Factors influencing the origins of colour categories

Tony Belpaeme

Artificial Intelligence Laboratory
Vrije Universiteit Brussel

Proefschrift voorgelegd voor het behalen
van de academische graad van doctor in de wetenschappen,
in het openbaar te verdedigen op vrijdag 8 maart 2002.

Acknowledgements

I started as a research assistant in the Artificial Intelligence Laboratory in autumn 1996. My first interests were into behavioural robotics and robot ecosystems. As a continuation to my "licentiaats" thesis I started building a camera system to extend the sensory perception of the lab's robots (Belpaeme and Birk, 1997a,b; Belpaeme, 1998; Birk and Belpaeme, 1998; Birk et al., 1998, 1999; Belpaeme and Birk, 2001). It was around that time when Luc Steels got interested in the origins of language. His early experiments formed the seed for what is now one of the most important paradigms for exploring linguistic interactions with computer simulations. Luc soon wanted more and had plans to implement a language experiment in the real world, for which I delivered the visual perception (Belpaeme et al., 1998; Belpaeme, 1999). This got me interested in visual features, and my research soon switched to the artificial evolution of visual feature detectors. As the robot research in the lab was simmering, most resources and brainpower were put on the origins of language research.
In 1999 Luc had the idea to take colour as theme for one of our courses on artificial intelligence. With the help of Erik Myin speakers from different backgrounds were invited: engineers, philosophers and even an artist came by to give their view on colour. During these lectures it dawned on me that there is more to colour than meets the eye. I there and then decided to delve into the splendid field of colour research and to use Luc's simulation paradigms to study colour categorisation. Two years later, this thesis is the result of my explorations.

I am most indebted to my supervisor, Luc Steels, for the opportunities and intellectual freedom he gave me. He provided the ideas and guidance for this work and though his physical presence is every now and then lacking, his influence and inspiration are always felt.

I am also indebted to Barbara Saunders, Erik Myin, Johan Lammens, Paul Kay and Michael Bach for answering my questions, for introducing me to the literature and for their kind advice.

My gratitude goes out to all my friends and colleagues of the AI Lab whom I ever had the pleasure of working and socialising with: Karina Bergen, Andreas Birk, Joachim De Beule, Bart de Boer, Edwin de Jong, Bart de Vylder, Sabine Geldof, Bart Jansen, Holger Kenn, Dominique Osier, Jo-

han Parent, Peter Stuer, Joris Van Looveren, Dany Vereertbrugghen, Paul Vogt, Tante Emma, Thomas Walle and Jelle Zuidema.

I am also very grateful to all who proof read and criticised the manuscript: Joachim De Beule, Bart Jansen, Claudine Lesaffre, Bart de Vylder, Peter Stuer, Joris Van Looveren and Jelle Zuidema.

I would also like to thank all the people who kept my academic activities into perspective, I will not sum up your names here, there's too many of you. You know who you are. Thanks!

Of course, I would like to thank my parents and my sister for always stimulating and encouraging me to learn and to explore. The values you thaught me form the roots of this work.

Finally, my warmest appreciation goes out to my girlfriend Sarah for her support, for the warmth she gave me, and for putting up with all the time I spent on my thesis and not on more important things.

Summary

In the evolution of language field a large body of work exists on mapping meaning to form, but little attention has been paid to the formation of meaning itself. Nevertheless, the debate on the nature of perceptually grounded concepts such as colour categories, and the impact of language on this, is central to cognitive science and linguistics. One point of view considers perceptual concepts as universal to all humans and argues that genetical determinism is responsible for this. Opponents adhere to a relativist account, which claims that concepts are learned and thus are ecologically and culturally specific.

Colour categorisation presents an opportunity for testing both perspectives. Berlin and Kay, in the late sixties, and Rosch, in the early seventies, provided convincing evidence for the universalist position through linguistic and memory experiments. Field data collected and interpreted over the last decades show a remarkable agreement between colour categories from different cultures. The relativist position, it is argued, is unable to account for these phenomena. Backed up by results from neuropsychology on the opponent character of colour perception the universalist stance has held strongly. Only recently, evidence from anthropology and close scrutiny of experimental results and the deductions based on these undermine the authority of the universalist model. The question arises if the relativist position can account for the strong facts brought forward by experimental data.

This work investigates outstanding issues in the colour categorisation debate using computational modelling. It reports new insights on the plausibility of accounts of the origins of colour categories. Four simulations have been constructed, two in which colour categories are evolved through a process of natural selection, the other two in which colour categories are formed with a learning approach using environmental, ecological and cultural constraints. Both genetic evolution and learning are studied with and without the influence of language.

The basic entity of the simulations is the *agent*. The colour perception of an agent is modelled as a mapping from spectral measurements to an internal three-dimensional colour space. Categories are defined on this internal space as adaptive network. These adaptive networks consist of locally tuned receptors sensitive to small region in the space. The output of a net-

work is the weighted sum of the reactions of all its locally tuned receptors. Categories can also be associated with words, needed for communicating colour meaning to other agents. The association between a category and a word is modelled by a scalar value representing the strength of the association.

Four simulations have been implemented; (1) agents learn categories individually under the influence of environmental and ecological constraints, (2) agents acquire categories through cultural learning, where agents use language to learn the colour categories, (3) agents genetically evolve categories with selection happening on the ability of the agents to discriminate colour stimuli, and (4) agents genetically evolve categories with selection happening on the communicative abilities of the agents.

The categorical repertoires of an agent in simulations (1) and (3) are judged on a discrimination task, in which an individual has to discriminate one colour stimulus from a set of other colour stimuli. The repertoires of agent in simulations (2) and (4) are judged on their performance on a communication task. This communication task is evaluated with two agents: one agent has to name one colour from a set of colours and the second agent has to guess what colour the first agent meant.

The results show how genetic evolution and learning are both able to come up with a set of colour categories satisfying the discrimination task. Also genetic evolution and cultural learning arrive at colour categories shared throughout the population. The evolutionary approach arrives at identical categories because evolved individuals are all offspring of a few fit ancestors and thus all agents share the same "colour genes". On the other hand, a population using cultural learning also reaches shared colour categories; not because the categories are genetically passed on, but because the linguistic interactions push the agents towards having similar categorical repertoires needed for communication. So this shows how individuals can reach coherent categories by way of an indirect coupling through the environment. The linguistic task boosts the conceptual coherence, showing a causal influence of language on category acquisition. We argue that an adaptive model together with a shared ecology and linguistic sharing of category labels provides a viable explanation for sharedness of colour categories, dispensing the need for a universalist account.

Samenvatting

Het onderzoek naar de evolutie van taal besteedt veel aandacht aan het afbeelden van betekenis op woordvormen, maar vaak wordt er weinig aandacht besteed aan het eigenlijke ontstaan van betekenis. Niettemin is het debat over het karakter van perceptueel verankerde betekenis en de invloed van taal hierop, cruciaal voor de cognitieve wetenschappen en voor de linguïstiek. Omtrent het karakter van perceptuele concepten bestaan er twee verschillende stromingen. De ene stroming beweert dat verankerde concepten genetisch bepaald en dus universeel zijn voor iedereen. Tegenstanders verdedigen een relativistische standpunt en beweren dat concepten geleerd worden en zodus ecologisch en cultureel specifiek zijn.

Kleurcategorisering vormt het ideale domein om beide perspectieven te onderzoeken. Berlin en Kay, in de late jaren zestig, en Rosch, in de vroege jaren zeventig, hebben door middel van taalexperimenten en geheugenexperimenten sterke aanwijzingen gevonden voor het universele karakter van kleurcategorieën. Het veldonderzoek van de laatste decennia heeft steeds een merkwaardig sterke overeenkomst aangetoond tussen de kleurcategorieën van verschillende volkeren. Er wordt beweerd dat een relativistische theorie nooit in staat zal zijn om deze verschijnselen te verklaren. De universele verklaring wordt nog gesterkt door resultaten uit de neurofysiologie, waar het opponent karakter van kleurperceptie werd gemeten in de hersenen van apen. Slechts recent werden nieuwe bewijzen geleverd voor het relativistisch karakter van kleurcategorieën; onder andere door nieuw anthropologisch onderzoek (bijvoorbeeld Davidoff) en een kritische evaluatie van oude resultaten en conclusies (onder andere van Saunders en van Brakel, of Lucy). De vraag blijft echter of relativisme de sterke feiten kan verklaren waarop het universalisme gestoeld is.

Dit proefschrift onderzoekt de verschillende kwesties in het kleurendebat door middel van computermodellen. De resultaten van de simulaties verschaffen nieuwe inzichten over de waarschijnlijkheid van de verschillende argumenten in het debat. Er werden vier verschillende computersimulaties gebouwd, in twee daarvan worden kleurcategorieën genetisch geëvolueerd door middel van natuurlijke selectie. In de andere twee worden kleurcategorieën geleerd op een adaptieve wijze onder druk van omgevingsfactoren, ecologie en cultuur.

Een simulatie bevat een populatie van individuen (vanaf nu *agents* genoemd). De kleurperceptie van een agent wordt gemodelleerd als een afbeelding van spectrale metingen van kleuren op een interne driedimensionale kleurenruimte. Categorieën worden voorgesteld door middel van adaptieve netwerken. Deze adaptieve netwerken bestaan uit receptoren die gevoelig zijn voor een bepaald gebied in de kleurenruimte. De reactie van een adaptief netwerk bestaat uit de gewogen som van de reacties van alle receptoren die behoren tot het netwerk. Categorieën kunnen ook geassocieerd worden met een woord. Deze associatie wordt voorgesteld door een scalaire waarde die de sterkte van de associatie weergeeft.

