sum of the evaluation for the grammar and data components. In the case of the learned grammar, the sum of the evaluations given for each component does not exactly equal that for the overall evaluation, but this is simply due to rounding error, as all the evaluations are given to an accuracy of one decimal place.) Dowman (1998) used this same learning system (without any modifications except to the maximum number of non-terminal symbols) to also learn aspects of French, Japanese, Finnish and Tigak.

	Initial state of learning	Learned Grammar
Overall Evaluation	406.5 bits	329.5 bits
Grammar	160.3 bits	199.3 bits
Data	246.2 bits	130.3 bits

Table 9.5. Evaluations for English Grammar.

9.3 Learning Verb Subcategorizations

Given Dowman's (1998) success in learning simple syntactic systems, it was decided to investigate whether the same model could be used to learn some of the kinds of phenomena which it has been argued are especially problematic for theories of learning. In particular it was investigated whether the distinction between sub-classes of ditransitive verbs such as *gave* and *donated* could be learned.

There were three key results which the model aimed to replicate. Firstly, children eventually learn a distinction between verbs which can appear in both the double object and prepositional dative constructions, and those which do not show this alternation. Secondly, when children encounter a previously unseen verb they use it

productively in both constructions. Finally, during learning, before children have seen many examples of an irregular verb which only occurs in a subset of the possible constructions of other verbs, they use that verb productively in constructions in which it is not grammatical[92].

9.3.1 Data Used for Learning

The same model was used as in Dowman (1998), but this time the data consisted of two types of sentences, prepositional datives such as (9.2a) and (9.2b), containing one of the verbs *gave*, *passed*, *lent*, or *donated*, and double object datives such as (9.2c), containing *gave*, *passed* or *lent*, but not *donated*. Each of these four verbs occurred with roughly equal frequency, and the alternating verbs (*gave*, *passed* and *lent*) were just as likely to appear in either construction. In addition, the sentence (9.2d) was added, containing the only example of the verb *sent*. This was so that it would be possible to see if the model would place a newly seen verb in the regular or irregular class of verbs (assuming that it learned two such classes).

Noun phrases consisted of either one of two proper nouns, or one of the two determiners *a* or *the*, followed by either *painting* or *museum*. There were no biases as

---

[92] Throughout this chapter donate is described as irregular, because it does not participate in the dative alternation, but it is one of several verbs for which this is the case. Hence its non-participation in the alternation might best be described as a sub-regularity, an analysis likely to be preferred by many theorists who have worked on the alternation, such as Mazurkewich and White (1984). However, whether or not that is in fact the case, the argument presented here is applicable to any situation in which a lexical exception prevents a word from being used in a regular construction. There is further discussion concerning the correct analysis of the alternation in section 9.4.

to which noun phrase was most likely to occur in which position. Clearly, this leads to a situation in which many of the sentences in the corpus, while syntactically correct, are semantically very strange. For example, (9.2b) is unusual, because John would normally refer to a person, which is not something which could typically be donated to a museum. Overall the data consisted of 150 sentences, and the full corpus is given in Appendix D.

This data set was created completely artificially, and is in many ways unrealistic. However, it enables us to focus on the central question of whether the dative alternation can be learned. Crucially, Pinker's (1989) arguments, concerning the learnability paradox created by the dative alternation, are equally applicable to this simple data set as to real languages. Whether the corpus is realistic, is not in itself important in resolving the issue of whether the non-occurrence of a particular construction could be used as evidence by children that that construction is not grammatical. Whether the model is capable of overcoming the learnability problem was tested by applying it to this data set, and investigating whether the resulting grammar accounted correctly for the subcategorizations of all the verbs. No modifications were made to the model used to learn the data in Table 9.1, except that in order to cope with the more complex data set the maximum number of non-terminals was increased to 14, and the number of iterations in the search was also increased.

(9.2) a. John gave a painting to Sam.
b. Sam donated John to the museum.
c. The museum lent Sam a painting.
d. The museum sent a painting to Sam.

9.3.2 Results

The initial and final evaluations of the grammars are given in Table 9.6. (Again, the overall evaluation is not exactly equal to the sum of the grammar and data components for the learned grammar, due to rounding error.) This grammar was produced by running the model twenty times, and selecting the grammar which received the lowest overall evaluation on any of the runs. (Most of the runs produced different grammars, but each of those grammars had a worse overall evaluation). The model was implemented in SICSTUS PROLOG, and was run on two Sun UNIX workstations. It took approximately three weeks to obtain twenty complete runs of the program on these machines, but this overestimates the total execution time, as each run of the program was started manually, and hence there would be periods of time when the program was not running. Dowman (1998) showed that the program is likely to execute much more slowly when significantly more complex grammars need to be learned. Further details of the program, including all the source code, appear in Appendix E.

	Initial state of learning	Learned Grammar
Overall Evaluation	3445.6 bits	1703.4 bits
Grammar	190.3 bits	321.0 bits
Data	3255.3 bits	1382.3 bits

Table 9.6. Evaluations for Ditransitive Verbs Data.

We can see from Table 9.6 that, as with the case of learning from the English data of Dowman (1998), a more complex grammar has been learned, and it accounts better for regularities in the data than the original grammar did. The grammar is shown in Table 9.7, where the arbitrary symbols have been replaced with more interpretable ones, and where the rules appear in a different order to improve clarity. (V_r is an abbreviation for regular verb, and V_i for irregular verb.) Examination of the learned

grammar showed that the verbs had been divided into two classes. (They have different symbols on the left hand sides of the rules producing them.) *passed*, *gave*, *sent* and *lent* have all been placed in one class, while *donated* appeared in a class of its own. The grammar is able to generate only grammatical sentences, so *gave*, *passed*, *lent* and *sent* may appear in both double object and prepositional dative constructions, while *donated* may occur only in the prepositional dative construction. This has been learned even though there was no data explicitly indicating that *donated* did not follow the regular pattern, and even though *sent* only occurred once, and in the prepositional dative structure.

$S \rightarrow X\ NP$	$DET \rightarrow a$
$X \rightarrow NP\ Y$	$DET \rightarrow the$
$Y \rightarrow V_r\ NP$	$N \rightarrow museum$
$Y \rightarrow V_r\ Z$	$N \rightarrow painting$
$Y \rightarrow V_i\ Z$	$V_r \rightarrow passed$
$Z \rightarrow NP\ P$	$V_r \rightarrow gave$
$P \rightarrow to$	$V_r \rightarrow sent$
$NP \rightarrow DET\ N$	$V_r \rightarrow lent$
$NP \rightarrow John$	$V_i \rightarrow donated$
$NP \rightarrow Sam$	

Table 9.7. Grammar Learned from Ditransitive Verbs Data.

The structures which the grammar assigns to sentences are, however, clearly not correct, as can be seen from the parse shown in Figure 9.1 below. The model has identified a clear noun phrase category (*NP*), and has placed all the individual words into appropriate classes, but it has not correctly identified other phrasal constituents, such as verb phrases or prepositional phrases. Instead there are three phrasal constituents, *X*, *Y* and *Z*, which clearly do not correspond to any phrasal constituent in English. This analysis should not be surprising, as there was no information in the data presented to the model, to demonstrate that this type of structure is incorrect. For example, there was no evidence to show that the preposition forms a constituent with the following noun phrase, rather than the preceding one. Dowman (1998) showed

that the model can learn structures such as verb phrases when presented with data which contains evidence supporting the existence of such phrases, so we should expect that if the ditransitive verbs data were augmented with further appropriate data, the model would then learn a grammar which assigns correct structures to all the sentences, as well as learning the correct subcategorizations for the verbs.

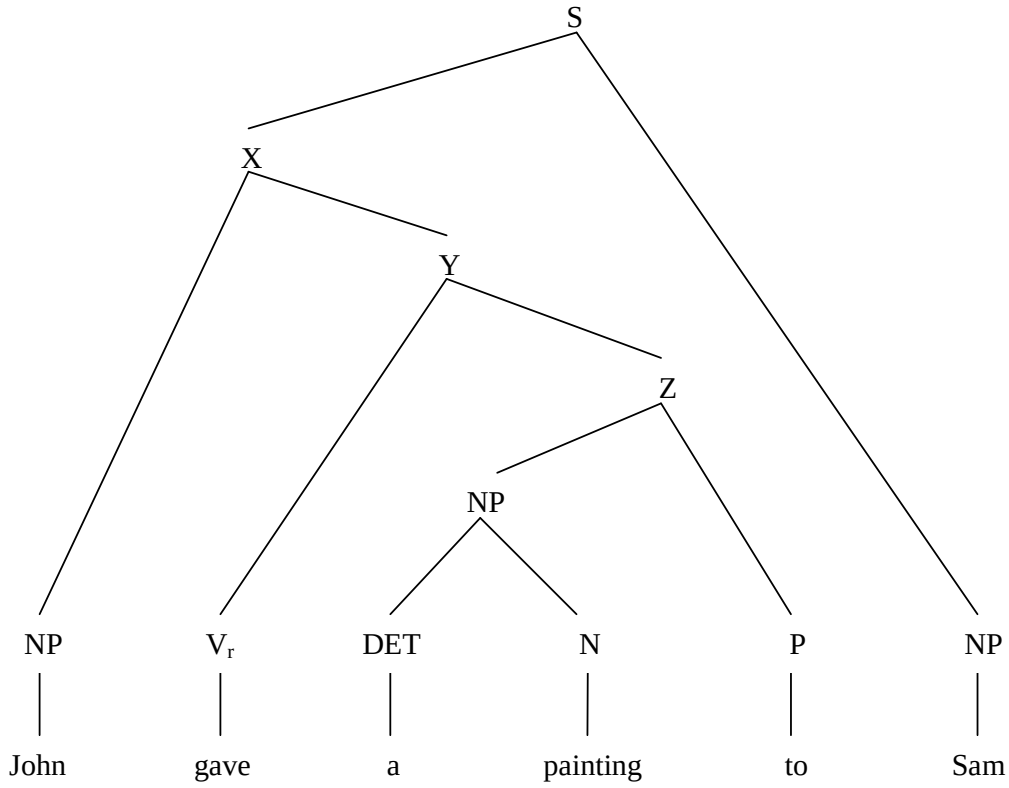


Figure 9.1. A Structure Assigned by the Learned Grammar

It is interesting to investigate exactly why *sent* was placed in the class of regular verbs, rather than being grouped in the irregular class along with *donate*. There was no clear evidence showing in which of the two classes *sent* occurred, so it is not obvious why the model preferred placing it in the regular class. The grammar was changed so that *sent* was placed in the irregular class by changing the symbol on the left hand side of the rule introducing it to be the same as that which introduces *donate*. Table 9.8

shows the evaluations for the new grammar (as well as those for the grammar learned by the system, for comparison), and we can see that the overall evaluation is very slightly worse, thus confirming that the model prefers the grammar which places *sent* in the regular class. (The model could have arrived at that classification simply by chance.) We might think that this result is surprising, however, because the learned grammar predicted that *sent* could be used in the double object dative construction, and so that sentences of types which were not observed in the training corpus were grammatical. The preferred grammar (shown in Table 9.7), therefore did not model the regularities in the training data as well as the one which classes *sent* as irregular, and this can be seen by comparing the evaluations given to the data, for which the preferred grammar was given a slightly worse evaluation.

	Grammar with *sent* in Irregular Class	Grammar with *sent* in Regular Class
Overall Evaluation	1703.6 bits	1703.4 bits
Grammar	322.2 bits	321.0 bits
Data	1381.4 bits	1382.3 bits

Table 9.8. Evaluations for Ditransitive Grammars with *sent* in Irregular Class or Regular Class.

The reason that the model prefers to classify *sent* as regular, is because this results in a lower evaluation for the grammar component, and, as can be seen by comparing Table 9.6 and Table 9.8, the improvement in the evaluation of the grammar component is greater than the deterioration in the evaluation of the data component. The question that now arises is why does the grammar receive a lower evaluation when *sent* is regular as opposed to irregular? After all, both grammars contain exactly the same number of rules, and the only difference is the symbol on the left hand side of the rule introducing *sent*. The small difference in the evaluation of each grammar must therefore be due to this symbol. The key to understanding this issue is to remember that, not only is the data encoded probabilistically in terms of the grammar,

but that the grammar itself is also coded probabilistically. The frequency of each symbol is recorded, and these frequencies are used to determine the probability of using each symbol in constructing the grammar. Infrequent symbols thus have a higher coding cost than frequent ones. As can be seen in Table 9.7, other than the symbol on the left hand side of the rule introducing *sent*, there are five occurrences of the regular verb symbol (V_r) and only two occurrences of the irregular verb symbol (V_i). There is therefore a lower cost to placing *sent* in the regular class, which involves the addition of another V_r, rather than in the irregular class, which would involve the addition of another, more costly, V_i. Examination of this issue reveals that ensuring that the grammar itself was encoded statistically added to the generalization ability of the model, because it effectively added a bias to prefer regular (and hence more frequent) constructions over irregular (less frequent) ones.

The results above account both for eventual learning of the distinction between syntactically distinct verbs such as *gave* and *donated*, and the productive use of novel verbs in regular constructions. The final phenomenon which I aimed to demonstrate was that, at earlier stages of learning, children overgeneralize, and use verbs such as *donated* productively in constructions in which they are ungrammatical. In order to investigate this phenomenon, the total amount of data was reduced, to simulate a stage of acquisition where children had not been exposed to so many examples of each kind of verb. When the model learned from this data it failed to maintain a distinction between sub-classes of verbs, allowing all verbs to occur in both constructions. This was because there were not enough examples of *donated* to justify making the grammar more complex by creating a separate syntactic class, and so it was simply placed in the regular class.

9.4 Discussion

These results concerning the acquisition of regular and irregular verb subcategorizations show that an aspect of syntax which many other theories would have difficulty accounting for is learnable. Dowman (1998) compared the performance of the model described here to that of connectionist models of syntactic acquisition, such as Elman's (1993) model, and argued that some of the generalization 'successes' could, in the light of the results of the present model, be alternatively interpreted as a failure to learn exceptions.

Elman's network learned a language containing only 23 words, and yet 50,000 sentences were used to train the network. (The training sentences were generated artificially using a grammar.) This means that every word could have been observed in every syntactic position may times over, greatly reducing the need to form generalizations. Christiansen and Chater (1994) investigated to what extent this kind of model was able to generalize to predict that a word observed in one syntactic position would also be grammatical in another position. In order to do this, they trained a similar connectionist network on a more complex language containing 34 words, again using 50,000 sentences. In the training data they did not include *girl* and *girls*, in any genitive contexts, and, *boy* and *boys* in any noun phrase conjunctions. After training they found that the network was able to generalize so that it would allow *boy* and *boys* to appear in noun phrase conjunctions, but it did not generalize to allow *girl* and *girls* to occur in genitive contexts. Christiansen and Chater considered the learning to have been successful in the case of *boy* and *boys*, but not in the case of *girl* and *girls*.

However, the account of the acquisition of verb subcategorizations presented in this chapter relies on statistical properties of the data, and in particular the non-occurrence of certain forms. So, given 50,000 sentences of a language with only 34 words, in which two words did not appear in a given construction, it would seem that a learner would predict that this could not simply be due to chance. Given this perspective, it seems that Christiansen and Chater's network has learned correctly in the case of *girl* and *girls*, but not in the case of *boy* and *boys*[93].

In order to account for distinctions between *gave* and *donated*, it seems that neural networks must be more sensitive to quantitative information in language. The degree to which recurrent neural networks generalize is partly dependent on the fixed architecture of the networks, and in particular on the number of hidden nodes. Bayesian learning methods for neural networks (MacKay, 1995) should be able to solve this problem, by placing a prior probability distribution on network structures and parameter values, although I am not aware of any applications of such networks to modelling language acquisition.

Redington et al's (1998) system for learning word classes is capable of making very fine distinctions between sub-classes of verbs, but unlike the system described here, it

---

[93] Clearly the number of times with which such neural networks are presented with each training example is unrealistically large, and may best be seen as representative of a much smaller number of examples, or of a much more diverse set of examples containing a larger number of different words, so Christiansen and Chater's interpretation of their results is hardly unreasonable. However, a better way to have tested for generalization might have been to have inserted a novel verb in one, or perhaps a few, syntactic positions, and to have looked for generalization of that word to other positions. I suspect that such an experiment would not have been successful given the kind of neural network used.

is not able to decide when the distributions of two words are dissimilar enough that they should be placed into separate classes, and when the difference in distributions is simply due to chance variation within a class. However Boulton (1975) describes a program which does incorporate a Bayesian based metric into this kind of clustering system, and so demonstrates that it is possible to learn discrete classes automatically.

Certainly evaluation procedures based on simplicity metrics are not new to linguistic theory. Chomsky's (1965) theory of syntactic acquisition relied on such a measure to choose between alternative grammars. However, it is possible to identify some key differences which make Chomsky's theory very different from the Bayesian approach suggested here. Firstly Chomsky considered syntax to be fundamentally non-statistical. He had earlier argued that 'Despite the undeniable interest and importance of semantic and statistical studies of language, they appear to have no direct relevance to the problem of determining or characterizing the set of grammatical utterances….[P]robabilistic models give no particular insight into some of the basic problems of syntactic structure.' (Chomsky, 1957, p17). It seems hard to explain how any system which did not monitor the frequencies with which verbs such as *donated* and *gave* are used would be able to account for how the different subcategorizations of these verbs could be acquired.

Probably an even more important difference between the kind of simplicity measure proposed in Chomsky (1965) and the kind used here, is that Chomsky did not incorporate a measure of goodness of fit to data into his simplicity metric. (Interestingly, as was discussed above, this was also true of Langley's (1995) model of syntactic acquisition, though he added such a measure to his Langley and Stromsten (2000) model.) Chomsky's metric simply looked for the grammar which

was shortest, in terms of the number of symbols which it contained. The theory relied on innate constraints on what forms grammar could take in order that ‘significant considerations of complexity and generality are converted into considerations of length, so that real generalizations shorten the grammar and spurious ones do not.’ (Chomsky, 1965, p. 42). Ultimately any notion of a simplicity metric was dropped from syntactic theory, because little progress seemed to be being made in understanding grammar selection in this way.

Interestingly however, Chomsky’s (1965) theory shows that simplicity metrics are not necessarily incompatible with theories which postulate very strong innate constraints on grammar. It seems that even within a parameter setting model of language acquisition, statistical inferences would make the task of learning much easier, especially given the presence of noise in the data from which people learn (due primarily to grammatical errors, and exposure to data from children who have not mastered certain aspects of grammar). In fact, the use of minimum description length within a parameter setting framework was modelled by Briscoe (1999), who used it as part of a model of phylogenetic language evolution. However, in the light of the results reported here, the use of minimum description length together with Universal Grammar may seem odd, because it would seem that if minimum description length is used to learn languages, then there is no need for Universal Grammar. The arguments supporting the existence of Universal Grammar are primarily based on the claim that languages would not be learnable without it, so, if it can be shown that this is not the case, it removes the main piece of evidence supporting the theory of Universal Grammar. However, having said this, showing that Bayesian inference can be useful in explaining language acquisition does not necessarily mean that it is actually used. Essentially it allows us to return the degree to which language is determined by innate

principles of grammar to an empirical question, allowing the possibility that there is a much greater degree of learning in the process of syntactic acquisition than had previously been thought.

However, postulating that a Bayesian mechanism is used in acquiring syntax, results in very different predictions about what form syntactic knowledge takes, than would be the case if we presume that language is largely determined by universal principles. Chomsky has argued that the language faculty of the mind should satisfy 'general conditions of conceptual naturalness that have some independent plausibility, namely, simplicity, economy, symmetry, nonredundancy, and the like' (Chomsky, 1995, p. 1). While Chomsky notes this is 'a surprising property of a biological system' (Chomsky, 1995, p. 5) he argues that this view is justified because, throughout the history of syntactic research, systems conforming to this kind of principle have turned out to be the right ones. However, if language is learned with a Bayesian system we would not expect it to conform to such principles. Grammars could contain a lot of irregular rules if these accounted well for regularities in observed language. Even the principle of lexical minimization is not so clear cut within a Bayesian based account of learning, as Bayesian metrics will favour grammars which associate a lot of information with individual words if this allows them to account better for regularities in the data. Hence, one prediction of Bayesian theory is that the most commonly occurring words may be very idiosyncratic and irregular in their behaviour, while very rare ones must conform to regular patterns.

It is interesting to compare the Bayesian account of acquisition of subcategorizations presented here to Pinker's (1989) theory. Pinker's theory predicts that universal innate principles relate the meaning of a word to its syntactic subcategorization. Instead of

the syntactic subcategorization of a verb being determined empirically by a learner based on observations of patterns of occurrence, it is determined by the meaning of that verb. Certainly Gropen et al (1989) have shown that children are sensitive to correlations between semantic and phonological characteristics of verbs, and which subcategorization frames they are most likely to occur in. Researchers including Mazurkewich and White (1984) had already proposed that children use these correlations to determine verbs' subcategorizations. However, it is quite possible that these patterns were learned by the child in much the same way as we have proposed that syntactic subcategorizations may be learned.

One way of resolving the question of how people actually do learn these verbs, and hence of evaluating more directly the correctness of the model, would be to perform psycholinguistic experiments in which participants were taught novel verbs. The participants could be introduced to novel, nonce, verbs, which would be incorporated into passages of text. The texts could either be presented to the participants orally, or given to them to read. As people continue to learn new vocabulary throughout their lives, it should be unproblematic to conduct such experiments with adult participants. The novel verbs could variously either occur a single time, in either the double object or the prepositional dative construction, or could appear multiple times, in either just one, or both of these conditions. Furthermore, the verbs could be chosen to provide the semantic and phonological cues that Mazurkewich and White (1984) and Pinker (1989) have identified as being indicative of verbs which either follow the alternation, or which do not alternate. The subjects would then be required to perform some exercise that would cue the production of the verbs, such as being asked questions that would normally be expected to produce an answer containing the verb. It would then be possible to see what subcategorizations the participants had learned for each

of the verbs, and to determine whether the participants had learned to alternate any of the verbs, so that they used them in both constructions.

If some of the novel verbs were of the semantic and phonological type that normally alternates, but appeared only in the prepositional dative construction, it would be possible to determine which cues ultimately take precedence, at least if sufficient examples of the verb had been provided during the experiment. It would also be possible to determine whether participants would learn to use verbs in only the prepositional dative construction if they were presented in both constructions, but their semantic and phonological cues indicated that they should occur only in the prepositional construction. Such experiments ought to be able to determine conclusively whether the semantic and phonological cues are the ultimate determinant of the subcategorization of verbs, or whether they just provide probabilistic cues, that can be overridden given sufficient distributional data. It could even be possible to investigate whether subjects could be cued to learn new rules relating semantic and phonological cues to verb subcategorizations. This might be possible if examples were given of a sufficient number of verbs for which there was a correlation between semantic and phonological properties, and subcategorization.

The main limitation of the computational model described here is that it can only learn from small artificial data sets. There is no reason in principle why it cannot operate on naturally occurring language; it is simply that it would take an extremely long time to run on this kind of corpus. This is clearly a limitation that is shared with connectionist approaches, though Redington et al (1998) and Grünwald (1994) demonstrated impressive results learning from real language corpora. This is probably the main criticism that is likely to made of this approach, especially by researchers

aiming to defend the universal grammar hypothesis. However, the primary evidence in support of the universal grammar hypothesis has been that languages are not learnable, due to the poverty of the stimulus. This is a different issue to whether or not a search mechanism can be found that will be able to identify the correct grammar. As the amount of data increases, it becomes necessary to consider larger and more complex grammars, and so there is a combinatorial explosion in the number of possible candidate grammars. In these circumstances, a universal grammar could certainly help to reduce the size of the search space.

It seems that there are really two separate learnability problems. Firstly, it is unclear whether there is sufficient information in the input data available to language learners for them to identify the correct grammar without strong innate constraints being available to them to rule out some of the candidate grammars. Secondly, it is unclear whether it is possible to complete a search through sufficient of the candidate grammars for a correct (or nearly correct) one to be identified in the amount of time that children take to acquire language, without strong innate constraints being available to them to constrain the search. The work of this chapter has gone some way to showing that there is sufficient information in language to enable acquisition (although it has only shown this as regards the acquisition of one very small aspect of language, so there could well be others which such a method is not able to acquire). However, this chapter has done relatively little to address the second learnability problem, that of combinatorial explosion[94]. In order to show that that problem can be overcome by such a method, it would be necessary to demonstrate that a much more

[94] This problem is, however, discussed at more length in Dowman, 1998.

complex grammar could be learned, covering a much larger range of language than that investigated in this chapter.

A task for further research will be to investigate ways in which the search procedure could be made more efficient, so that learning from more realistic corpora becomes possible. It seems worth acknowledging, however, that we are modelling a process which takes place over many years, and that the human brain, while operating in a very different way from man-made computers, probably has a much greater overall processing capacity. This suggests that it may not be possible to learn the syntactic systems of real languages in their full complexity on a present day computer.

The work described here has not been the only research which has investigated the use of minimum description length as a psychological model for learning verb subcategorizations. Onnis et al (2002) used minimum description length to try to account for the way in which we learn regular linguistic rules, whilst also learning that there are exceptions to those rules. He looked at a similar verb alternation to that investigated in this chapter, the alternation in English between intransitive verbs, transitive verbs which must take an object, and verbs which can appear in either of these forms. An example of the first type of verb would be *arrive*. We can say *The train arrived* but not **John arrived the train*. However this contrasts with verbs of the second type which must take an object, such as *cut*. We can say *John cut the string* but not **The string cut* or **John cut*. These two types of verb both contrast with verbs of the third type, such as *bounce*. We can say both *John bounced the ball* and *The ball bounced*.