Er werden vier verschillende simulaties gebouwd: (1) de agents leren hun categorieën individueel onder invloed van omgevingsfactoren en ecologische factoren, (2) de agents leren hun categorieën cultureel, wat inhoudt dat de agents taal gebruiken tijdens het leren van categorieën, (3) de categorieën worden genetisch geëvolueerd, de selectie gebeurt op de bekwaamheid van de agents bij het onderscheiden van kleuren en (4) de categorieën worden genetisch geëvolueerd, maar nu gebeurt de selectie van de agents op hun communicatieve bekwaamheid.

De categorieën van een agent in de simulaties (1) en (3) worden beoordeeld aan de hand van een discriminatietaak. Hierbij moet het individu één kleur kunnen onderscheiden van een verzameling andere kleuren. In simulaties (2) en (4) worden de agents beoordeeld op een communicatietaak. Deze taak vereist twee agents: de ene agent benoemt een kleur uit een verzameling kleuren en de andere agent moet raden welk kleur bedoeld werd.

De resultaten in dit proefschrift tonen hoe genetische evolutie en leren allebei kleurcategorieën opleveren die voldoen om te discrimineren. Ook laten ze zien hoe genetische evolutie en cultureel leren kleurcategorieën opleveren die gedeeld zijn in de populatie, dit wil zeggen dat elke agent na een tijdje min of meer dezelfde kleurcategorieën bezit. Ook met genetische evolutie wordt bereikt omdat na een aantal generaties alle agents afstammen van slechts enkele aangepaste voorouders, en daardoor allemaal dezelfde "kleurgenen" delen. Anderzijds slaagt ook een populatie waarin categorieën cultureel geleerd worden erin om gedeelde categorieën te bereiken; dit wordt niet veroorzaakt door het doorgeven van genetisch materiaal, maar door de linguïstische interacties die de categorieën dwingen om meer gelijkenis te vertonen. Want, praten over iets lukt slechts als iedereen hetzelfde bedoelt met de woorden die gebruikt worden. Dit toont aan dat individuen erin slagen om coherente categorieën te verkrijgen door middel van een indirecte koppeling in hun taal en hun omgeving. De linguïstische taak drijft de conceptuele coherentie op, wat aantoont dat een causaal verband tussen taal en het verwerven van categorieën theoretisch mogelijk is. Hierbij verdedigen we dat een adaptieve leermethode samen met een gemeenschappelijke ecologie en het linguïstisch delen van woorden een

waardevolle verklaring biedt voor het gedeeld zijn van kleurencategorieën, en zo een alternatief vormt voor de universele verklaring waarom mensen een hechte interpretatie hebben van kleur.

Contents

List of Figures

Chapter 1

Introduction

What a drab place the world would be without colour. Imagine living in a world devoid of all colour, where everything would be ashen and dull. One realises quickly that colour adds life and lustre to our world. But colour not only has an aesthetical quality, it also enhances the way in which we perceive and recognize objects in our everyday world. It is used to segment complex visual input into regions aiding us to differentiate objects from the background (Callaghan, 1984). When recognizing objects and scenes, colour is an important cue and facilitates classification. For example, objects in their typical colour —a red fire engine— are more rapidly recognised than objects in artificial colours —a blue fire engine (Wurm et al., 1993). Recently, is has been discovered that colour perception in primates has primarily evolved to distinguish fruit and young leaves from foliage (Sumner and Mollon, 2000; Regan et al., 2000), showing the importance of colour to humankind's closest relatives and plausibly also to us humans.

Though we share much more with other primates than just our colour vision, there is one eminent trait settings us apart from apes and monkeys: our ability to use language. We use language to communicate about our world and thoughts. As we do, we convey words to each other, or more generally: we communicate symbols. A first requirement is that these symbols are shared within our language community. If not, conversation would be impossible. A second requirement comes from the fact that symbols are associated with meaning and so not only the symbols have to be shared, also the meanings of the symbols have to be agreed upon by all speakers of a language. Only then can we truly communicate.

One of the things we quite often talk about is colour. Colour is a perceptual phenomenon, and so colour categories are perceptually grounded categories. These are the categories we talk about when using words such as "red" or "green". We all seem to agree on the meaning of these words, but where does this agreement come from? Are we born with colour categories already in place? Do we learn colour categories by interacting with

our environment during our childhood? Or do our parents teach us colour categories while teaching us colour names? These questions, though so simple in nature, have led to much controversy. Some believe that perceptual categories are innate, i.e. we are born with these categories already in place. Others believe that we learn these categories during nurture. This thesis attempts to shed light on these issues by means of computer simulations of colour categorisation and colour naming. Different simulations are explored, each corresponding to the main dividing lines in the discussion on the origins of colour categories.

1.1 Categories, concepts and meanings

Concepts and categories are two notions often used in the same context without clearly stating what exactly is meant by the concept of a "concept" and concept of a "category". Concepts and categories are central to cognitive science and this work dwells on a specific field that benefits from a clear definition of both notions. There exist many different interpretations of concepts (Wilson and Keil, 1999, p. 176), but for our purposes the definition from behavioural psychology fits best: a *concept* represents a category that has a particular relation to other categories. An example of a concept is the number THREE[1], which stands in relation to other concepts (for example, THREE is one more than TWO and one less than FOUR). Now if a concept adds relations to categories, what then is a category?

For defining a category let us again restrict and adopt a workable notion from behavioural psychology for the purposes of this work: a *category* refers to specific things on the basis of their properties. A category is often defined by the principle of similarity: things are in the same category because they are more similar to each other than to things in categories they do not belong to. One problem with this view is the fact that similarity is ill defined (Medin, 1989). Consider for example two categories, CAT and DOG, the differences within each category are enormous (think about the differences between labradors and chihuahuas) and both categories have much more similarities than differences, they both contain animals having four legs, fur, a tail, and etcetera. Things only seem similar because they belong to the same category, so the roles seem to have changed here, since the categories now define what similarity is.

For our purposes, investigating perceptual colour categorisation, we can limit ourselves to saying that perceptual categories are indeed defined by similarity and that a consequence of this is that the world, our perception of it and our implementation of similarity will define what a category is. Cognitive scientists always refer to colour categories as opposed

[1]Written in uppercase to denote the difference between concepts or categories and their lexical labels.

to colour concepts. According to our definition a colour category refers to colours based on perceptual similarity. Colour concepts would then be the combination of a colour category and its relation to other colour categories.

When we perceive or think about the world we handle categories. When we communicate we intentionally convey *meaning* (of which categories and concepts form the basis) using words and sentences. But how do words get their meaning? We all seem to agree on the meaning of words, but where does shared meaning originate? Limiting ourselves to grounded, perceptual categories (such as colour categories), we can expose some important controversies.

1.1.1 Are categories genetically determined?

On the origins of meaning there exist two opponent views, meaning either is genetically determined or they can be acquired through learning. Students of cognition such as Chomsky (1965), Fodor (1998), Pinker (1994), Jackendoff (1992; 1993) and Shepard (1987; 1994) subscribe to the former.

The innate view on meaning seems very appealing. At first sight it seems that meaning is shared among all cultures and that only linguistic labels are different. Words are remarkably easily translated between languages, one can look up the translation of almost any word in a dictionary, which gives only more weight to the nativist view on meaning and categories. A strong nativist position is difficult to hold, as a newborn does not have all categories and concepts pre-programmed from basic gravitational concepts to concepts on new technology. But we might be given a genetic head start by having innate categories and concepts for basic notions to which all others are related.

Jackendoff (1993) proposes exactly this. He claims that we have an innate basis of concepts that can be used to construct an unlimited variety of other concepts. These basic concepts are considered to be building blocks and have to be present in order to achieve more elaborate conceptualisations. We cannot just think anything we want; we are restricted by these building blocks provided by our genetic make-up.

The psychologist and philosopher Fodor (1998) claims that a system able to represent higher concepts must express these in terms of basic concepts, inherited as part of the structure of the brain. His argument is based on studying logics in the Representational Theory of Mind framework, where he concludes that systems possessing less powerful logics cannot express and therefore cannot build systems that embody more powerful logics (Drescher, 1991, p. 38).

Chomsky (1965) also takes a radical nativist position of cognitive development. For him it is hard to believe that cognitive universals —concepts normally acquired by all humans (such as physical properties of objects)— are learned. There is an innately present mechanism which, interacting

with the normal environment, inevitably leads to concepts shared by all. Chomsky compares cognitive development with the nearly inevitable development of limbs by a zygote: the limbs are not present at conception, nor does the zygote contain a model of limbs; but still they do develop. It is an innately specified maturational process: the genome, under normal circumstances, specifies the growth of limbs. Chomsky decides that cognitive universals are no different from physiological universals.

Shepard (1994) argues that invariants in the world have been internalised by evolution, much in the same way as genes shape an individual's size or shape to be better adapted to its ecological niche. Just as the rotation of our planet has been internalised in diurnal species, other ecological constants have been hard-coded in our genetic composition. One example is our three-dimensional colour perception, which can be explained through our need for colour constancy under natural perceptual conditions. However, Shepard does not go as far as saying that colour categories are internalised under comparable ecological pressure.

Obviously there is a strong innate component to human cognition and language (Hawkins, 1992), as humans invariably acquire parallel cognitive and linguistic capabilities and other primates, even if intensively trained, hardly get any further than learning a handful of concepts (Savage-Rumbaugh, 1986). So the important question is: what mechanisms are innate and what components of the cognitive system develop during ontogeny.