Onnis et al simulated this pattern by creating a very simple artificial language which contained only two word classes, nouns (*N*) and verbs (*V*). Sentences in this language

could then take one of two forms, either *NV* or *VN*. (*NN* and *VV* were both disallowed.) They created 36 verbs and 36 nouns. 16 of the verbs could occur in either the *NV* or the *VN* construction, but 10 of them could only occur in the *NV* construction, and another 10 in only the *VN* one. This simulates the situation in English as regards transitivity, in which some verbs can occur in both transitive and intransitive constructions, but in which others are restricted to occurring in only one of these forms. The actual number of verbs of each type in the artificial language was based on statistics extracted from the CHILDES English corpora (MacWhinney, 2000), and reflected the number of verbs used exclusively in an intransitive or transitive context, or in both contexts, in adult speech in the corpus.

Onnis et al devised a coding scheme, which could code sentences in this language, but which could optionally record that some verbs were 'exceptional' and so could not occur in one of the two constructions. Recording these exceptions had a cost in terms of coding length, and so including this information would only result in a lower overall coding length if the corpus contained a sufficient number of occurrences of these exceptional verbs that a significant gain in terms of reduced coding length would result from the knowledge that they can only occur in one of the two constructions. Onnis et al showed that when less than 16,000 sentences had been observed, a shorter coding length was achieved when no distinction was made between the verbs which could occur in both constructions and those which were restricted to only occurring in one construction. However, when a slightly greater number of sentences had been observed, recording the restrictions on exceptional verbs did result in shorter overall coding lengths. In common with the research reported here, Onnis et al argued that their model explained why children at first over-generalize alternations, extending their scope to verbs to which the alternation does

not apply, yet later are able to recover from this overgeneralization, once they have observed a greater number of verbs.

Onnis et al's model is very similar to the one presented here in that both models aim to explain the acquisition of verb alternations from positive evidence, and thus resolve the problem of how children know that some verbs cannot occur in particular constructions. Both models use minimum description length as their learning mechanisms, and both explain early overgeneralization followed by later learning of exceptions in the same way; that is that with only a few examples of a verb, there is insufficient benefit in terms of shorter coding length to justify recording the exceptions. However the way in which the models represent grammars is completely different. The model described here is much more general, in that it uses context free phrase structure grammars, and so can learn a wide variety of linguistic structures. Onnis et al's model is restricted to coding two word languages, and the differences between the grammars in his model consisted only of lists of which words fitted into each of the two exceptional categories.

While the model described here includes a search mechanism for finding grammars as well as a coding scheme, Onnis et al did not provide a search mechanism which would determine which verbs were exceptional and which were not, but instead simply compared the case where no exceptions were recorded to the one in which all the exceptions were recorded. Hence Onnis et al's model was not able to learn as such, but instead simply evaluated two grammars, and decided which was best. It would of course be possible to add a search strategy to Onnis et al's model, and perhaps the lack of one should not be thought of as a deficiency, especially when compared to the model presented here, because no claim is made that the search

strategy of the present model closely reflects the mechanism which children use to find possible grammars. There does, however, remain the problem with Onnis et al's model that it has only compared two possible grammars, and so it is not possible to be certain that another grammar would not have received a better evaluation than either of the two considered for any particular corpus. Because the range of allowable grammars is so restricted, this is probably not the case, but this kind of phenomenon commonly occurred in experiments using the model presented in this chapter. (For example there is a grammar which both specifies the verb subcategorizations correctly, and assigns correct structures, but this grammar has a worse evaluation than the one actually learned. However, if a comparison had simply been made between this grammar, a similar one which classed *sent* as irregular, and the initial grammar, the correct grammar would probably have received the lowest evaluation, and so we could have been tempted to argue that the model had shown that it could learn not only the subcategorizations correctly, but also the correct structures, which is clearly not the case.)

That Onnis et al's model is simpler than the one presented here could in itself be considered an advantage. If two models explain the same data then it would seem that the simpler one would be preferable. However, the model presented here is able to learn a lot more generally than Onnis et al's model, as it can learn aspects of syntax besides subcategorizations, so this direct comparison is not completely applicable. However, a clear advantage of Onnis et al's work is that the artificial language used there is modelled closely on empirical data derived from a corpus, so the actual frequencies of each type of verb would be more realistic. Onnis et al also used a much larger number of verbs than were used in the research presented in this chapter, so, in this way, their model more closely reflects the real process of language acquisition.

The model presented here could not cope with such a large number of verbs because the search mechanism used is very slow, even when learning on only the four verb corpus. However there should be no difficulty in simply evaluating grammars able to parse languages containing significantly greater numbers of verbs, as the time needed to evaluate a grammar for a corpus is much much less than that needed to learn a grammar from a corpus. It would seem that if the model described here were presented with Onnis et al's corpus, then it would make a similar prediction, in that with only a small corpus it would predict that all verbs followed the alternation, but that with a bigger corpus it would learn the exceptional cases of verbs which do not alternate. However, it would be interesting to investigate just how large a corpus was needed to achieve such a result. While Onnis et al's model did not prefer grammars which accounted for the exceptional verbs until it had observed over 16,000 sentences, it would seem likely that the model described here would do so when it had seen many fewer sentences, because it was able to learn the distinction between verbs which can appear in the double object dative construction and those which do not from a corpus of only 150 sentences.

It is worth noting that, in common with the model described here, Onnis et al's model made no attempt to use non-distributional cues, such as semantic or phonological evidence, to determine the correct subcategory for each verb, despite there being evidence which suggests that people do make use of such evidence (Gropen et al, 1989). However Allen (1997) showed how semantic cues could play a part in acquisition of the rules governing the mapping of verb arguments to semantic roles, using a neural network model. The architecture of his model is shown in Figure 9.2.

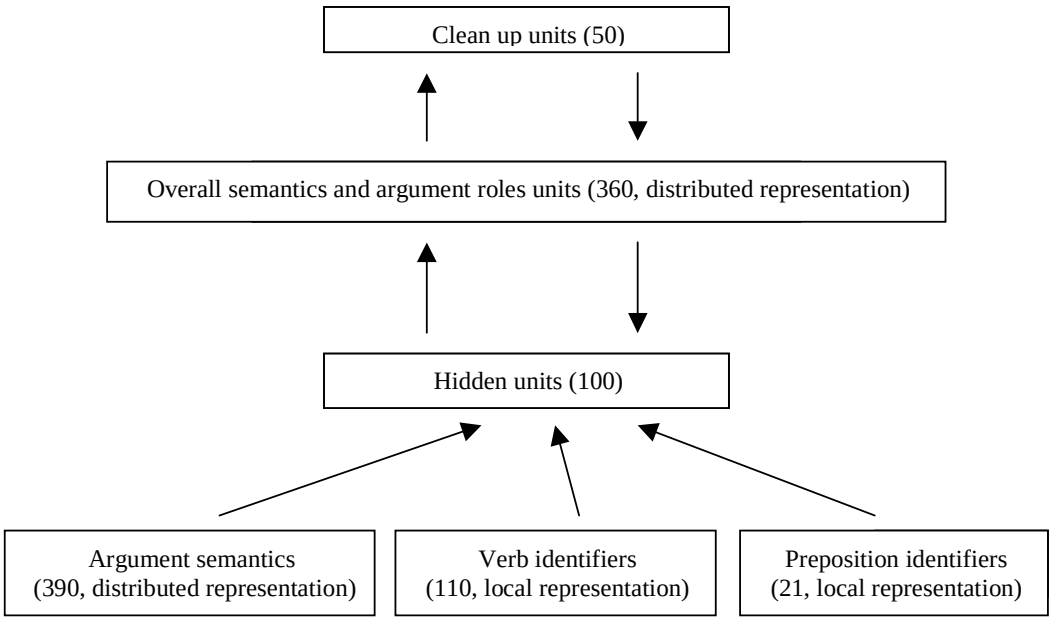


Figure 9.2. Architecture of Allen's Model.

(The numbers in brackets show the number of each type of node.)

The network was trained based on child-directed speech occurring in the CHILDES corpora (MacWhinney, 2000). The utterances containing the 110 most frequent main verbs were found, and the verbs together with any associated preposition were then presented to the network together with semantic representations of their nominal arguments, an overall semantic representation of each utterance, and a representation of the thematic roles filled by the nominal arguments. Verbs and prepositions were represented with a localist scheme, so that there was one specific node corresponding to each verb, and one for each of the 21 prepositions found in the corpus. The nominal arguments were represented as a pattern of activations across all 390 nodes. The nodes corresponded to WordNet features (Miller et al, 1990; Miller, 1990), so they coded elements such as *+human*, *+male*, and *–vehicle*. The coding referred to the semantics of the words, not the words themselves, so two different words which had similar meanings might receive the same encoding. For example, the two proper

nouns *John* and *Peter* would have identical feature assignments. The overall semantics of the sentence were coded in a similar way, but the features used mainly related to the verb's meaning. For example, an utterance including the verb *eat* would include the features *+act, +cause* and *–communicate*. Finally, the argument roles fulfilled by the nominal arguments corresponded to traditional thematic roles (such as *patient*, *experiencer*, and *instrument*), supplemented with nodes for features specifying more specific aspects of the role concerned.

Only one nominal argument and its corresponding thematic role could be presented to the network at one time, so if an utterance contained more than one nominal argument, each was presented in turn, in the order in which they occurred in the utterance, and this sequence was then repeated two more times. The verb, preposition, and overall semantic arguments were presented continuously throughout this period. The network was trained using back-propagation, with 1,200 distinct utterances, for 100,000 iterations.

Testing of the model involved presenting it with novel grammatical and ungrammatical utterances, and comparing the results. The network was judged to have 'accepted a novel sentence if it computes a role for all and only the number of arguments in the novel sentence.' (p. 303). Allen was able to demonstrate that the model was able to generalize to use verbs such as *kick* in the double object dative construction, even though it had not been presented with *kick* in a double object dative sentence in the training data. The model was also able to reject as ungrammatical the sentence *John put the book*, because this resulted in the model assigning three thematic roles, even though the original sentence contained only two noun phrases. However the model rejected *John carried Mary the basket*, even though this sentence

would be grammatical for at least some speakers of English. Allen explained that the model allowed *kick*, but not *carry* to appear in the double object dative construction because '*kick* is closer [to other verbs which show the alternation] along semantic dimensions which are *relevant* to getting this set of argument role assignments' (Allen, 1997, p303). More specifically, Allen states that the alternation is activated by verbs for which the *+instantaneous force* feature is activated, which would clearly include *kick* but not *carry*. We should note, however, that in attempts to replicate Allen's model, Smith (1999) concluded that the rules which the model learns are complex, and in many cases very idiosyncratic, and so cannot generally be given a concise explanation in terms of, for example, the presence or absence of a single semantic feature. This suggests that there may be no coherent grammar underlying languages at all, and that a coherent explanation of languages at a level above the neural one *may* not be possible.

It is interesting to make a more specific comparison between Allen's model and the one presented here. The program of this chapter relied purely on distributional cues to learn verb subcategorizations, and, while it clearly showed how learning can be achieved using distributional cues alone, it did not demonstrate that children do not make use of other available cues, in particular the semantic and phonological cues identified by Mazurkewich and White (1984) and Pinker (1989). Allen's model did not make use of phonological cues either, but semantic cues were central to its learning mechanism. However, this did not mean that Allen's model did not make use of distributional cues; it had to use these cues as well, because it did not have any *a priori* knowledge of how semantics related to subcategorizations. The rules it used to determine subcategorizations from verb semantics were learned based on correlations between semantics and syntax in the example utterances.

This way of learning actually has a lot of similarities to Pinker's (1989) proposal, in that Pinker also proposed that children learn correlations between verbs' semantics and their subcategorizations, and that they then use these correlations as the basis of general rules which predict the allowable subcategorizations for individual verbs. The key difference is that Pinker proposed that children learn formal rules, and that the general form of those rules is determined innately. However, he did not provide an explanation of how those rules were learned. In contrast, in Allen's model, all such rules are implicit in the connection strengths in the neural network, and the model learns the rules linking semantics and subcategorizations at the same time as learning which semantic features are relevant for placing verbs in particular categories. The model does not just use semantic bootstrapping (where semantic cues are used to learn syntax, as Pinker (1989) proposed), but can also perform syntactic bootstrapping (where syntax can be used to help predict word meaning, as Landau and Gleitman (1985) proposed). It can use the correlations between syntax and semantics which it has learned so that if it is given a nonce verb, it is able to guess appropriate verbal semantic features for the verb, and thematic roles for its arguments.

However, Allen did not address one of the basic questions addressed in this chapter – namely why do children overgeneralize verb alternations, and then subsequently recover from those overgeneralizations. Allen did not report any investigations concerning the predictions of his model as regards the course of learning, and hence there is no indication of whether his model replicates the pattern of early overgeneralization followed by later acquisition of correct subcategorizations. Allen did demonstrate correct acquisition of the rule which allows generalizations to be made concerning which verbs can appear in the double object dative construction, and he showed how multiple cues to syntactic structures could be integrated. However, he

did not report whether his model was able to reliably infer that verbs appearing in the prepositional dative construction can also appear in the double object dative construction.

Another approach to computational modelling of the acquisition of verb subcategorizations was that of Brent (1994). The task attempted by his system was in many ways more complex than that attempted here and by Onnis et al's (2002) system, because Brent's model learned from transcriptions of child directed speech obtained from the CHILDES corpora (MacWhinney, 2000), instead of learning from artificially created data sets. Central to Brent's model is the hypothesis that children learn a considerable number of function morphemes and proper names at a very young age, and then use these morphemes to aid in the acquisition of verb subcategorizations. This hypothesis seems reasonable, because in all languages there exist a small number of function morphemes that occur with very high frequency, and so it would seem likely that children learn these morphemes early. It is widely reported that function morphemes are often absent from the speech of young children, but research has shown that children learn to recognise many of these words from a young age, even thought they might not initially use them productively (Hirsh-Pasek and Golinkoff, 1996). Children will also be exposed to a limited number of proper names, each of which they will observe with a fairly high frequency, so again it seems likely that these words would be acquired at a young age.

A problem remains, however, which is that children cannot learn verb subcategorizations until they have first identified the verbs. Brent's model identified verbs as those words which occurred in plain form and with an *–ing* suffix, except that he excluded these words when they immediately followed a preposition or determiner.

He then looked at all the instances of each verb found, and tried to identify phrases which were complements to the verb, and to determine which kind of phrase they were. In general, if a proper noun, pronoun or determiner followed immediately after a verb, this would be taken to be a potential noun phrase complement, while if a preposition and then one of these elements followed a verb this would be taken as a potential prepositional phrase complement. Using this kind of rule, Brent's model classified each of the candidate verbs according to whether it occurred in one of several types of subcategorization frames used with English verbs, such as those requiring just a single noun phrase complement, a finite clause complement, or the double object or prepositional dative subcategorizations discussed above.

The model now had a list of verbs, and for each would have a table which recorded both the frequency of the verb, and how often it seemed to have occurred in each of the subcategorization frames. This table would be expected to contain some errors, primarily because the cues used to identify subcategorizations were only clues, and could not be relied upon to always identify the underlying linguistic structure correctly. (For example, prepositional phrases following verbs in English are frequently adjuncts, which are not subcategorized for by the verb.) For each verb, a statistical rule was then used to decide in which subcategorization frames it could reliably be said to occur. If a verb occurred in a particular subcategorization frame only rarely, relative to the verb's overall frequency, then that was not considered to provide sufficient evidence to establish that that was an allowable subcategorization for that verb.

Using this procedure, Brent's model was able to assign one or more subcategorization frame to 76 of the 126 verbs found, and it generally did so with a high degree of

accuracy. There were seven verbs which show the dative alternation, *give, show, bring, feed, roll, throw* and *read*, but the alternation was only learned for the first two of these. (That is, only those verbs were classified as taking both the double object and prepositional dative subcategorizations.) This is presumably because of the fairly limited amount of data available. The prepositional dative structure was, however, learned for *bring, roll* and *throw*, and the double object dative structure for *read*.

A key result of the model reported in this chapter is that it was able to demonstrate generalization of the dative alternation from one verb to others, but Brent's system was unable to make any such generalization, because it treated each verb individually. It is clear that children do sometimes make such generalizations, so ideally Brent's system would be extended so that it too could do so.

Another problem with Brent's system is that it is specific to English, and would be completely unable to learn verb subcategorizations in any other language. The *a priori* incorporation of some aspects of English, such as knowledge of functional morphemes, can be justified by arguing that these are acquired at a stage of learning which must take place before the stage simulated by the model, but the justification for other parts of the model is less clear. The model had *a priori* knowledge of the types of element for which English verbs subcategorize, even though it was the acquisition of subcategorization frames for which the model aimed to account. Furthermore, much knowledge of English is implicit in the various tests for complement phrases, so it seems that the model is provided with a considerable amount of knowledge concerning English verb subcategorizations prior to its exposure to the data. This is a problem, because a computer model of acquisition must

learn from only those information sources which are available to children, if it to accurately account for the acquisition process.

Ideally a computer model of language should be equally able to learn any human language, as all normal children are able to do this, and it is they that such computer models are supposed to mimic. Clearly Brent's model gives an insight into the kind of cues that children might use to learn verb subcategorizations, but it leaves unanswered what may be a more difficult question – that of how children learn to identify the relevant cues and the appropriate rules for applying them. Given the wide cross-linguistic diversity in the ways in which verbal arguments are expressed, it would seem that this is likely to be a very difficult task.

Brent's (1994) system had many similarities to a system developed by Manning (1993) which could produce a verb subcategorization dictionary by analysing text corpora. Manning's system also incorporated considerable *a priori* knowledge of English, in the form of a finite state parser and stochastic tagger used by the system. Such factors would make Manning's system a poor psychological model of language acquisition, but this is hardly surprising, because Manning's system did not aim to model language acquisition. Its aim was simply to provide a tool which could be useful in the development and maintenance of language technology systems.

There has been a limited amount of work which has combined the minimum description length approach to grammar learning with the expression-induction methodology described above. In particular, the work of Kirby (2002) discussed in section 2.4 above has served as the basis for work which has replaced the simple learning mechanism used there with a more sophisticated one based on minimum description length. Teal and Taylor (2000) created a system which could represent

languages, whose sentences consisted of strings of six letters, using finite state automata[95]. The system would then produce a random sample of sentences to pass on to the next generation. Another automaton would then be constructed by incorporating those sentences into an initially empty automaton, so that that automaton would allow all and only the observed sentences. Nodes in the new automaton were then merged if that made an improvement to the automaton according to a minimum description length evaluation measure. The resulting automaton was then used to produce a sample of sentences from which the automaton in the next generation could learn. They showed how their system could model change in language over several generations, and concluded that simpler languages change more slowly than more complex ones (at least within the bounds of their simulation).

Brighton and Kirby (2001) and Brighton (2002) created a similar model which also used finite state automata and learned in a very similar way, again by merging nodes under the guidance of a minimum description length evaluation measure. Their system was more sophisticated however, because, like Kirby's (2002) system, it could express meanings using the languages it learned, and did not just reproduce syntactic patterns like Teal and Taylor's system did. The system incorporated a meaning space, defining the range of possible meanings which the system could express. A range of meanings would be expressed using one of the languages under investigation, and the resulting sentence-meaning pairs would be observed, and incorporated into a new

[95] Finite state automata are a simpler kind of grammar than context free phrase structure grammars. They consist of a number of nodes linked by arcs which are labelled with words or symbols, and they define allowable sentences as those which lie along a series of arcs starting at a particular start node. See Chomsky (1957) for a discussion of their use in linguistic theory.

automaton, which would then be generalized using a minimum description length evaluation measure, in a similar way to that in which this was done in Teal and Taylor's system. However, in contrast to Teal and Taylor's and Kirby's approaches, the new automaton did not express a range of meanings to a new generation, because no attempt was made to simulate evolution over several generations. Instead a calculation was made to determine the proportion of the meanings expressible in the original language that the new automaton could express. This provided a measure of 'language stability', because it corresponded to how accurately the language had been passed on to the next generation. Brighton and Kirby calculated these values for both compositional and non-compositional languages under a range of conditions, so that they could determine in which situations there would be a selective advantage for compositional languages. They concluded that compositional languages are more stable than non-compositional ones when there is a sufficiently large meaning space, and only a small proportion of the meanings expressible in a language are ever heard by people learning the language. These are the conditions under which human languages evolve, and Brighton and Kirby's model hence supports the hypothesis that syntax emerges due to communicative pressures acting over several generations.

9.5 Conclusion

This chapter has shown that Bayesian inference (in the form of minimum description length inference) is able to provide a simple and plausible account of how a number of aspects of syntax could be learned. In particular the computational model described here can learn verb subcategorizations where one verb is grammatical in only a subset of the structures in which another can appear, and yet predicts that newly encountered verbs are used productively in regular patterns. The model also accounts for overgeneralization, and hence the use of irregular items in regular constructions

during early stages of acquisition. While it is not logically necessary that children must make use of Bayesian inference in learning language, there is potential to incorporate Bayesian inference into theories as diverse as recurrent neural networks and Universal Grammar.

Chapter 10

The Nature of Language

This thesis has so far attempted to address specific issues within linguistics, relating to colour terms and to syntactic acquisition. However, a further contribution of the thesis is that it has examined several different ways of accounting for linguistic phenomena, which should help to address the issue of what is the best way to understand language. This relates closely to the question of what concepts of language are appropriate for use in linguistic research.

The part of the thesis on colour terms has suggested that we may not be able to explain linguistic typology if we simply view language as a psychological phenomenon. The typological patterns that were apparent in the colour term systems which emerged in the simulations cannot be explained just in terms of which kinds of system are learnable by the artificial people, and which are not. They are, however, clearly a product of the nature of the artificial people's conceptual colour spaces and the learning mechanism which they use. This contrasts with the view that colour term typology is the product of the uses to which language is put. (For example, it could be proposed that the types of colour term evolving cross-linguistically are those which are most useful to people communicatively.)

An interesting property of the evolutionary colour terms model is that it does not take any account of communicative success. The model could be regarded as incorporating functional pressures, because it required the artificial people to name colours, and a correlation between frequency of colour naming and complexity of colour language was demonstrable. Hence the results of the simulation were affected by how often people needed to use colour words, which is a kind of functional pressure. However, the function of language would have had a more direct effect on the simulations if people had been rewarded when they were able to communicate successfully, and penalised when they failed to do so. In the colour terms models, no attempt was ever made to see if the hearer of any utterance would be able to deduce the colour being referred to, or whether they would be able to discriminate it from any other colour[96]. Hence, clearly the emergent languages were not shaped by feedback coming from the people who heard the utterances. Therefore it seems that, while the languages emerging in the simulations were shaped by their function, they would not be adapted in such a way as to maximise their potential for achieving effective communication. This seems to place the model of colour term evolution somewhere in between the functionalist and formalist extremes of the functionalist-formalist debate (Newmeyer, 1998).

The typological patterns arise when many successive generations of artificial people attempt to learn the colour term systems used by the previous generations. This

[96] Interestingly, Belpaeme (2002) incorporated a discriminability test into his model, in which it was investigated whether or not the term that had been used had allowed the hearer to distinguish between the target colour and a neighbouring context colour. The hearer's colour lexicon would then be modified, based on whether communication had been successful (see section 2.4. for more details).

suggests that the concept of *meme*, which was proposed by Dawkins (1976), may be useful in understanding the simulations, and, by analogy, language in general. Dawkins proposed that we can think of any information which is transmitted between people as a *meme* (named on analogy to *gene* in genetics). The memes which come to be known by people will be those which are passed on from generation to generation. Dawkins noted that whether memes are useful to the people who learn them will influence whether they are passed on, but that memes which are not useful might well be passed on as well, simply because of some property they might have which aids in their transmission. For example, Dawkins suggested that the idea of *blind faith*, that is that religious beliefs should be held unquestioningly, and not justified by rational inquiry, is self-perpetuating (as it itself forms a part of some religious belief systems). Hence, perhaps the reason for its continued existence is that it encourages people to go on believing that it is true, and discourages them from questioning it. This would tend to result in it surviving in the belief systems of individual people, therefore increasing the chances that they would pass the idea on to other people. We should note that this argument is valid, regardless of whether the idea of blind faith itself is beneficial or detrimental to the person who believes in it.

The colour words and their denotations within the colour model can also be viewed as memes, and this generalisation may be extended to other aspects of language which are learned based on input from other people. While the simulations do not take account of whether colour terms are useful to people, if colour words were not useful to people, then they would never use them, and so the colour words would never be transmitted to the next generation of learners, and so they would be lost from the language. Hence, it would seem that colour term systems will only exist in languages if the speakers of those languages have cause to refer to colours. However, showing

that it is necessary for people to have reason to use colour terms, in order for them to exist in a language, does not necessitate that the colour terms which emerge will be those that are most useful.

The evolutionary colour terms model might be best understood in terms of a proposal made by Kirby (2000), which is closely related to Dawkins' concept of memes. Kirby suggested that it may be best to view languages as independently evolving adaptive systems, which use people to aid their transmission. We can then consider languages to be collections of memes which exist in an environment which consists of a chain of human hosts[97], and languages will come to adapt so as to maximise their potential for survival in this environment. Given this perspective, it would seem that languages may exist simply because there is a suitable medium available in which they can survive (human hosts), and a replication mechanism through which they can propagate themselves (transmission via E-language[98] and an acquisition mechanism). Languages would, however, appear to be beneficial to their hosts, and so they should probably be viewed as symbionts[99].

However, the model of colour term evolution does in fact suggest that some aspects of language may not be useful to their hosts, because the absence of any measure of

---

[97] Of course languages can also exist in some other forms, for example on paper (in the form of writing), or on audio tapes, but people are probably the most important kind of host for languages.

[98] E-language was defined in section 2.2.