1.1.2 Are categories learned?

When assuming that categories and concepts are learned, one is confronted with some important questions. If concepts are individually learned, how do concepts become shared between individuals? Is there a certain bias, environmental or biological, in concept learning which might account for concept sharing? And what is the role of language in all this?

Elman et al. (1996) claim that cognition is largely the results of dynamics in the interaction between the environment and the individual, and growth and learning of the individual. They are supported by progress made in recent years in genetics, embryology, and developmental neuroscience which has demonstrated a higher degree of cortical plasticity than was accepted in earlier research. Elman's research framework is that of computation modelling, or connectionism, where he investigates how programs modelling a cognitive function could learn certain behaviour, without this behaviour being pre-programmed.

Piaget (1977) investigates how knowledge develops in children and through doing so he elucidates the nature of knowledge in general. For this he takes a constructivist approach to epistemological relativism: all knowledge is constructed during lifetime; no structures are given in advance, in

the mind or in the external world. Children, while growing up, constantly build and try constructions and so adapt their knowledge to fit the external environment. According to Piaget, social interaction or collaboration is very little involved in this.

These *empiricist* psychologists claim category learning to be individualistic (only based on interaction between the learner and the environment) and observational (there is no feedback from another human being). Everyone still arrives at the same categorisations because of the bias present in the learning algorithm and in the environment: everyone shares the same cognitive mechanisms and the environment is highly structured so all learners inductively arrive at the same categorisations of the world.

In the machine learning field many different approaches exist to category acquisition. Examples include decision tree learning (Quinlan, 1993), in which discrete decisions for classifying something are modelled in tree-like structures, or connectionism (Rumelhart and McClelland, 1986), where categories are modelled in networks inspired by natural neural networks. However, these approaches require a training set of labelled examples and therefore implicitly delineate each category beforehand. The categories are given in advance, and so these approaches do not explain how categories originate. In relation to colour categories, Lammens (1994) has shown that a computational representation of colour categories can indeed be optimised with the aid of training data to resemble human colour categories.

Other approaches in machine learning deduce structure from the environment from statistical regularities in the input. These approaches are grouped under the denominator of unsupervised learning. No categories need to be defined beforehand, as they are extracted from the input data. An example is principal component analysis (Jolliffe, 1986) where structure is deduced from the environment by extracting the dimensions on which variance is maximal.

1.1.3 Are categories cultural?

If we would agree that perceptual categories are learned, this brings us of course to the influence of language on perceptual categorisation. When discussing the influence of language on cognition we arrive at the theories of Benjamin Lee Whorf an his mentor Edward Sapir. An often quoted passage from Whorf's work illustrates his ideas,

> The categories and types that we isolate from the world of phenomena we do not find there because they stare every observer in the face; on the contrary, the world is presented in a kaleidoscopic flux of impressions which has to be organized by our minds —and this means largely by the linguistic systems in our minds. We cut nature up, organize it into concepts, and

> ascribe significances as we do, largely because we are parties to an agreement to organize it in this way— an agreement that holds throughout our speech community and is codified in the patterns of our language. (Whorf, 1956, p. 213)

The Whorfian hypothesis has been drawn apart in a strong version and a weak version. The strong version states that language fully determines cognition: anything were no word exist for can not be thought. Of course, the strong thesis is impossible to uphold; this is immediately suggested by the fact that we can think about simple acts or facts which are very hard to describe linguistically. Or by the fact that a dog has the power of simple thoughts, even though it does not have language. The weaker version — known as linguistic relativism— is however much more attractive, it states that language has a causal relation, however weak, on our cognition. Even though this might seem attractive and sensible, it often contested and even mocked (Pinker, 1994). For the weak version there is also the problem left of defining the direction of the causal relation. Whorf claims that it is language which has an influence on cognition, but might it be cognition that has an influence on language? This causal relation is very hard to determine, and often empirical evidence for the former can also be interpreted as evidence for the latter (Lucy, 1996).

Applying the Whorf hypothesis to perceptual categories (e.g. colour, facial expressions, sound, odour) means that categories are not as much established because there is a perceptual similarity to them, but rather linguistic similarity is the critical factor driving perceptual categorisation (Davidoff, 2001). This does not mean that language alone specifies perceptual categories. There are constraints linked to the properties of our perceptual systems, for example we can not hear sound higher than 20kHz or the colours we can perceive are limited by our visual system. But besides these constraints, language still forms the basis upon which humans build their categorisation of the world.

The acquisition of knowledge in this framework is captured under the heading of *cultural learning*. Cultural learning agrees upon the fact that categories are learned and shared in a culture, but disagree with empiricists by saying that the structure of the environment alone is not enough to attain shared categories. Instead, social interaction is needed for "tuning" categorisation (e.g. Tomasello, 1999).

1.2 Colour categories and colour naming

The controversy surrounding the Sapir-Whorf thesis has been most prominent in the field of colour categorisation and colour naming. Colour terminology and its possible influence on the way individuals perceive, remem-

ber and make judgements of colour stimuli has been passionately investigated in the last fifty years by psychologists, anthropologists and linguists. Colour studies lend themselves very well to studying linguistic determinism. The physics of colour are known and described, mainly because of interest from technology and industry. Colour perception is a tightly defined domain, meaning that all relevant stimulus variables can, up to certain level, be accurately controlled. Moreover, much has been discovered on the physiological and neural substrate underlying colour perception. Furthermore, anyone can relate to colour; without much specific background a reasonable overview can be gained from studying the literature, making the field accessible to many, often very different, disciplines.

The research in this work builds on a consensus that has emerged in the field of colour categorisation and colour naming. The main points are summarised here,

- *Colour categories have a focal point and an extent with fuzzy boundaries.* This was established in diverse naming experiments and memory experiments.
- *Colour categories can be named.* E.g. every language has at least two words used exclusively to name colour categories.
- *Different languages use different colour words.* E.g. a red sample is called "rood" in Dutch, "red" in English and "vermell" in Catalan.
- *Colour categories aid our visual perception.* It has been shown that categories aid both discrimination of the visual world and higher-level vision. For example, colour facilitates the recognition of structurally similar objects (e.g. differencing between lemons and limes).

These banal agreements are in shrill contrast to the many controversies still present in the field. Some of the compelling questions are,

- Are colour categories shared within one language community? The categories are obviously shared to a certain degree, otherwise communication about colour would be impossible, but is the sharing absolute?
- Are colour categories shared by humans in different cultures and by humans speaking different languages?
- Are colour categories learned or innate?
- If colour categories are learned, what are the constraints under which they are learned? Will environmental and biological biases be enough to explain the nature of human categorical repertoires. Or do we need extra, maybe linguistic, constraints?

- If colour categories are learned, how can we explain a supposed sharedness of categories?
- Does language have an influence on colour categorisation or is language a separate entity with no influence on cognition whatsoever?

The broader scope of this work is the study of tenets of linguistic and cultural determinism and the rivalling theories of universalism. As a case study we take colour categorisation and naming; because, as mentioned before, the colour domain is pre-eminently the field were these discussions are held. As Gellatly (1995) states: "Because of the swing of opinion against the Whorfian hypothesis derives in large part from work on colour, any attempt to rehabilitate linguistic and cultural relativism must be able to accommodate the results of that work".

1.3 The aims of this work

This thesis tries to shed light on the question and controversies mentioned in the previous section by aid of computer simulations. By studying different variations of simulation models running under different conditions we hope to provide more insight on the active issues in universalism, empiricism and culturalism.

The common part of the simulations is built upon the consensus already reached. Colour categories serve a purpose, which in simulations is evaluated using a *discrimination* task. This is in accord with the primordial and most important function of colour perception, which is the quick and reliable discrimination of objects. Next, a mechanism is needed to acquire colour categories: here the consensus ends, and so we implement two strategies concordant with universalism and relativism. In the former, we let colour categories genetically evolve in a simulation. In the latter, colour categories are acquired using a learning mechanism. We first examine if both approaches are able to acquire a category set which is sufficient and adequate to perform well on the discrimination task. If one of these approaches should fail to yield a discriminative category set, this would mean that the approach is a less likely candidate to explain the acquisition of human colour categories. We then look into the sharing of colour categories: can individual learning and genetic evolution account for sharing of categories within a population and across a population? Finally, we need to answer two more questions: How do both strategies cope with a changing environment? And what is the nature of the category sets they produce?

As languages invariably lexicalise colour categories, we should also implement a form of language and study the influence between categorisation and lexicalisation and vice versa. This is where we introduce two more

simulation variants, each extending the previous two. We construct a simulation where categories are learned and lexicalised, and we construct a simulation where categories are genetically evolved and lexicalised. Again we need a task on which to evaluate to lexicalisation. This task comes in the form a communicative interaction where one individual has to name a colour from an array and the other individual has to guess what colour was meant. This task captures the essential rationale of categorisation and lexicalisation and is one of the simplest linguistic interactions between two humans. Both simulation models are thus augmented with lexicalisation and the possibility to evaluate the categorisation and lexicalisation on a "guessing task". We again investigate if the categorical repertoires thus obtained are shared, and if language now has an influence on categorisation.

1.3.1 Justifying computer simulations

The results, suggestions and conclusions presented in this thesis depend heavily on computer simulations. The use of simulations to investigate real world phenomena is common practice in a branch of artificial intelligence known as Artificial Life (for an introduction see Langton, 1989). Some readers might question the use and validity of simulations. Artificial Life (or AL for short) as used in this thesis, is concerned with bottom-up, individual-based modelling. For the purpose of this work we distinguish between a model and a simulation. A *model* predicts or clarifies a phenomenon; often, but not necessarily, on the level of the individual entity; scientific tradition would demand that a model uses mathematical descriptions, such as $F = ma$. A model however can also be specified physically, computationally or verbally; as long as the model is justifiable as being a simplified description of real-world phenomena. A *simulation* is "a model that unfolds over time" (Bullock, 1997). Instead of manually tracing the behaviour of the model over time, the model is run using computer simulations. Simulations allow the researcher to examine larger parameter spaces and allow the study of phenomena at a higher level than at which the model has been designed. These higher-level phenomena are often called *emergent* phenomena (Bhaskar, 1978). Noble, using simulations to study the evolution of animal communication systems, justifies AL simulations as follows (Noble, 1998).

> Imagine that we have observed some real-world phenomenon in animal behaviour, call it E_R, and that we are interested in using an AL simulation to generate a model or theory that might account for it [...] AL simulations model a low level of description explicitly; thus, there will be a set of assumptions or axioms A_M that exhaustively specify the lower-level description of the model. When the simulation is run, let us assume

> that the assumptions A_M give rise to some higher-level emergent phenomenon; call it E_M, and thus $A_M \rightarrow E_M$. The critical point is this: if E_M, the emergent outcome of the simulation, is sufficiently similar to E_R, the real world phenomenon, then it seems reasonable to advance the empirical claim that there exist A_R, real-world analogues of the low-level simulation assumptions, and that, through a similar emergent process, these factors give rise to the real-world phenomenon E_R. That is, $A_R \rightarrow E_R$. The theory is, of course, not assumed to be true, but referred back to empirical biology as a working hypothesis.

The scientific methodology followed in this work is radically different from the descriptive methodology usually followed in human sciences. We use a *synthetic methodology* (Pfeifer and Scheier, 1999), where we try to build a minimal artificial system that mimics functionality observed in nature. Through analysing the components and identifying the minimum requirement for certain phenomena to appear, we hope to shed some light on possible explanations for phenomena observed in nature (Di Paolo et al., 2000). Astronomers have been observing planetary orbits to build an understanding of how the sun and the planets stand in relation to each other. These observations have led to the Ptolemaic system, which placed the Earth in the centre of the universe, and was superseded fourteen centuries later by the heliocentric system of Copernicus. Opposed to this observational methodology, our approach would study planetary motion not through direct observation but through simulating the solar system using a model. The components of the model are based on observations and might be axiomatic in nature. If the simulation exhibits behaviour similar to observed natural phenomena, the model is a candidate for a possible explanation for these phenomena.

There is a strong tradition in Artificial Life to take a constructive approach towards explaining natural phenomena. A typical example would be the work of Reynolds (1987) in which the flocking behaviour of large groups of animals is studied. Flocking behaviour emerges by programming three simple rules in a simulated individual: steer to avoid crowding local flockmates (separation), steer towards the average heading of local flockmates (alignment) and steer to move towards the average position of local flockmates (cohesion).

Recently, computational modelling of complex systems has been used to investigate aspects of language and the evolution of language. Computational models have been constructed to study the emergence of vowel systems (de Boer, 2001), the emergence of signal–meaning mapping (Hurford, 1989; Oliphant, 1996; Steels, 1996a; Oliphant and Batali, 1997; de Jong, 1999), the emergence of compositionality (Steels, 1998b; Kirby, 1999; Batali, 1999; Nowak et al., 2001; Zuidema, 2001). The mapping from signal to

meaning has also been modelled on robotic systems (Yanco and Stein, 1993; Billard and Dautenhahn, 2000; Steels and Kaplan, 2002; Vogt, 2001).

1.3.2 Self-organisation of language

The thesis provides support for the work of Steels (1996a,b, 1997c, 2001b) and others (for example, Hurford, 1989; Oliphant and Batali, 1997; de Boer, 2001) in the field of computational modelling of linguistic phenomena. Steels views language as a self-organising adaptive system. Language is not an individual phenomenon (as individuals do not have perfect knowledge of a language), instead language is sustained in a population of language users and adapts to biological, environmental and cultural constraints. Language is a continuous adaptation and stands in co-evolution to these constraints. Steels does not believe in the innate presence of a language faculty (Chomsky, 1965) or in a strong co-evolution of an innate language faculty and language (Durham, 1991), but believes that language relies on recruiting existing brain faculties. The interactions between perception, cognition and language are highly complex and non-linear, and therefore difficult —if not impossible— to formalise or predict. For that reason computer simulations are the main tool to investigate Steels's claims on the origins of language and related behaviour.

1.4 Contributions

This work is largely interdisciplinary and therefore contributes to many different fields; if not by providing new insights, it does by providing food for thought and discussion.

I propose a computational representation for colour categories. This extends the work of Lammens (1994), where colour categories where represented as parameterised Gaussian bells. In this work, colour categories are represented using adaptive networks, which allow non-convex category shapes and allow for learning and evolution of categories. The categories are defined over a perceptual space, in concurrence with Lammens the CIE LAB space was taken. The chromatic data used for the experiments consists of real-world measurements of colour samples, allowing full control of the conversion to a representation space.

Four variations of the acquisition of colour categories have been implemented, allowing an extensive study of different positions in the debate. This is different from other research into the acquisition of form-meaning (Hurford, 1989; Oliphant, 1996; Steels, 1996a; Oliphant and Batali, 1997; de Jong, 1999) where experiments have always been restricted to studying one approach towards the acquisition of form-meaning associations, thereby leaving issues on the origins of categories unanswered.

The results on individual and cultural learning of colour categories show that it is possible to autonomously learn perceptual categories. Even more, social interaction drives individuals towards having identical perceptual categories. This confirms research by e.g. (Steels, 1997a) and (de Jong, 1999), who have already demonstrated that identical categories can be acquired through population interactions and that assumptions on an innate presence of identical categories are not needed. If categories are not innate, this does not mean that category meaning needs to be copied between individuals to arrive at identical categories. Individuals can reach shared perceptual categories through the indirect coupling formed by linguistic interaction. These results thus support the Sapir-Whorf thesis of linguistic relativity.

Most work in the computational modelling of the evolution of language uses highly abstract meaning representations. Meaning is often represented by a symbol (Hurford, 1989; Oliphant and Batali, 1997; de Jong, 1999) or by a logical expression of atomic symbols (Steels and Kaplan, 1999b; Steels et al., 2002). In the case where meaning is actually grounded in the world, the representation often takes a highly stylised form (Steels and Kaplan, 1999b), there is only a handful of categories to be learned (Vogt, 2001) or the perception is artificial, not resembling human perception, while the environment is modified to reduce noise and real-world artifacts (Steels et al., 2002). This work is different in two aspects. First, the input consists of real noisy measurements of actual colour chips and is therefore more realistic than the input used in most other work on signal–meaning mapping. Second, the representation of the categories allows for fuzzy membership. As mentioned before, most category representations are discrete in nature and can therefore not represent a continuous membership function. In these works either something belongs to a category or it does not, thereby subscribing to the classical view of categories (Armstrong et al., 1983): the binary membership function does not allow for subtle membership evaluation as encountered in perceptual categories.

The results also show that if colour categories are innate, there is the theoretical possibility that the nature of the categories is influenced by their capacity to be linguistically communicated. This is a new viewpoint that has not been investigated yet in the field.

The methods and results described in this work might also be relevant for research on other perceptual categories, such as olfactory categories (Dubois, 2000) or facial expressions (Etcoff and Magee, 1992). It should be interesting to see if other perceptual categories are universal or if there are cross-cultural categorical differences, and if so, if we might find an influence of culture on these categories.

This work also contains some minor contributions:

- Adaptive networks, based in radial basis function networks known

from machine learning, are used to represent perceptual categories. Together with an instance-based learning strategy, they prove to work well for representing categories with fuzzy membership functions.

- A metric has been designed to compute the psychological distance between two point sets of unequal size. The metric performs subjectively better than other measures (such as the undirected Hausdorff metric).
- Measures have been designed to compute the quality of form-meaning repertoires for continuous representation spaces. These already exist for discrete form-meaning repertoires (de Jong, 1999).
- The representations and algorithms to learn colour categories and colour names can also be used to learn from human data.

1.5 Outline

Chapter 2 provides the necessary background for understanding the context of the research. First it introduces the basics of human colour perception: the neural pathways of colour perception, the human eye and the link between trichromacy and opponent-colour theory. The second part of the chapter provides an overview of the positions held in the discussion on colour categories and colour naming and contains an overview of the most influential research in the field.

Chapter 3 gives a concise description of how colour is technically represented, as dealing with a numerical representation of colour is essential for the simulations.

Chapter 4 describes the internal organisation of individuals. An individual is modelled by an *agent*, which is an autonomous simulated entity. The chapter describes how an agent perceives colour stimuli. How perceptual categories, which behave in a prototypical manner, are implemented and what choices have been made in doing so. Finally, it is described how words are connected to categories; words can be uttered by the agents and are of course essential for communicating colour meaning to others.

Chapter 5 is one of the most important chapters. It explains the two simulation variations for *learning* colour categories: one variant uses individual learning, the second variant includes the influence of language and therefore implements cultural learning. The chapter ends by defining measures for judging the quality of the results of all simulations. These measures are used throughout the thesis to compare the different simulation models.

Chapter 6 introduces the two other simulations in which categories are *genetically evolved*. One variant does so without language, the other evolves categories under influence of the linguistic performance of the agents.

Chapter 7 presents results from running two simulations *without* the influence of language. In one simulation the categories are individually learned, in the other categories are genetically evolved. The chapter first introduces the colour stimuli sets used as input for the agents. The chapter also introduces the reader to the mode of presentation, and illustrates how the different measures should be read. Finally, it discusses in how far learning or genetic evolution without the influence of language can account for the phenomena observed in human colour categorisation.