[99] A *symbiont* is an organism involved in a *symbiotic* relationship (a relationship in which two or more organisms live together or interact, and from which they both benefit).

communicative success within the model shows that the evolution of particular language forms need not be dependent on their usefulness. This leaves open the possibility of completely useless words or constructions entering the language, simply because they tend to be learned and passed on between generations. Any such aspects of language should be regarded as parasites, and not symbionts. Clearly, in the evolutionary model of colour terms, the colour term systems only existed because the people used colour terms, which would imply that the colour terms fulfilled some purpose for those people. However, there is no reason to suppose that all the emergent colour words would be useful to a real user of the language. It is quite possible that people might in reality be better off with fewer colour terms, in which case the terms which are not beneficial would be parasites. This suggests that many aspects of language, such as irregular constructions, may well be parasitic, as they would appear to make language needlessly complex[100]. Viewing languages as partly parasitic phenomena, would help to explain why they seem to be needlessly complex, and why they contain so many irregularities which seem to complicate the task of communicating, without having any benefit for language users.

It does, however, seem sensible to point out some problems with the concept of memes, at least in so far as it is applied to language. Firstly, people are not simply passive repositories through which language may pass, but are quite capable of modifying language, and creating entirely new expressions, both consciously and

[100] We should note that there can also be benefits arising from irregularity. If, for example, a particular complex meaning is expressed very frequently, we might expect it to be lexicalized (Hurford, 2000), which would reduce the overall number of morphemes needed in communication, which would be beneficial for both speakers and hearers.

unconsciously. The artificial people in the evolutionary simulation simply tried to mimic other speakers, which is probably what real people do most of the time, but they need not always do so. However, Pinker (1997) has argued that complex memes arise ‘because some person knuckles down, racks his brain, musters his ingenuity, and composes or writes or paints or invents something’ (p209), and not as a result of an evolutionary process involving cumulative copying errors, and the differential survival of memes. Certainly Pinker is to some extent right, as many kinds of idea could presumably only be created as a result of rational thinking and the application of foresight. (Think, for example, of the design of an aeroplane. It would seem that it would only be possible to design a working aeroplane by understanding the principles by which aeroplanes work. A working design for an aeroplane could never arise simply as the result of a memetic process.) However, I think that Pinker is incorrect when he dismisses the idea of memetic evolution so completely, especially in relation to language. Languages are shared systems of conventions, and so words and constructions can only become established if they are adopted by large numbers of speakers. Pinker argued that ‘when ideas are passed around, they aren’t merely copied with occasional typographical errors; they are evaluated, discussed, improved on, or rejected.’ (p210). This is probably true for some kinds of idea, but it would seem likely that language is transmitted largely as a result of mimicry, rather than as a result of any more directed cognitive process. This is exactly the kind of process to which meme theory is applicable.

A further problem with the concept of memes, is that they are not passed on intact to the next generation in the same way that genes are, because the copy of a meme may

not be exactly the same as the original from which it was learned[101]. (While genes do occasionally mutate, in the vast majority of cases, an identical copy is passed on to the next generation (Dawkins, 1976).) In the simulations, the representations of the colour terms themselves were passed on unchanged, which probably reflects the situation in real languages, as usually a child would use exactly the same phonemes to represent a word as the person who they learned it from did[102]. However, the same cannot be said about the words' denotations, as each artificial person would normally learn a slightly different denotation for each word, depending on exactly what examples of its denotation they had observed. This is probably also what happens when real people learn colour terms, and most other kinds of word. Some words, however, have a precise meaning, and so we might argue that the memes corresponding to their meanings are replicated perfectly when those words are acquired. Examples of such words might include determiners such as *both*, or nouns such as *Tuesday*. When a

---

[101] It should be noted that Dawkins (1976) did acknowledge and discuss this problem.

[102] This is not to say that they would pronounce the word identically, as even if two people pronounced a word in very similar ways, we could still expect that there would be some difference between the pronunciations of each person, no matter how small. However, phonemes are not speech sounds, but more abstract representations, which code the underlying distinctions between sounds. (We can view phonology as a digital system, in contrast to the speech sounds themselves which are analogue.) For example, English has both /l/ and /r/ phonemes, each of which has a distinct pronunciation. We would expect a child learning English to maintain this distinction, so that he or she would pronounce /l/ differently to /r/, and hence a word such as *lob* would be distinguishable from *rob*. This distinction would be maintained even if the child's exact pronunciation of each phoneme differed slightly from that of other people. It is the representation of words at this level of distinctive contrasts that we would expect to be passed on unchanged, rather than the speech sounds themselves.

meaning is passed on unchanged between generations, it is unproblematic to consider it to be a meme, but whether the meanings of words such as colour terms should really be thought of as memes is somewhat questionable. Perhaps a modified version of meme theory is needed to cope with situations like this, in which there is fuzzy transmission of ideas, but it seems that the concept of meme is never-the-less useful in understanding many aspects of language.

While meme theory may be useful in explaining the properties of colour term systems, it is largely inconsistent with Chomsky's view of language. Meme theory only applies in situations in which there is some sort of learning device which can learn concepts, or which can learn to reproduce some form of behaviour. However, Chomsky's theory, which proposes that the language people acquire is determined largely by Universal Grammar, leaves little room for learning. Language forms appear in successive generations primarily because they are specified by Universal Grammar, and learning plays only a relatively minor role in choosing which of a limited range of devices is used to express particular meanings. We could still regard language as being memetic, but only in those aspects which are determined based on exposure to the E-language produced by other speakers. Hence it would seem that the meme concept will be of relatively little help in explaining linguistic phenomena, if the Universal Grammar hypothesis is correct.

If people are born possessing an innate Universal Grammar, then we should be able to explain most syntactic phenomena simply in terms of individual psychology, because it would be individual people's versions of Universal Grammar which are largely responsible for determining the syntactic system which they acquire. Hence, if Chomsky is correct about the mechanism by which syntax is learned, the focus on I-

language will be well justified. However, the model of syntactic acquisition described in Chapter 9, suggested that learning, and not Universal Grammar, could be the primary determinant of the syntactic systems which people acquire. Hence, if the model is accurate, and people do learn syntax with Bayesian inference (or with any other mechanism in which learning is more important than innate structure), then the concept of memes would be much more applicable, and I-language would appear to be too narrow a concept to explain linguistic phenomena. The syntactic systems of languages would be determined to a large extent by the nature of the arena of language use, and by cultural evolutionary processes taking place over several generations.

If syntactic structure is largely learned, and not determined by innate structure, then syntactic phenomena may be best explained as the products of evolutionary processes (Ellefson and Christiansen, 2000), in much the same way as the properties of colour terms were. If this is the case, typological patterns may reveal little about the underlying I-languages of individual people, because they may be due largely to process which operate above the level of individuals, and over longer time spans than an individual person's lifespan. Only some aspects of language may be explainable by reference to ontogenetic processes, while others may require a diachronic explanation, or one in terms of whole populations of speakers, not just individuals. For example, while the model explains how verb subcategorizations can be learned, it does not explain why we have two different subcategorizations for some verbs, but only a single one in other cases. It would seem that such phenomena cannot be explained with a program which just models individual people, because the determinant of whether a verb has one or two subcategorizations is the input that is received by the model. In reality this input would be produced by other people, and so the

subcategorizations of the verbs are determined based on input received from other speakers, and not as a result of any property of individuals (although clearly such phenomena can only exist in languages if people are capable of learning them). However, this suggests that the correlation between syntax and semantics evident in English verb subcategorizations may not have a simple psychological explanation, but may be a result of social processes involving whole communities of speakers[103]. Hence, we should be wary of approaches to this problem in which an attempt is made to explain all of the data within a purely psychological account.

Chomsky has proposed that we can make inferences about Universal Grammar based on typological regularities. However, if we had applied this methodology to the results of the simulations of colour term systems, then we would have had to postulate many principles of Universal Grammar which simply do not exist. (In the case of the simulations, we can be sure of this, because we know exactly how the artificial people were constructed.) For example, there is no aspect of the Bayesian learning mechanism which prevents the acquisition of *blue-red-yellow* colour terms. However, the results of the simulations would appear to suggest that this is the case, because no such term had been learned by the artificial people alive at the end of any of the

---

[103] This would appear to be likely, as most of the verbs which do not alternate are Latinate (originating from a romance language) and multisyllabic, while most of those which do alternate are phonologically native, and monosyllabic (Mazurkewich and White, 1984). This suggests that the distinction between non-alternating and alternating verbs originated when verbs were borrowed from romance languages, a process which clearly involves social interaction, and which is therefore not explainable at the level of individual psychology.

simulations[104]. Hence, if we were only able to observe the emergent colour term systems, and did not have knowledge of the mechanism which produced them, we might postulate that there is a principle of Universal Grammar which prevents *blue-red-yellow* colour terms from being acquired, which is clearly not the case (at least as far as the artificial people in the simulations are concerned). We should therefore be suspicious of the large body of work which makes claims about Universal Grammar based on language typology, as many processes not related to Universal Grammar could be responsible for the creation of typological patterns.

Overall, I think that it can be said that this thesis highlights how important it is to consider the theoretical status of regularities apparent in E-language. Most approaches to explaining colour term typology (including Berlin and Kay, 1969 and Kay and Maffi, 1999) have proposed an evolutionary account, but in general these accounts have been somewhat vague as to the exact details of the evolutionary process. By modelling communities of people using artificial people who learn from one another, we can create a model which incorporates both E-language and I-language, and which makes both the processes of language acquisition and language evolution explicit. If learning plays an important part in language acquisition, then it is not appropriate to neglect diachronic processes, but neither can we have an adequate model of language evolution which neglects the process through which children construct an I-language based on E-language input obtained from other speakers. It would seem that language

---

[104] Or at least no such term was learned by a majority of the people alive at the end of any of the simulations. Because the results presented above were based on analyses of communities overall, we cannot be sure that no individual artificial person had learned such a term.

function is likely to play an important role in shaping the form of languages, but that the mechanism through which this is achieved is somewhat indirect. We should be cautious about attributing properties of particular languages to either innate knowledge or to functional pressures, as they may in reality be a product of an interaction between both of these factors, or they may simply exist because they have an inherent survival value, and so are good memes.

References

Aarts, E. H. L. & Korst, J. (1989). *Simulated Annealing and Boltzman Machines: A Stochastic Approach to Combinatorial Optimisation and Neural Computing*. Chichester: Wiley.

Allen, J. (1997). Probabilistic Constraints in Acquisition. In A. Sorace, C. Heycock, and R. Shillcock (eds.) *Proceedings of the GALA '97 Conference on Language Acquisition.* Edinburgh: Human communication research centre, University of Edinburgh.

Anderson, J. R. and Matessa, M. (1991). An Incremental Bayesian Algorithm for Categorization. In D. H. Fisher, Jr., M. J. Pazzani, and P. Langley (eds.) *Concept Formation: Knowledge and Experience in Unsupervised Learning*. San Mateo, CA: Morgan Kaufmann Publishers Inc.

Andrieu, C., de Freitas, N., Doucet, A. & Jordan, M. I. (2003). An Introduction to MCMC for Machine Learning. *Machine Learning*, 50: 5-43.

Bailey, A. C. (2001). On the non-existence of blue-yellow and red-green color terms. *Studies in Language*, 25(2): 185–215.

Barnett, V. (1982). *Comparative Statistical Inference*, Second Edition. Chichester: John Wiley & Sons Ltd.

Baxter, R. A. (1996). *Minimum Message Length Inference: Theory and Applications*. Doctor of Philosophy Thesis, Monash University, Australia.

Bayes (1763). An Essay Towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions*, 53: 370-418.

Belletti, A. & Rizzi, L. (2002). Editors' Introduction: some concepts and issues in linguistic theory. Introduction to Chomsky (2002).

Belpaeme, Tony (2002). *Factors influencing the origins of color categories*. PhD Thesis, Artificial Intelligence Lab, Vrije Universiteit Brussel.

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

Bornstein, M. H. (1973). Color Vision and Color Naming: A Psychophysiological Hypothesis of Cultural Difference. *Psychological Bulletin*, 80(4): 257-285.

Boulton, D. M. (1975). *The Information Measure for Intrinsic Classification.* PhD Thesis, Monash University, Melbourne.

Boynton, R. M., and C. X. Olson. (1987). Locating Basic Colors in the OSA Space. *Color Research and Application*, 12(2): 94-105.

Brent, M. (1993). Minimal Generative Explanations: A Middle Ground between Neurons and Triggers. *Proceedings of the 15th Annual Conference of the Cognitive Science Society*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Brent, M. R. (1994). Surface cues and robust inference as a basis for the early acquisition of subcategorization frames. In L. Gleitman and B. Landau (Eds.) *The Acquisition of the Lexicon*. 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.

Brighton, H. (2002). Compositional Syntax from Cultural Transmission. *Artificial Life*, 8(1): 25-54.

Brighton, H. & Kirby, S. (2001). The Survival of the Smallest: Stability Conditions for the Cultural Evolution of Compositional Language. In J. Kelemen & P. Sosík (Eds.), *Advances in Artificial Life*. Berlin: Springer.

Briscoe, E. J. (1999). The Acquisition of Grammar in an Evolving Population of Language Agents. *Electronic Transactions on Artificial Intelligence*, 3(B):44-77.

Carroll, G. & Charniak, E. (1992). Two Experiments on Learning Probabilistic Dependency Grammars from Corpora. In *AAAI-92 Workshop Program: Statistically-Based NLP Techniques*.