Chapter 8 shows results from running simulation on the influence of language. The learning and the genetic evolution is now under direct influence of the communicative performance of the agents. The simulations are explored under different settings and some of the most important results, concerning the causal influence of language on categorisation, are presented. The chapter concludes by presenting a discussion of how the results might relate to human colour categorisation and colour naming.

Finally, chapter 9 present a discussion of the models and the results obtained by running these. It summarises the main points of the thesis and it critically considers the validity of the simulation and judges its possible contribution to the interdisciplinary debate.

The thesis contains two appendices. Appendix A contains a list of symbols used throughout the thesis and appendix **??** contain colour prints of some of the figures in this thesis.

Chapter 2

Human colour perception, categorisation and naming

2.1 Human colour perception

This section introduces the anatomy, neurophysics and functionality of the human visual apparatus relevant for this thesis; readers familiar with these topics might skip this section.
The eye is the entry-point of visual perception, and often it is –somewhat inexactly– compared to a camera that relays an image of the world to the visual processing machinery in the brain[1]. The physical build of the eye strongly influences visual perception, and an insight into the basics is needed to understand how humans perceive the world. It will also help appreciate the difficulties experienced by scientists in modelling visual perception and colour vision specifically. The first subsection covers the anatomical structure of the human eye. The second subsection describes the neural apparatus relevant to colour vision, which starts at the retina that converts light stimuli to neurophysical signals. These signals are then conveyed along the optic tract to the visual cortex.

Figure 2.1 shows a horizontal cross-section of the brain with the visual pathway. Light falls on the retina in the eye, and initiates a signal travelling down the optic nerve. The optic nerve is a bundle of ganglion cell axons and it preserves the visual field relationships. Humans have about one million of fibres in each optic nerve. At the optic chiasm the optic nerves of both eyes cross. Fibres from the nasal side of the retina (the nasal hemiretina) go to the opposite side of the brain, while fibres from the temporal

[1] Although the camera metaphor is considered wrong, it still proves useful to think of the eye as the input device of visual perception. However, already at the retinal level there is a considerable amount of simple visual processing going on. Only the fact that the information from 130 million photoreceptors of the retina is relayed to the brain using one million nerve fibres, suggest that a considerable amount of data reduction processing is going on at the retinal level.

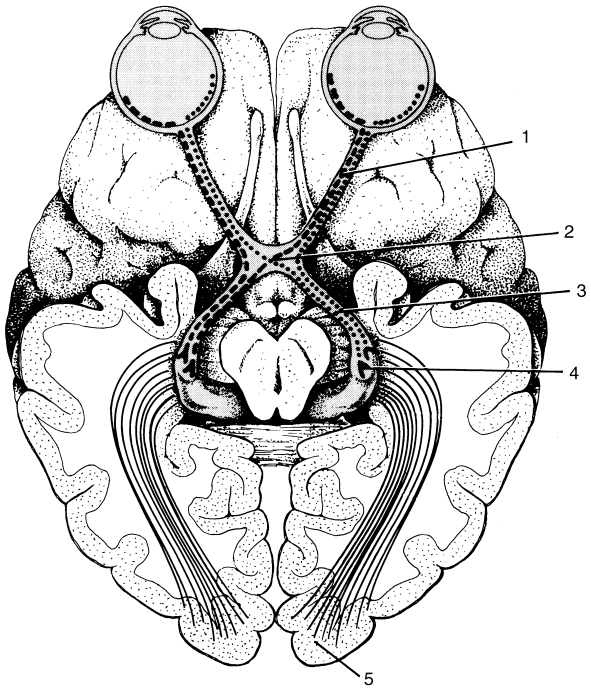


Figure 2.1: The visual pathways, (1) the optic nerve, (2) the optic chiasm, (3) optic tract, (4) lateral geniculate nucleus and (5) the visual cortex. Adapted from (Cotter, 1990).

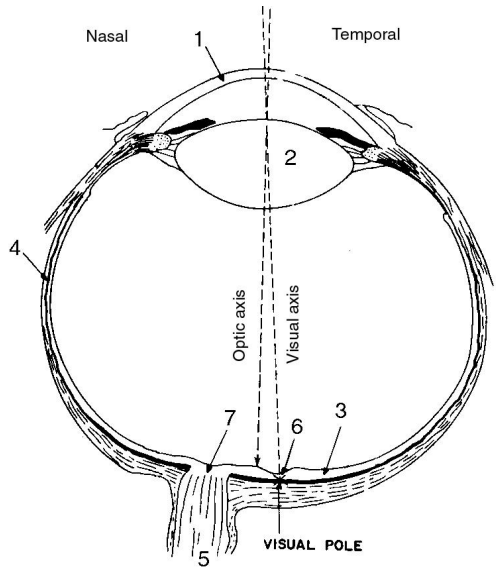


Figure 2.2: Schematic horizontal section of the human eye, (1) cornea, (2) eye lens, (3) retina, (4) sclera, the outer coat of the eyeball, (5) optic nerve, (6) fovea, (7) papilla, where the blind spot is located. From (Kaiser and Boynton, 1996).

side of the retina do not cross. The fibres that cross and the ones that do not cross are brought together in the optic tract, which projects predominantly to two lateral geniculate nuclei (LGN), one in each hemisphere. From there a number of postsynaptic fibres go to the visual cortex. In the visual cortex we find more parallel visual streams, but they are not of concern for this work; detailed accounts can be found in for example (Cornsweet, 1970; Cotter, 1990; Leibovic, 1990; Zeki, 1993; Arbib, 1995; Kaiser and Boynton, 1996).

2.1.1 Optics of the eye

Figure 2.2 shows a schematic cross section of the human eye. The eye optically functions as a camera to project images of the world onto the retina. The cornea is the outer shell of the eye through which the light first passes. The index of refraction of the cornea with respect to air is large and the cornea is curved, making it the most important part of the eye for focussing an image of the world onto the back of the eye. The lens is a flexible structure that has a gradient index of refraction: the middle of the lens refracts more than the sides of the lens, this is used to reduce chromatic aberrations that occur in lenses having a constant refractive index. The shape of the lens is controlled by the ciliary muscle, which makes the lens thicker when fo-

cusing on nearby objects, and thinner when focussing on far away objects. The lens gradually loses its flexibility and by the age of fifty the lens is often less able to focus on nearby objects, a condition known as presbyopia.

The lens also plays a role in colour vision. It is translucent in newborns but gradually becomes more yellow with age, filtering out light of shorter wavelengths. This is why older people are less sensitive to blue and purple light. Older subjects when asked to describe colour samples therefore often describe them as being redder.

The iris is the sphincter controlling the aperture of the eye. Its colour is due to the melanin within the iris. The iris is perforated by the pupil, which allows for light to enter the eye. The size of the pupil depends mainly on the illumination level of the scene. However, pupil size can also be influenced by nonvisual phenomena, such as a certain state of mind (e.g. arousal or fear) or intoxication (e.g. drugs). The pupil can vary from 3 mm diameter to 7 mm. This corresponds to a factor five change in pupil surface, and accordingly to the same amount of retinal illuminance. However the sensitivity change caused by opening the iris is limited by the refraction of marginal rays, only central rays reach the retina undistorted[2].

The area between the cornea and the lens is called the aqueous humour, and the largest volume of the eye is taken up by the vitreous humour, which lies between the lens and the retina. Both are optically clear, and slightly under pressure to guarantee a stable shape of the eyeball.

The retina

The light that enters the eye is focussed on the retina: a thin structure formed by light sensitive cells and connecting neural circuitry[3]. The light sensitive cells or *photoreceptors* are responsible for converting light stimuli into chemical and electrical stimuli, which can be relayed to further stages of visual processing in the retina and the brain. Behind the retina is the pigmented epithelium, a dark layer that serves to absorb any light not caught by the photoreceptors, preventing the light from reflecting back and blurring the retinal image.

The photoreceptors convert absorbed light energy into a neural signal. After being grouped and processed by a number of layers of neurons, the visual signal is then transmitted to the visual cortex. Two kinds of photoreceptors exist: *rods* and *cones*, the names are derived from the shape of the cells. The rods and cones contain pigment molecules, or *opsins*, each with a particular sensitivity to a part of the visual spectrum. Rod and cone receptors serve very different functions. Rods are involved in low light vision (called *scotopic* vision), their reaction time is slow, but their sensitivity is

[2] This is called the Stiles-Crawford effect.

[3] (Dowling, 1987) thoroughly describes the human and other vertebrae retinas.

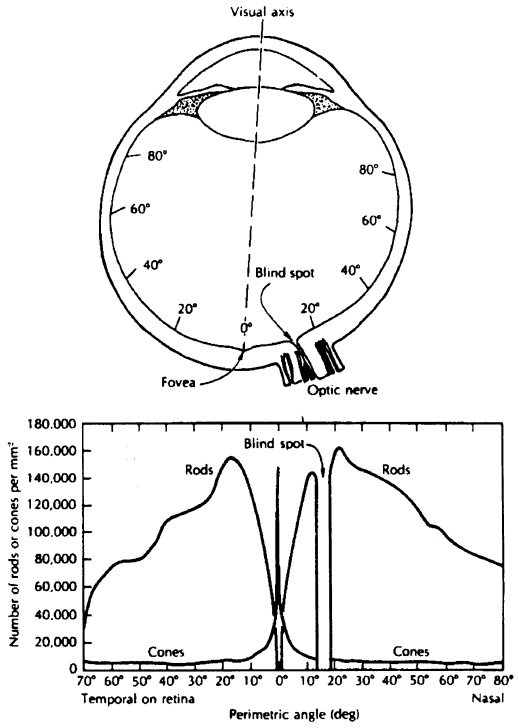


Figure 2.3: Rod and cone density as a function of the retinal location; clearly showing how there are hardly any rod photoreceptors in the fovea, while the cone density is very high. (from Cornsweet (1970)).

magnitudes higher compared to cones. Cones operate at much higher levels of intensity (dubbed *photopic* vision), have faster response characteristics and are used to convey colour information. Though the characteristics of rods and cones are quite apart, their signals are merged in the first neural layers of the retina in all vertebrates and transmitted over the same pathways to the brain. Only the fovea is an exception: hardly any rod receptors are present and foveal vision is performed almost exclusively by cone receptors, figure 2.3 illustrates this. The fovea is the high-resolution region of the retina providing the best spatial and chromatic visual perception. When a subject fixates, the head and eyes are moved such that the object of fixation falls on the fovea. For more on these topics the reader is referred to (Wyszecki and Stiles, 1982; Dowling, 1987; Slaughter, 1990).

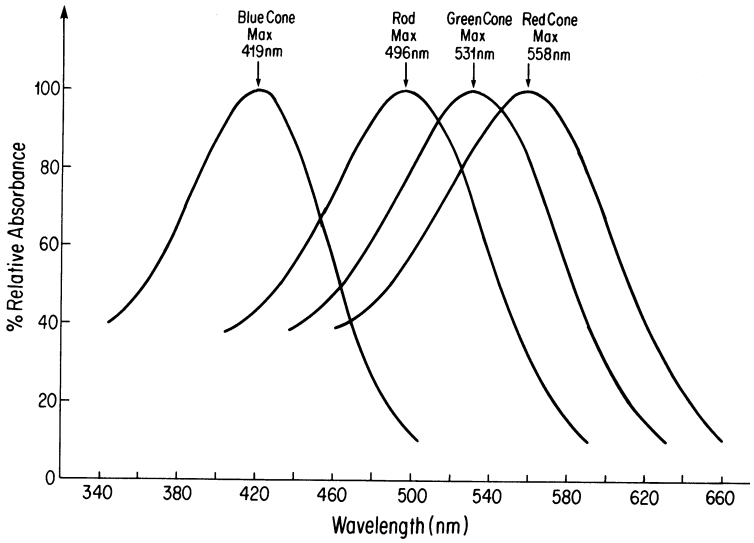


Figure 2.4: The relative absorbance of the four human photoreceptor pigments. The rod pigment has a maximal absorbance at 496nm, the cone pigments each have a maximum corresponding to a sensitivity for reddish, greenish and bluish light. From (Leibovic, 1990, p. 19).

2.1.2 Mechanisms of colour perception

The basis of colour vision lies in the different spectral sensitivities of the three cone receptors. This fact was already recognised about 200 years ago, and has been dubbed the *trichromatic theory* of colour vision. It relates how colour can be discriminated through relative activation of each of three receptors, the condition for this being that each photoreceptor is sensitive to a different part of the visual spectrum. The three cone receptors are called the L, M and S cones, designating their sensitivity to long, middle and short wavelengths. This roughly corresponds to red, green and blue. Figure 2.4 shows the relative absorbance for the four human photoreceptor pigments. The L receptor has its maximum sensitivity at 558nm, the M receptor at 531nm and the S receptor at 419nm. The spectral tuning of the three classes of cones is very broad. In fact, the L and M cone are sensitive to the whole range of the human visual spectrum. The spectra of the cones overlap each other such that each wavelength in the visual spectrum will yield a unique ratio of excitation for each of the three cones.

Visual perception requires a source of illumination and a surface on which the light is reflected. It is the partially reflected light which gives surfaces its colour appearance. Here it should be noted that colour only contributes only for a limited part in our visual perception, texture and

shading alone are enough to perceive most of the visual world (otherwise black and white television would have never sold). Surfaces vary widely in their spectral reflectance characteristics. Human colour perception relies on the diffuse spectral reflectance of surfaces, this is the percentage of diffusely reflected light in function of the wavelength. Every surface also has a specular component, which reflects the source of illumination unaltered. A mirror, for example, is a surface with full specularity. The specularity however does not contribute to the chromatic content experienced by an observer. When subjects try to judge the colour of surface, they actively try to avoid specular reflectance. The polished body of a car for example contains specular reflection which observers first must avoid before making a judgement on the colour of the car (Boynton, 1990).

2.1.3 From trichromacy to opponent channels

The trichromatic theory of colour perception is a based on physiological facts, but at the end of the nineteenth century Hering proposed a psychological theory of colour perception the *opponent-process* (or opponent-colours) theory. In this theory hue is described in terms of its redness or greenness and its yellowness or blueness (Wyszecki and Stiles, 1982, p. 451). Hering noted that in describing colour perception one never uses terms like reddish-green or bluish-yellow, while other combinations of green, red, blue en yellow are all possible. After experimenting with afterimages (the afterimage of red is green and of blue is yellow and vice versa) he concluded that red is opponent to green and that blue is opponent to yellow.

Jameson and Hurvich (1955) described the opponent-colours theory quantitatively through hue cancellation experiments with normal observers which allowed measurement of the relative sensitivity of the opponent hue responses. Together with proof from other disciplines a consensus emerged to reconcile the trichromatic theory with the opponent-process theory. The receptors are indeed trichromatic, each sensitive to a different part of the spectrum, and are then recombined in the retina and perhaps in later stages of the brain to form three opponent channels (figure 2.5). The output of the three receptors is summed to produce an achromatic response ($L + M + S$). Differencing the L, M and S-signals produces the red-green ($L - M + S$) and yellow-blue ($L + M - S$) opponent channels[4].

So, opponent-colour processing is important for decorrelating the cone signals: as the responses of the cone cells are not sufficient for seeing colour,

[4]It should be noted that this depiction of opponent-processing is very simple and therefore also in many ways incorrect. In the field of neurophysiology there is still no clear picture available of how cone and rod cells connect to ganglion cells. For example, for the blue/yellow pathway nine different ways have been found in which the cone and rod outputs are connected to it (Gouras, 1984)

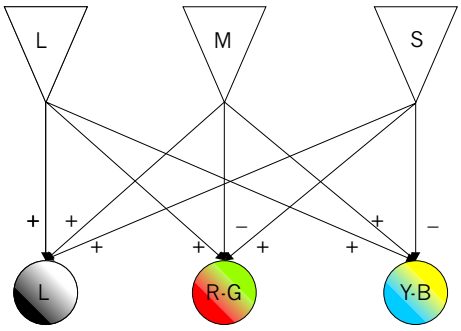


Figure 2.5: Encoding the cone signals into opponent-process signals. Adapted from (Fairchild, 1998).

the difference between the cone responses is what makes us see colour. The opponent coding also provides an efficient coding and reduces noise in the transmission of the signals to the brain. Opponent processing cells have been found in the lateral geniculate nucleus (De Valois et al., 1966), in the retina (DeMonasterio and Gouras, 1975) and in the visual cortex (Vautin and Dow, 1985). Opponent coding has been reproduced in computational experiments by Wachtler et al. (2001). They found that dimensionality reduction of visual scenes yields channels that correspond to opponent colour processing. Opponent-colour also features prominently in the formulation of colour appearance models. These are models that capture the psychological experience of human colour perception (Fairchild, 1998; Brainard, 2001).

2.2 Colour categorisation and colour naming

Research into the division and the naming of the spectrum takes off with Rivers's turn of the century expeditions[5]. Rivers undertook an extensive study of colour naming and colour vision of Australasian, African and Indian natives (Rivers, 1901, 1903, 1905). After observing how languages of certain primitive tribes have less colour terms than Western languages, he suggested a parallel between mental and social development of cultures and their colour lexicalisation. In over a century of colour categorisation research, two major schools can be distinguished: the *relativist* and the *universalist* school. The relativists believe that colour categorisation and naming is not innate but learned by every individual. The spectrum visible to hu-

[5]Even earlier, the German ophthalmologist H. Magnus studied colour vocabulary (see (Berlin and Kay, 1969; Kay et al., 1991; Dedrick, 1998)). He reports that some languages might name two fundamental colours with one single name.

mans is continuous, and there are no natural or physical divisions present in the spectrum. So any division of the spectrum must be the product of our environment, our culture and our language.

> There is a continuous gradation of colour from one end of the spectrum to the other. Yet an American describing it will list the hues as red, orange, yellow, green, blue, purple, or something of the kind. There is nothing inherent either in the spectrum or the human perception of it which would compel its division in this way. (Gleason, 1961, quoted in Berlin and Kay, 1969, p. 159).

The universalists, who find themselves at the other side of the anthropological battlefield, believe that the nature of human colour perception is such that colour categorisation is not at all arbitrary. They explain the apparent panhuman divisions of the spectrum through the genetically specified neurophysiological structure of colour perception. Of course, as in most cases the truth probably lies somewhere in middle. Critical voices in the last years advocate a view wherein the neurophysiological structure of colour perception as well as cultural influences are used to explain colour categorisation and naming. This section describes some of the milestones in colour categorisation and colour naming research.

2.2.1 Universalism

The universalist school sticks to the epistemological position of phenomenal absolutism, in which the division of the spectrum and the relation to colour words only depends on the underlying biological structure of colour perception. Universalism is embraced by many students of colour categorisation (see (Allott, 1974) for an overview) and colour categories and the corresponding colour terms have become text book examples for nativism (Clark and Clark, 1977; Durham, 1991; Crystal, 1997). The following quote illustrates the universalist position well.

> [R]egularities in the linguistic encoding of color result from regularities in the neural coding of color in the brain, with the implication that this is a case of genetic mediation. (Durham, 1991, p. 218).