Chater, N. (1999). The Search for Simplicity: A Fundamental Cognitive Principle? *The Quarterly Journal of Experimental Psychology*, 52A (2): 273-302.

Chen S. F. (1995). Bayesian Grammar Induction for Language Modeling. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton & Co.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Cambridge, MA: MIT Press.

Chomsky, N. (1972). *Language and Mind*. New York, NY: Harcourt Brace Jovanovich Inc.

Chomsky, N. (1984). *Lectures on Government and Binding: the Pisa Lectures*. 3rd Revised Edition. Dordrecht: Foris Publications.

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin and Use*. New York: Praeger.

Chomsky, N. (1995). *The minimalist program.* Cambridge, MA: MIT Press.

Chomsky, N. (2002). *On Nature and Language*. Cambridge: Cambridge University Press.

Christiansen, M. H. & Chater, N. (1994). Generalization and Connectionist Language Learning. *Mind and Language, 9*, 273-287.

Cleland, T. M. (1937). *A Practical Description of The Munsell Color System with Suggestions for its use*. Baltimore, MD: Munsell Color Co.

Conlan, F. (2002). Searching for the semantic boundaries of the Japanese color term 'ao'. Talk given at the 27th Annual Congress of the Applied Linguistics Association of Australia, 12-14 July, 2002, Macquarie University, Sydney, Australia.

Dawkins, R. (1976). *The Selfish Gene*. Oxford: Oxford University Press.

de Boer, B. (1999). Evolution and self-organisation in vowel systems. *Evolution of Communication* 3(1): 79–103.

de Marcken, C. (1996). *Unsupervised Language Acquisition*. PhD thesis, Massachusetts Institute of Technology.

de Saussure, F. (1959). *Course in General Linguistics*. Eds. C. Bally and A. Sechehaye in Collaboration with A. Reidlinger, trans. Wade Baskin. London: Peter Owen. (Originally published in French in 1916.)

De Valois, K. K. & De Valois, R. L. (2001). Color Vision. In N. J. Smelser and P. B. Baltes (Eds. in Chief), *International Encyclopedia of the Social and Behavioral Sciences*. Amsterdam: Elsevier.

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

Dowman, M. (1998). *A Cross-linguistic Computational Investigation of the Learnability of Syntactic, Morpho-syntactic, and Phonological Structure* (Research Paper EUCCS-RP-1998-6). Edinburgh, UK: Edinburgh University, Centre for Cognitive Science.

Dowman, M. (2000). Unsupervised Learning of Probabilistic Automata for Language Modelling. *Proceedings of Inteligent Systems and Applications 2000*. Wetaskiwin, Alberta, Canada: ICSC Academic Press.

Ellefson, M. R. & Christiansen, M. H. (2000). The evolution of subjacency without Universal Grammar: Evidence from artificial language learning. In *Proceedings of the 22nd 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.

Elman, J. L. (1993). Learning and Development in Neural Networks: The Importance of Starting Small. *Cognition, 48*, 71-99.

Elman, J.L, Bates, E.A., Johnson, H.A., Karmiloff-Smith, A., Parisi, D. & Plunkett, K. (1996). *Rethinking innateness: a connectionist perspective on development*. Cambridge, MA: MIT Press.

Fass, D. and Feldman, J. (2002). Categorization under complexity: a unified MDL account of human learning of regular and irregular categories. Presented at *Advances in Neural Information Processing Systems 2002*. Available on-line at http://ruccs.rutgers.edu/~jacob/papers.html. Downloaded on 12 May 2003.

Feldman, J. A., Gips, J., Horning, J. J., & Reder, S. (1969). *Grammatical Complexity and Inference* (Tech. Rep. CS 125). Stanford, CA: Stanford University: Computer Science Department.

Fitzgibbon, L. J., Dowe, D. L. and Allison, L. (2002). Univariate Polynomial Inference by Monte Carlo Message Length Approximation. *Proceedings of the Nineteenth International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann.

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

Francis, W. N. & Kucera, H. (1979). *Brown Corpus Manual*, revised and amplified edition. Published on line at http://www.hit.uib.no/icame/brown/bcm.html. Download on 28 May, 2003.

Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: MIT Press.

Gleason, H. A. (1961). *An Introduction to Descriptive Linguistics*. Revised Edition. New York, NY: Holt, Rinehart and Winston.

Gold, E. M. (1967). Language Identification in the Limit. *Information and Control, 16,* 447-474.

Goldsmith, J. (2001). Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics* 27(2): 153-198.

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.

Gropen, J., Pinker, S., Hollander, M., Goldberg, R. & Wilson, R. (1989). The Learnability and Acquisition of the Dative Alternation in English. *Language, 65,* 203-257.

Grünwald, P. (1994). A minimum description length approach to grammar inference. In G. Scheler, S. Wernter, and E. Riloff, (Eds.), *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language*. Berlin: Springer Verlag.

Guasti, M. T. (2002), *Language Acquisition The Growth of Grammar*, Cambridge, MA: MIT Press.

Halliday, M. A. K. (1985). *An Introduction to Functional Grammar*. London: Edward Arnold.

Hård, A. Sivik, L. & Tonnquist, G. (1996). NCS, Natural Color System – from Concept to Research and Applications. Part 1. *Color Research and Application*. 21(3): 180-205.

Hardin, C. L. (1988). *Color for Philosophers: Unweaving the Rainbow*. Indianapolis: Hackett Publishing Company.

Harrison, K. D., Dras, M. & Kapicioglu, B. (2002). Agent-Based Modeling of the Evolution of Vowel Harmony. In Masako Hirotani (ed.). *Proceedings of the Northeast Linguistic Society 32*. Amherst, MA: GLSA.

Hawkins, J. A. (1988). Explaining Language Universals. In J. A. Hawkins (Ed.) *Explaining Language Universals*. Oxford: Basil Blackwell.

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

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

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

Hendriks, P. (2000). The Problem with Logic in the Logical Problem of Language Acquisition. 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.

Hering, E. (1964). *Outlines of a Theory of the Light Sense*. Trans. L. M. Hurvich and D. Jameson. Cambridge, MA: Harvard University Press.

Heylighen, F. (1997): Occam's Razor. In F. Heylighen, C. Joslyn and V. Turchin (eds.) *Principia Cybernetica Web*. Brussels: Principia Cybernetica. Published on-line at: http://pespmc1.vub.ac.be/OCCAMRAZ.html.

Hirsh-Pasek, K. & Golinkoff, R. M. (1996). *The origins of grammar: evidence from early language comprehension*. Cambridge, MA: MIT Press.

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

Hurford, J. R. (1990) Nativist and Functional Explanations in Language Acquisition. In I. Roca (ed.) *Logical Issues in Language Acquisition*. Dordecht, Holland: Foris Publications.

Hurford, J. R. (2000). Social Transmission Favours Linguistic Generalization. In C. Knight, M. Studdert-Kennedy and J. R. Hurford (eds.). *The Evolutionary Emergence of Language: Social function and the origins of linguistic form*. Cambridge: Cambridge University Press.

Hurford, J. R. (2002). Expression/induction models of language evolution: dimensions and issues. In T. Briscoe (ed.) *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge: Cambridge University Press.

Indow, T. (1988). Multidimensional studies of the Munsell color solid. Psychological Review, 95(4): 456-470.

Kay, Paul. 1975. Synchronic variability and diachronic change in basic color terms. *Journal of Language in Society* 4:257-70.

Kay, P., Berlin, B., Maffi, L. & Merrifield, W. (1997). Color Naming Across Languages. In C. L. Hardin & L. Maffi (eds.) *Color Categories in Thought and Language*. Cambridge: Cambridge University Press.

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

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

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

Kay, P., Regier, T., Cook, R. & O'Leary, J. (n.d.). *Statistical tests of cross-language color naming*. Published on-line at http://www.icsi.berkeley.edu/wcs/study.html and downloaded on 22 May, 2003.

Keenan, E. & Comrie, B. (1977). Noun Phrase Accessibility and Universal Grammar. *Linguistic Inquiry*, 8:63-99.

Keenan, E. & Hawkins, S. (1987). The Psychological Validity of the Accessibility Hierarchy. In E. Keenan (ed.), *Universal Grammar: 15 Essays*. London: Croom Helm.

Kirby, S. (1999). *Function Selection and Innateness: The Emergence of Language Universals*. Oxford: Oxford University Press.

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. R. Hurford (Eds.), *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*. Cambridge: Cambridge University Press.

Kirby, S. (2002). Learning, Bottlenecks and the Evolution of Recursive Syntax. In E. Briscoe (Ed.), *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge University Press.

Kobayashi, I., Furukawa, K., Ozaki, T. & Imai, M. (2002). A Computational Model for Children's Language Acquisition using Inductive Logic Programming. In S. Arikawa & A. Shinohara (Eds.), *Progress in Discovery Science*. Berlin: Springer-Verlag.

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1): 1-7.

Kosko, B. (1994). *Fuzzy Thinking The New Science of Fuzzy Logic*. Glasgow: Harper Collins.

Ladefoged, P. (1975). *A Course in Phonetics*. New York, NY: Harcourt Brace Jovanovich.

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.

Land, E. H. (1977). The Retinex Theory of Color Vision. *Scientific American*, 237 (6): 108-128.

Landau, B. & Gleitman, L. (1985). *Language and Experience Evidence from the Blind Child*. Cambridge, MA: Harvard University Press.

Langley, P. (1995), *Simplicity and Representation Change in Grammar Induction* (Unpublished Manuscript). Palo Alto, CA: Institute for the Study of Learning and Expertise.

Langley, P., & Stromsten, S. (2000). Learning context-free grammars with a simplicity bias. In R. L. de Mantaras, and E. Plaza, (Eds.) *Proceedings of the Eleventh European Conference on Machine Learning*. Barcelona: Springer-Verlag.

Latimer, C. (1995). Computer Modeling of Cognitive Processes. *Noetica* 1(1). Published on-line at http://www2.psy.uq.edu.au/CogPsych/Noetica/.

Levinson, S. C. (2001). Yélî Dnye and the Theory of Basic Color Terms. *Journal of Linguistic Anthropology*, 10(1):3-55.

Levinson, S. C. (2003a). *Space in Language and Cognition: Explorations in Cognitive Diversity*. Cambridge: Cambridge University Press.

Levinson, S. C. (2003b). Spatial language. In L. Nadel (ed.), *Encyclopedia of Cognitive Scienc*e, Vol. 4:131-137. London: Nature Publishing Group.

Lewis, J. D. & Elman J. L. (2001). Learnability and the Statistical Structure of Language: Poverty of the Stimulus Revisited. In B. Skarabela, S. Fish, and A. H.-J. Do (Eds.) *Proceedings of the 26th annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press.

Lucy, J. A. (1992). *Language Diversity and Thought A Reformulation of the Linguistic Relativity Hypothesis.* Cambridge: Cambridge University Press.

MacKay, D. J. C. (1995). Bayesian Methods for Supervised Neural Networks. In Arbib, M. A. (Ed.) *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA: MIT Press.

MacLaury, R. E. (1995). Vantage Theory. In J. R. Taylor and R. E. MacLaury (eds.), *Language and the Cognitive Construal of the World*. Berlin: Mouton de Gruyter.

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

MacLaury, R. E. (1997b). Ethnographic evidence of unique hues and elemental colors. Commentary on Saunders and van Brakel (1997). *Behavioral and Brain Sciences,* 20(2):202-203.

MacLaury, R. E. (1999). *Vantage Theory in Outline*. Published on-line at http://www.sas.upenn.edu/~maclaury/VT-Outline.pdf. Downloaded on 22 November 2002.

McNeill, N. B. (1972). Colour and colour terminology. *Journal of Linguistics,* 81: 21-33.

MacWhinney, B. (2000). *The CHILDES Project Tools for Analyzing Talk*, third edition. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Maffi, L. & Hardin, C. L. (1997). Closing Thoughts. In C. L. Hardin and L. Maffi (Eds.), *Color Categories in Thought and Language*. Cambridge: Cambridge University Press.

Manning, C. D. (1993). Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. *31st Annual Meeting of the Association for Computational*

Linguistics Proceedings of the Conference. Association for Computational Linguistics.

Mazurkewich, I. & White, L. (1984). The acquisition of the dative alternation: Unlearning overgeneralizations. *Cognition* 16:261-83.

Miller, G. A. (1990). Nouns in WordNet: A Lexical Inheritance System. *International Journal of Lexicography*, 3(4), pp 245-264.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4), pp 235-244.

Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw-Hill.

Morris W.C., Cottrell, G.W., & Elman, J.L. (2000). A connectionist simulation of the empirical acquisition of grammatical relations. In Stefan Wermter and Run Sun (Eds.) *Hybrid Neural Symbolic Integration*. Berlin: Springer-Verlag.

Newmeyer, F. J. (1998). *Language Form and Language Function*. Cambridge, MA: MIT Press.

Nowak, M. A., Komarova, N. L. & Niyogi, P. (2002). Computational and Evolutionary Aspects of Language. *Nature*, 417:611-617.

Oliver, J. J. and Hand, D. (1996). *Introduction to Minimum Encoding Inference*. Technical Report 205 (amended version). Walton Hall, UK: Department of Statistics, Open University.