Berlin and Kay

In 1969 Brent Berlin and Paul Kay published their influential monograph "Basic Color Terms: Their Universality and Evolution" (Berlin and Kay, 1969). In this work, which rekindled the colour discussion in anthropology, psychology and philosophy, Berlin and Kay set out to show that colour

terms are universal across languages and that there is a specific order in the emergence of colour terms. The climate in (American) anthropology and linguistics at that time favoured extreme linguistic relativism, a principle where every language user forms its own unique coding of experience. Semantic universals do not exist in the view of extreme linguistic relativism. For colour categorisation this would mean that every individual and every culture segments the colour continuum in its own particular way. Berlin and Kay however suspected that colour naming was not as variable between different languages as the extreme relativists wanted the community to believe. They tested their hypothesis and came up with remarkable results and conclusions which still, after thirty years, provoke a good deal of debate.

Berlin and Kay used procedures described by Lenneberg and Roberts (1956). They first established the operational definition of a *basic colour category* or BCT for short. A BCT is a colour word that is monolexemic (its meaning may not be predicted from its parts), its significance is not included in any other colour term, its application can not be restricted to a narrow class of objects, it must be psychologically salient to all informants, terms should have the same "distributional potential" (an example of this would be that in English reddish, whitish and greenish are valid words, but aguaish is not. So red, white and green are candidate BCTs and agua is not), colour terms can not refer to objects (so gold and lemon are no BCTs), recent foreign loan words are no BCTs, and a BCT should be morphologically uncomplex (so blue-green is no BCT) (Berlin and Kay, 1969).

The basic colour terms centre on the chromatic visual experience and are purely non-contextual; no reference to any object or situation is present. This excludes colour words such as crimson, blond, bluish, gold, lemon-coloured, etc. The maximum number of BCTs found was eleven (English, for example, has eleven BCTs). Some languages have less BCTs, sometimes as few as two or three. This does not mean that they cannot discriminate more colours, only that they have split up the spectrum in fewer categories.

For each language under investigation Berlin and Kay first identified the basic colour terms and informants were then asked to (1) indicate all colour chips that would, under all conditions, be named with a particular basic colour term, and (2) to identify the best example of that colour term. They thus examined twenty native-speakers[6] of unrelated languages. The informants used a colour chart with 329 colour chips, as in figure 2.6. The chips were selected from the Munsell colour space (Munsell, 1976): 320

[6]The twenty languages are Arabic (educated Lebanese), Bahasa Indonesian, Bulgarian, Cantonese, Catalan, American English, Hebrew, Hungarian, Ibibio, Japanese, Korean, Mandarin, Mexican Spanish, Pomo, Swa-hili, Tagalog, Thai, Tzeltal, Urdu or Vietnamese as their native language. Of the twenty languages studied, nineteen were studied by interviewing one bilingual speaker only: only Tzeltal was studied in the field (Durbin, 1972; Saunders and van Brakel, 1997).

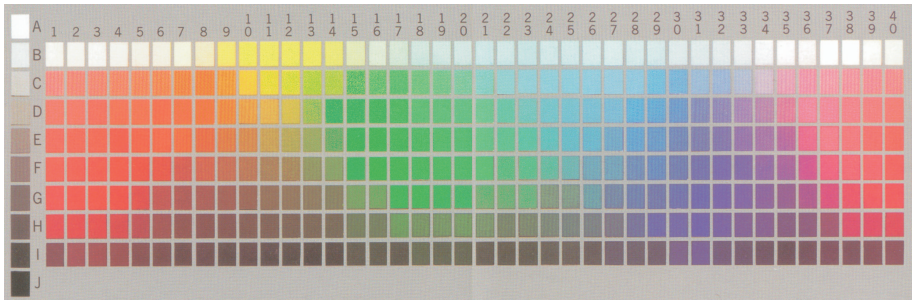


Figure 2.6: The chart with Munsell chips as used by Berlin and Kay. Scanned from the 1999 paperback edition of (Berlin and Kay, 1969).

chromatic chips (8 degrees of brightness of 40 equally spaced and fully saturated hues) and 9 non-chromatic chips (from white, over grey to black). Though there was little agreement between the limiting boundaries of the basic colour terms in each language, the results showed notable agreement between the colour chips selected as *best examples*. These best examples clustered in discrete regions on the colour chart, meaning that foci of basic colour terms agreed well for each language. Figure 2.7 shows the foci of the BCTs of the twenty languages reported. The concurrence is striking: not only do the foci lie in clustered regions; the areas enclosing the foci also don't overlap (only in the case of GREEN and BLUE there are BCTs which seem to encode for bluish green).

A second major finding was that there exists a particular order in which languages "evolve" basic colour terms[7]. Not all languages have the capacity of eleven BCTs, some have two, some have three or five BCTs. In the case where a language does not have all eleven possible BCTs, it is in general possible to predict which colour terms it will have, and which it will not have. Berlin and Kay summarised this in the *evolutionary order* of colour terms.

$$\left[\begin{array}{c}\text{white}\\ \text{black}\end{array}\right] < [\text{red}] < \left[\begin{array}{c}\text{green}\\ \text{yellow}\end{array}\right] < [\text{blue}] < [\text{brown}] < \left[\begin{array}{c}\text{purple}\\ \text{pink}\\ \text{orange}\\ \text{grey}\end{array}\right]$$

All languages have at least two BCTs, one for black and one for white. If a language contains three terms, it contains a term for red. If a language

[7] H. Magnus in 1877 was the first to note a certain evolutionary order in colour terms of primitive languages, commenting on his cross-cultural research he says "Practically every language has a name for red, nearly all have a name for yellow; but comparatively few have a conventional word for green, and still fewer have one for blue" (from (Bornstein, 1975))

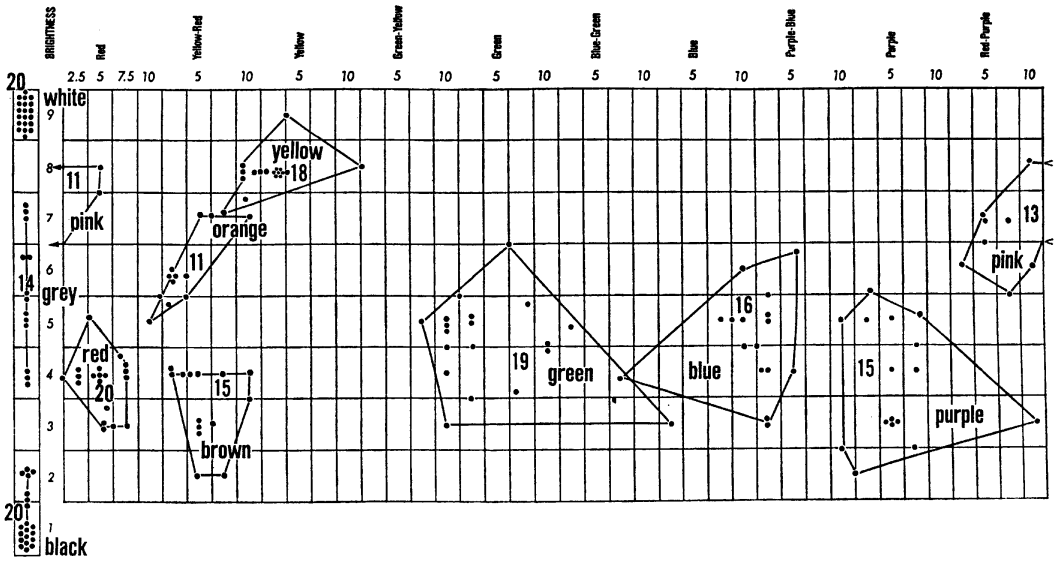


Figure 2.7: Overview of Berlin and Kay's results for 20 languages. The numbers and colour terms on the axes refer to Munsell hues and values. Points on the chart represent the centres of gravity of foci for each of the 20 languages; the lines show the hull of the foci for each colour term. The number inside the hull stands for number of languages encoding the colour category. From (Berlin and Kay, 1969).

contains four terms, it contains a term for green or yellow, but not both. If a language contains five terms, it contains terms for both yellow and green. If a language contains six terms, it will contain a term for blue. If it contains seven terms, it will contain a term for brown. If it contains eight or more terms, it will contain a term for purple, pink, orange or grey; or a combination of these.

It is worth noting that in languages having few colour terms the areas delimiting the applicability of each colour term are larger than in languages with more colour terms; although the foci stay more or less fixed. So in languages with few colour terms, the term for BLACK does not correspond with "black" as English speakers understand it. BLACK actually covers colours having qualities such as "dark" and "cool", while WHITE covers "light" and "warm".

Criticism on Berlin and Kay

The experimental procedures, the results and the conclusions of Berlin and Kay have been a lightning rod for criticism. Their work, together with that of Rosch (see 2.2.1), forms the basis for the universalist colour language tradition. Universalists believe that there is a limited set of basic colour terms, that some colours are psychologically salient while others are not and that colour categories are shared between all cultures because of the universal genetic coding of colour categories. Berlin and Kay belong to this tradition and took the basic colour system to be innate and biologically constrained, and tried to map every language onto their evolutionary sequence theory. The major points of criticism on their experiments and conclusions are summarised here.