Onnis, L., Roberts, M. and Chater, N. (2002). Simplicity: A cure for overgeneralizations in language acquisition? In W. Gray and C. Schunn (Eds.) *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Osborne, M. (1999). MDL-based DCG induction for NP identification. In M. Osborne and E. T. K. Sang, (Eds.), *CoNLL-99 Computational Natural Language Learning.* Association for Computational Linguistics.

Pinker, S. (1989), Learnability and Cognition: the Acquisition of Argument Structure. Cambridge, MA: MIT Press.

Pinker, S. (1994). *The Language Instinct*. New York, NY: William Morrow and Company.

Pinker, S. (1997). *How the Mind Works*. New York, NY: W. W. Norton and Company.

Poortinga, Y. H. & Van de Vijver, F. J. R. (1997). Is there no cross-cultural evidence in colour categories of psychological laws, only of cultural rules? Commentary on Saunders and van Brakel (1997). *Behavioral and Brain Sciences,* 20(2):205-206.

Ratliffe, F. (1976). On the Psychophysiological Bases of Universal Color Terms. *Proceedings of the American Philosophical Society*, 120(5): 311-330.

Redington, M., Chater, N. & Finch, S. (1998). Distributional Information: A Powerful Cue for Acquiring Syntactic Categories. *Cognitive Science*, *22*, 425-469.

Rissanen, J. (1978). *Modeling by shortest data description*. Automatica, 14: 465-471.

Rissanen, J. (1983). A universal prior for the integers and estimation by MDL. *Annals of Statistics*, 11(2): 416-431.

Rissanen, J. & Ristad, E. S. (1994). Language Acquisition in the MDL Framework. In E. S. Ristad, (Ed.), *Language Computation*. Philadelphia: American Mathematical Society.

Roberson, D., Davies, I. & Davidoff, J. (2000). Color Categories are Not Universal: Replications and New Evidence from a Stone-Age Culture. *Journal of Experimental Psychology: General*, 129(3): 369-398.

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

Rumelhart, D. and McClelland, J. (1986). On learning the past tenses of English verbs. In McClelland, J., Rumelhart, D., and the PDP research group (eds.). *Parallel Distributed Processing: Explorations in the microstructure of cognition. Volume 2: Psychological and Biological Models*. Cambridge, MA: MIT Press.

Sacks, O. (1996). *The Island of the Colour-blind*. Sydney: Pan Macmillan Australia.

Saunders, B. A. C. (1992). *The Invention of Basic Color Terms*. Utrecht: ISOR.

Saunders, B. A. C & van Brakel, J. (1997). Are there nontrivial constraints on color categorization? *Behavioral and Brain Sciences*, 20(2):167-228.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 27,* 379-423 & 623-656.

Shepard, R. N. (1987). Towards a Universal Law of Generalization for Psychological Science. *Science*, 237:1317-1323.

Shepard, R. N. (1992). The Perceptual Organization of Colors,: An Adaptation to Regularities of the Terrestrial World? In J. H. Barkow, L. Cosmides and J. Tooby (eds.) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford: Oxford University Press.

Shepard, R. N. (1994). Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin and Review*, 1(1):2-28.

Siskind, J. M. (1997). A computational study of cross-situational techniques for learning word-to-meaning mappings. In M. R. Brent (Ed.) *Computational Approaches to Language Acquisition*. Cambridge, MA: MIT Press.

Smith, K. (1999). *Cognitive Linguistics and Connectionist Models of Language Acquisition.* MSc Dissertation, Centre for Cognitive Science, University of Edinburgh.

Solomonoff, R. J. (1960). The mechanization of linguistic learning. In *Proceeding of the 2nd International Congress on Cybernetics*. Namur, Belgium: Association Internationale de Cybernétique.

Solomonoff, R. J. (1964a) A Formal Theory of Inductive Inference, Part I. *Information and Control*, 7(1): 1-22.

Solomonoff, R. J. (1964b) A Formal Theory of Inductive Inference, Part II. *Information and Control*, 7(2): 224-254.

Starkie, B. (2001). Programming Spoken Dialogs Using Grammatical Inference. In Brooks, M., Corbett, D., and Stumptner, M., (Eds.), *AI 2001: Advances in Artificial Intelligence*. Berlin: Springer.

Steels, L. & Kaplan, F. (1998) Stochasticity as a Source of Innovation in Language Games. In C. Adami, R. Belew, H. Kitano and C. Taylor (eds.). Proceedings of Artificial Life VI. Cambridge, MA: MIT Press.

Stelmach, G. E. & Vroon, P. A. (Eds.). (1988). *Fuzzy Sets in Psychology*. Amsterdam: North-Holland.

Stolcke, A. (1994). *Bayesian Learning of Probabilistic Language Models.* Doctoral dissertation, Department of Electrical Engineering and Computer Science, University of California at Berkeley.

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

Teal, T. K. & Taylor, C. E. (2000). Effects of Compression on Language Evolution. *Artificial Life*, 6: 129-143.

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

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

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

Thompson, E. (1995). *Colour vision: a study in cognitive science and the philosophy of perception.* New York: Routledge.

Travis, L. (1989). Parameters of phrase structure. In M. R. Baltin and A. C. Kroch (Eds.) *Alternative Conceptions of Phrase Structure*. Chicago, Illinois: University of Chicago Press.

Ungerer, F. & Schmid, H.-J. (1996). *An Introduction to Cognitive Linguistics*. London: Longman.

Venkataraman, A. (2001). A Statistical Model for Word Discovery in Transcribed Speech, *Computational Linguistics*, 27(3): 352-372.

Wallace, C. S. & Boulton, D. M. (1968). An information measure for classification. *Computer Journal*, 11(2): 185-194.

Wallace, C. S. & Dowe, D. L. (1999). Refinements of MDL and MML Coding. *Computer Journal*, 42(4): 330-337.

Wolff, J. G. (1987). Cognitive Development as Optimisation. In L. Bolc (Ed.) *Computational Models of Learning*. Berlin: Springer-Verlag.

Wolff, J. G. (1991). *Towards a theory of Cognition and Computing*. New York, NY: Ellis Horwood.

Xiao, Y., Wang, Y. & Felleman, D. J. (2003). A Spatially Organized Representation of Colour in Macaque Cortical Area V2. *Nature*, 421:535-539.

Zadeh, L. A. (1965). Fuzzy Sets. *Information and Control*, 8:338-53.

Appendix A

Source Code for Model with Continuous Colour Space

The source code is contained in three files, *community.cpp, person.cpp* and *person.h*, which are on the accompanying CD. The compiler used was Borland C++, so the files should conform to the ANSI standard, and hence be compatible with any ANSI C++ compiler. The file to be compiled is *community.cpp*, which will in turn compile the other two files. The programs are menu driven, and documentation is included as part of the source code. A number of parameters at the beginning of the file *community.cpp* can be adjusted, but most of these can also be adjusted while the program is running, using the menu system. The model can output graph data, which is saved in a simple ASCII format suitable for loading into Microsoft Word's graph drawing component. (These files have a *.gra* suffix.)

Appendix B

Source Code for Model with Discrete Colour Space

This appendix also appears on the accompanying CD. The source code for the model is contained in three files, *conservative.cpp, dcperson.cpp,* and *dcperson.h.* Of these, the first should be compiled, as this will in turn compile the other two. Most of the observations made in Appendix A concerning the source code for the first model also apply here. In addition, there is a number of programs which were used in automating the analysis of data output by *conservative.cpp,* and which were used to obtain some of the results presented in Chapter 6 and Chapter 7.

conservative.cpp has an option which allows a summary of the colour term denotations known by all the speakers to be output to a file. (These files have a *.sum* suffix.) The output files are in a format similar to Figure 6.3, as it shows which colours are denoted by each colour term, and each line corresponds to a separate speaker. *collator.cpp* can then read in a collection of these files, and output them in a format which can be analysed by the other program. (These files have the filename *collation.txt*.) *balanced.cpp* is for investigating whether the unique hue points are evenly distributed between colour terms. *focicentrality.cpp* calculates how central the

prototypes of colour terms are to the category as a whole, while *fociuniquedistance.cpp* investigates whether category prototypes tend to be at the same locations as unique hue points. *focilocations.cpp* simply counts the number of colour terms with their prototypes at each colour. *mean.cpp* and *meannoisy.cpp* were used to calculate the mean number of colour terms in systems which had been run with varying life expectancies for the artificial people, and either with or without the presence of random noise (see Chapter 7). Finally, *typology.cpp* outputs files listing the number of colour terms of each type, and the number of overall colour term systems of each different type. There is also a compiled Windows executable for each of these programs, although some parameters can only be changed if the programs are recompiled. Further documentation can be found in the source code itself.

Appendix C

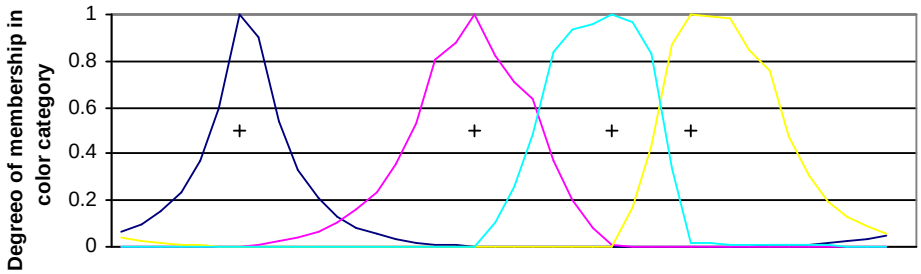
Results Obtained with Discrete Colour Model

This appendix also appears primarily on the CD. It contains files output by the discrete colour model, and some of the files output by the analysis programs. Most of the files are summaries of whole communities (in the form of *.sum* files), but there are also some *.gra* files, which give more detailed representations of the colour terms known by individual people, *.comm* files which list all the colour term examples observed by each person in a community, and *.txt* files which are output by the analysis programs, and which summarise various aspects of the emergent languages.

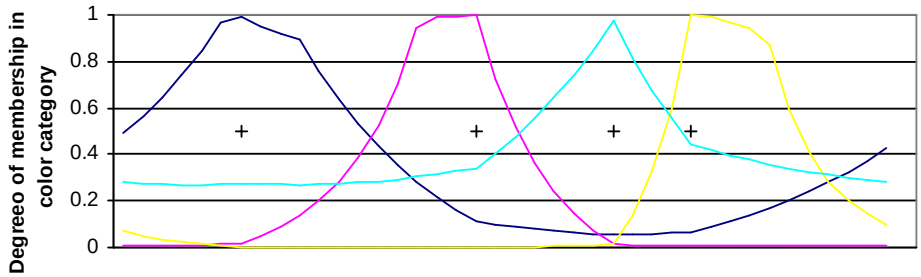
The files are divided into five subdirectories. The first of these contains data from which the Urdu colour terms graph of Figure 5.1 was plotted. The second directory contains the data which was used to plot the graphs of Figure 6.1, which show a cross section of the colour terms known by different speakers in a single community. Table C.1 below lists the number of examples which had been observed by each person in that community, and Figure C.1 contains graphs showing the denotations of the colour terms learned by each person. The graphs in Figure 6.1 are for people 1, 9, 5 and 7 (and they appear in that order in Figure 6.1).

Artificial Person	Number of Examples Remembered
1	68
2	38
3	2
4	29
5	113
6	75
7	32
8	100
9	71
10	24

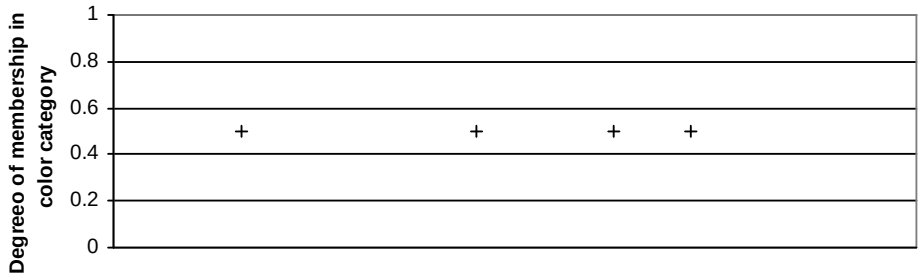
Table C.1. The Number of Examples Remembered by Each Artificial People in the Simulation Reported in Section 6.2. (The numbers of each person correspond to those in Figure C.1 below.)



(1)

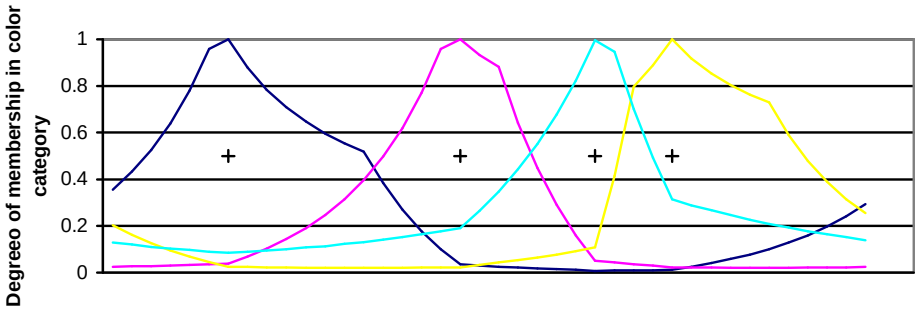


(2)

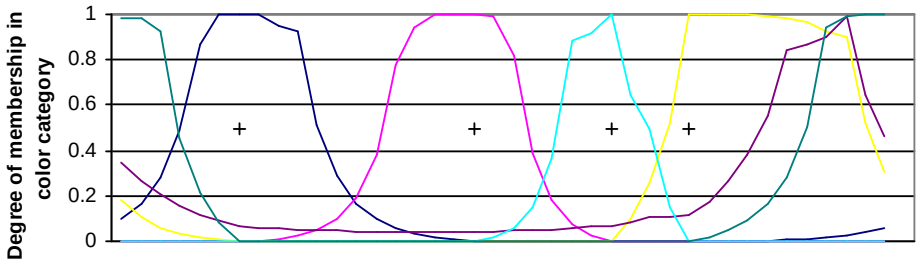


(3)

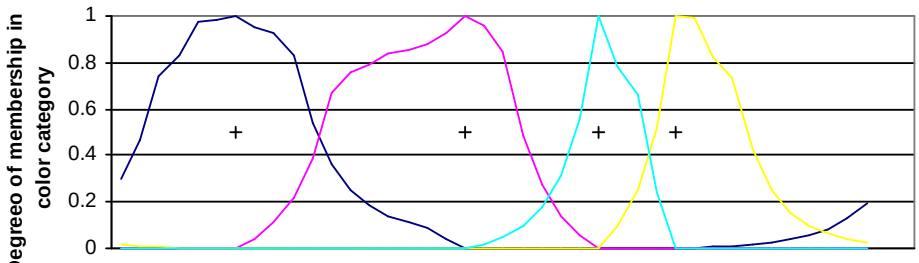
(4)


Degreeo of membership in color category
1
0.8
0.6
0.4
0.2
0

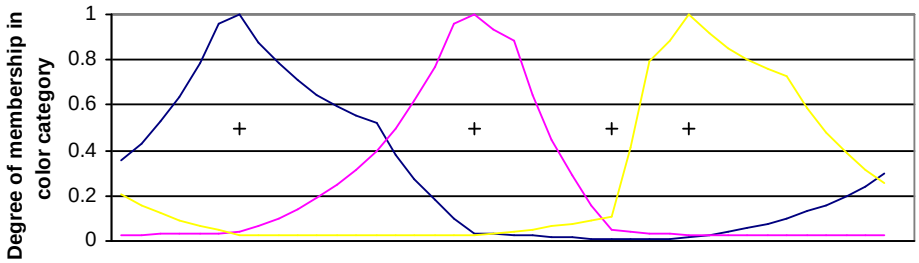
(5)


Degree of membership in color category
1
0.8
0.6
0.4
0.2
0

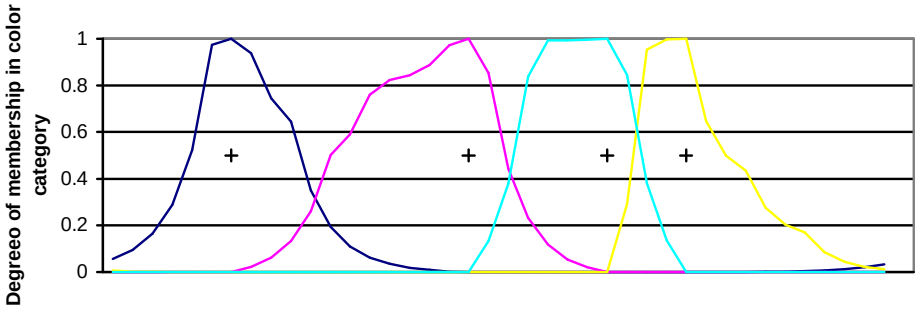
(6)


Degreeo of membership in color category
1
0.8
0.6
0.4
0.2
0

(7)


Degree of membership in color category
1
0.8
0.6
0.4
0.2
0

(8)


Degreeo of membership in color category
1
0.8
0.6
0.4
0.2
0

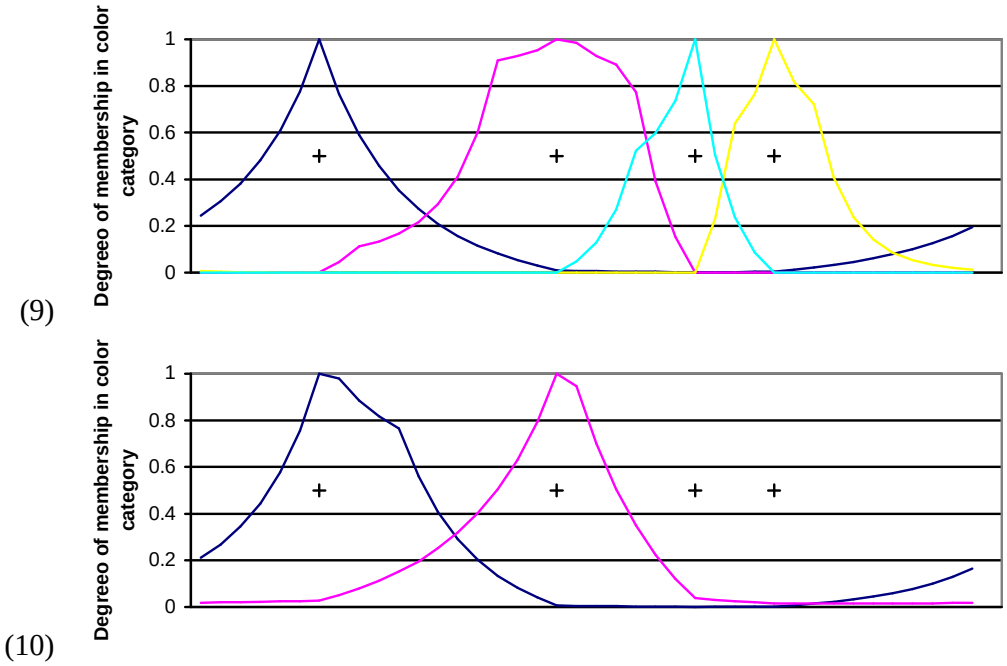


Figure C.1. The Basic Colour Term Systems of all the Artificial People from the Simulation Reported in Section 6.2.

(+'s mark unique hue locations, *p*=0.5, and people are creative one time in a thousand. Only colour terms for which people had remembered at least four examples are shown. Graph 3 is blank because person 3 had not seen four examples of any colour term.)

The final three directories contain summaries of the colour terms known by the artificial people present at the end of the evolutionary simulations, together with files containing analyses of this data. Each directory contains data for either the condition in which no unique hues were simulated, the unique hues were simulated but there was no random noise, or when both unique hues and random noise were simulated.

Appendix D

Ditransitive Verb Corpus

Below is the full corpus of data used for learning the dative alternation with the model described in Chapter 9. There are 55 sentences containing the double object dative structure, 55 containing the prepositional dative structure with *gave, lent, passed* or *sent* and 40 sentences containing the *donated* in the prepositional dative construction.

John gave a museum to the painting
Sam passed Sam to a painting
John passed John to Sam
the painting gave Sam Sam
John gave John a museum
Sam gave Sam to Sam
the painting lent Sam to John
Sam passed Sam John
John donated a painting to the museum
a painting donated the painting to Sam
a painting lent Sam a museum
Sam lent Sam a museum
Sam gave John Sam
a museum donated John to Sam
a painting lent John the painting
the museum donated John to Sam
John gave John to Sam
Sam lent a painting the museum
Sam lent John Sam
Sam gave a painting a museum
a museum gave John Sam
Sam donated Sam to Sam
a painting donated John to John
John gave John to John
John gave a museum to the painting
John passed Sam John
Sam lent John Sam
a painting passed John to John
John gave John John
John passed Sam to John

a museum lent John to John
Sam gave a museum to John
John lent John John
a painting passed the painting John
the museum lent John to Sam
Sam gave Sam the museum
Sam donated Sam to John
John passed John to Sam
John lent the painting to a painting
John gave Sam to Sam
John gave Sam Sam
John lent Sam to John
Sam passed Sam John
John passed Sam Sam
John lent Sam Sam
John passed John to a painting
John gave a painting a painting
the painting gave John the museum
Sam lent John Sam
a painting donated John to Sam
John donated John to Sam
John passed Sam Sam
John lent Sam to Sam
John lent John Sam
Sam donated John to Sam
John lent the museum Sam
John lent John Sam
Sam gave Sam the painting
Sam donated Sam to Sam
a museum donated Sam to a museum
Sam lent the museum the museum
John donated John to the painting
John lent the painting to the painting
John lent John Sam
John passed Sam a painting
Sam gave a museum to John
a painting donated Sam to Sam
a painting passed Sam to a painting
Sam gave Sam Sam
a museum gave the museum to the painting
the museum gave the painting to Sam
John gave a museum Sam
John lent John John
the museum lent the painting John
Sam lent a painting to Sam
Sam passed John John
Sam lent Sam to John
Sam passed John to Sam
John gave a painting the museum
John lent Sam Sam

John passed John to John
John passed Sam Sam
Sam passed Sam a painting
Sam gave Sam the museum
a painting passed John to John
a museum passed John to John
John donated John to the painting
John gave the museum to Sam
the painting donated John to John
Sam donated Sam to Sam
John donated Sam to John
a painting lent a painting to a museum
John passed the painting to Sam
the museum donated the museum to Sam
John gave Sam to Sam
Sam lent Sam to John
Sam gave a museum to John
John donated John to Sam
Sam donated Sam to Sam
Sam passed Sam to Sam
John lent John to John
the painting lent the museum to Sam
the museum donated Sam to the painting
Sam donated John to John
Sam donated John to a museum
the museum donated a painting to John
John passed John to Sam
John passed John to John
John donated the painting to a painting
Sam donated Sam to the painting
John passed Sam to John
Sam donated John to the painting
a painting donated a museum to the painting
John gave a painting to Sam
a painting lent Sam to John
John lent John to John
a painting donated Sam to Sam
the museum gave a museum to Sam
John passed John to John
the painting donated John to John
John donated John to the painting
a painting donated Sam to Sam
John donated John to the painting
a museum donated Sam to a museum
Sam donated Sam to Sam
Sam donated John to Sam
John donated John to Sam
a painting donated John to Sam
Sam donated Sam to John
Sam passed Sam Sam

John passed a painting John
Sam passed the painting a painting
John lent John the painting
the painting lent Sam Sam
Sam lent a painting Sam
the museum lent Sam a painting
Sam gave Sam the painting
the museum lent John the painting
a painting lent the painting John
the painting passed a museum to the painting
John lent Sam to a painting
Sam lent Sam to a painting
a painting gave Sam to Sam
John gave John to John
Sam gave the painting to a painting
Sam lent Sam to Sam
a painting passed John to Sam
John passed John to John
the museum gave a museum to Sam
the museum sent a painting to Sam

Appendix E

Source Code for the Syntax Model

This appendix appears on the accompanying CD, and contains program source code for the syntax learning model (written in SICSTUS PROLOG). It consists of four files, *searcher.pl*, *evaluator.pl*, *parser.pl* and *corpus.pl*. This program was originally written for research reported in Dowman (1998), and so does not form part of the work submitted for the PhD, but it is included here because it was used in the research reported in Chapter 9.

In order to run the programs, SICSTUS should be started in a directory where the files have been saved. Next, enter *compile(searcher).* , which will load in and compile all four files. Learning is initiated using *learn(LanguageName).* , where the name of a language, as given in the file *corpus.pl*, is substituted for *LanguageName*. (The corpus used to obtain the results reported in Chapter 9, and which is reproduced in Appendix D, is named *dative3* in *corpus.pl*.) Each language must be set up as exemplified in corpus.pl, with a list of the words, the non-terminal symbols (which must include root and s1), and all the sentences to be learned from. Most of the parameters which can be adjusted are at the beginning of *searcher.pl*, though the one concerning depth of parsing is hidden in *parser.pl*. To adjust which search moves are used with what probability, it is necessary to adjust the [1,2,4,4,4,5] in the rule below, which can be

found in the file *searcher.pl*. (The rule as given below is1/6 chance of adding, deleting or separating rules, and 1/2 chance of altering a rule, which is the setting which was used to obtain the results concerning the dative alternation.) Further documentation can be found in the code itself, or in Dowman (1998). A fifth file, *finalcorrectdative3grammarandevalutaion.txt*, gives the final grammar which was output by the program, and which was reproduced in an edited form in Table 9.7, together with its evaluation and the parses it assigns to sentences.

```
% 1=delete rule, 2=add rule, 3=merge rule, 4=alter rule, 5=separate rule.
learn(Language, GrammarNo, Temperature, Evaluation, Phase):-
    random_choice([1, 2, 4, 4, 4, 5], Random123456), % Decide whether to remove a rule or
    add a new one, merge rules, or alter a rule, or split a symbol.
```