The definition of BCTs, for most languages, excludes terms that would be considered as colour terms by the speakers of that language (Sahlins, 1976). Berlin and Kay regard BCTs as being detached from contextual semantics; while in many languages colours are not only distinguished according to their visual appearance, but also according to the object or objects they belong to. "Blonde" for example clearly is a colour, albeit only associated with hair and beer. Nevertheless "blonde" is very salient throughout the English speaking community. The definition of a BCT is sometimes vaguely interpreted. For example, some Slavic languages have two BCTs for blue; and Russian depending on the interpretation of the BCT definition has zero, one or five BCTs for purple (Corbett and Morgan, 1988). Dournes (1978) even reports that Jörai (spoken in Vietnam) has 23 BCTS. Every language was found to have at least two BCTs, while the maximum number of BCTs was found to be eleven. Not coincidental, English has eleven BCTs and is considered by Berlin and Kay to be at the most evolved stage of the sequence. This relation to American English permeates all experiments and results of Berlin and Kay. A very important criticism of Berlin and

Kay's methodology was their wish to fit the results to the evolutionary sequence of colour categories: if the informant did not pick out black and white as one of their first colour categories, the researchers assumed that a black and white category existed anyway. Furthermore, if a subject picked a colour sample of a green-blue shade (some languages do not have a separate term for green and blue), the researchers consequently noted that the subject had picked the GREEN category, since GREEN comes before BLUE in the evolutionary sequence. Also the choice of experimental subjects in the original experiment was poor: nineteen of the twenty informants for the original research were exchange students living in the United States and therefore spoke English well, which might have biased the results. Furthermore, many examples have been found that do not comply with the conclusions of Berlin and Kay. There exists a language, Shuswap (spoken at the northwest Pacific Coast of America), that has a word form which is used for yellow as well as green (MacLaury, 1987); which is not predicted by Berlin and Kay's theories. Another example is the saliency of red. As predicted by Berlin and Kay, red is one of the most important and early colour categories. But following their argumentation, blue should be a very salient colour as well. However, a term for blue is rather rare in the world's languages, which is not in accord with universalist theories or neurophysiology (Davidoff, 1991, p.153). Finally, the experiments and their presentation convey a general impression of sloppiness, which only undermines the confidence held in the conclusions (Saunders and van Brakel, 1997).

Rosch

Eleanor Rosch (at that time Eleanor Heider) (Heider, 1971, 1972; Heider and Olivier, 1972) developed the prototype theory of categorisation in reaction to the classical theory of categories, which stated that no member of a category has any special status. The classical theory stated that there is a delimiting set of necessary and sufficient attributes to define a category, but Rosch suspected the opposite. In her research on the Dani, a stone-age tribe from Irian Jaya (New Guinea), she used the fact that the Dani only know two colour terms, *mili* (dark-cool, including black, green and blue) and *mola* (warm-light, including white, yellow, red, orange, pink and red-purple), to investigate her theories on categories.

She showed that the Dani speakers had psychologically salient colour foci even though those colours were not named, thus challenging Whorf's thesis that language moulds the way in which one conceptualises the world. If Whorf's thesis would be correct then the Dani should not react to colour categories that are not lexicalised. In a series of experiments, Rosch showed that the Dani had colour prototypes that were not named. In one experiment it was demonstrated that the Dani were better at learning new, made-up colour terms for focal colours than for non-focal colour. If

colour categorisation is relativistic then the Dani should, on the contrary, have equal difficulty learning new colour terms. One group was taught eight new terms for focal colours, another group was taught terms for non-focal colours; learning terms for focal colours was easier. In another experiment, it was shown that it was easier to remember focal colours than non-focal colours (Heider, 1972). In yet other experiments the Dani categorise the colour continuum in much the same way as English speaker do (Heider and Olivier, 1972) and children seemed to have an clear preference for focal colours during random selection of colour cards and were better at remembering focal colours during memory games (Heider, 1971). These focal colours act as natural prototypes of a colour category: they have special status acting as the best example of the whole category (Rosch and Lloyd, 1978).

Rosch refined the conclusions and the notation of Berlin and Kay. In their stage I language, when only two BCTs are present, they have a term for BLACK and a term for WHITE with their focal colour on the black chip and on the white chip. Berlin and Kay saw the partition between BLACK and WHITE as a strict division between dark and light colours. Rosch refined these two categories to a category for DARK-COOL colours and one for WARM-LIGHT colours. She also observed that there is not one single focus for these categories; subjects, when asked to point out the best example of a category, choose different saturated colour chips.

Comments on Rosch

Particulars and conclusions of Rosch's experiments raise some questions (Ratner, 1989; Saunders, 1995; Saunders and van Brakel, 1997). It seems that the two colour terms mili and mola do not exactly translate as BLACK or WHITE, or even as DARK-COOL and WARM-LIGHT. Also, the foci of mili and mola are not the same for all subjects; for mola some subjects choose red (69% of the subjects), some yellow and some white as being the focus. It also seems that mili and mola are not uniquely used for referring to colour, but that both words have are more evaluatory meaning. Some methods of the experiments are also subject to criticism: the Dani were taught new colour terms, but they would not accept nonsense words, so they were taught indigenous words "reflecting the perceived world structure". Furthermore, during the memory experiments the Dani were markedly worse at remembering colour chips than Americans were, which just argues *against* the universality of prototypical colours.

Lucy and Shweder (1979) comment on the memory experiments by Rosch, in which subjects were asked to keep a colour chip in memory and then after a certain period of time point it out on an array of colour chips. She found that both for the short-term memory and the long-term memory experiments the subject remember focal colours better than non-focal

colours[8]. Lucy and Shweder claim that the colour array used by Rosch to demonstrate the influence of focality on memory is biased in favour of focal chips, meaning that the focal chips were easier to find in the test array, thus influencing the experiment. Lucy and Shweder removed the chips that were responsible for the fact that some chips were easier remembered and rearranged the array. With their new array they reimplemented Rosch's memory experiments and found that focality did "not relate at all to short-term memory".

Davidoff and his co-workers (Davidoff et al., 1999; Roberson et al., 2000) also replicated Rosch's experiments, using subjects from the Berinmo tribe of Papua New Guinea. The Berinmo have five colour terms: *mehi* (containing red, orange, pink colours), *wor* (yellow, green colours), *nol* (green, blue, purple colours), *kel* (brown, dark colours) and *wap* (light colours). The Berinmo make a distinction between yellowish green (*wor*) and bluish green colour (*nol*), a distinction that English speakers do not make. On the other hand, English speakers make a distinction between green (*green*) and blue (*blue*) colours, and Berinmo do not. In an experiment subjects were asked to remember a colour from around both the blue-green and nol-wor boundaries: the subject was shown a colour, it was taken out of view and after 30 seconds the same colour was shown together with a similar colour. The subject then had to select the colour it had seen 30 seconds ago. Berinmo subjects were good at cross-category nol-wor decisions, but performed poor at blue-green boundary colours. In contrast, English speaker performed well at green-blue, and poor at nol-wor decisions. This is inconsistent with the universal position on colour categorisation. In other experiments English speakers and Berinmo speakers were asked to learn new names for colour categories. Both groups of subjects did not have any trouble learning new names for colours that were already named in their language, but had significantly more problems learning names for colour that were not previously linguistically distinguished. Davidoff concludes that there is a "considerable degree of linguistic influence on colour categorisation", giving support to the relativist theories of categorisation.

The biological basis for universalism

The theories and facts about opponent processing of chromatic stimuli (see 2.1.3) form solid ground for universalist theories on the origins of colour categories (Bornstein, 1975; Ratliff, 1976). If colour perception is considered identical for all humans with normal colour vision, the conversion of thrichromatic perception to opponent channels in the lateral geniculate nucleus and the visual cortex (Zeki, 1983, 1993) strongly suggests that there

[8]With focal colours the eleven fundamental neural response categories are meant: BLACK, WHITE, RED, YELLOW, GREEN, BLUE, BROWN, ORANGE, GREY, PINK and PURPLE.

could be two elementary achromatic categories (black and white), and four elementary chromatic categories (red, green, yellow and blue). Appropriate mixing of these six elementary categories produces a total of eleven basic categories (the previously mentioned categories including orange, grey, purple, pink, brown).

Kay and McDaniel (1978) conclude that "the semantic structure of these [the basic colour] categories can be derived directly from the neural response functions". This is, however, wildly optimistic and supported by evidence from other disciplines they jump to conclusions that are difficult to hold. As Dedrick (1998) puts it "It involves a gigantic leap from the results of electrophysiology to the names that (some) people use for colours!". The relation between psychophysics and neurophysiology is a problematic one. For example, the opponent-process theory predicts that at a certain point in the spectrum neither red nor green should be perceived, as both red and green cancel each other out. As the green-red channel is not responding; the perceived colour should only be influenced by the yellow-blue channel. But work by Derrington et al. (1982) show that these cross-over points are at a very different location than what psychophysics predicts. For example, the cross-over point of L-M cells is expected to correspond to a yellow sensation, but actually corresponds to a colour which can be best described as chartreuse. The wavelengths corresponding to unique colours do not correspond to the wavelengths predicted by the maximum sensitivity of the different opsins in cone receptors (Saunders and van Brakel, 1997). Neurophysiology has only explained a very limited part of human colour perception, and until we know more on how the brain processes colour it seems presumptuous to draw any parallels with psychological and linguistic phenomena.

More support for universalism comes from experiments with infants (Bornstein et al., 1976). Infants, like adults, exhibit categorical perception of spectral colours, without even being able to speak a language. Even pigeons have been reported to categorise the spectrum, although not into the same categories as humans do (Bornstein, 1975).

2.2.2 Relativism

Relativists often adhere to cultural relativism, a perspective in which colour semantics are defined pragmatically and ecology and culture form the only pressure to categorise colour perception. The relative nature of colour categories fits within the framework of *linguistic relativism*, revived by the theories of Whorf (1940, 1956). Together with his mentor Sapir, Whorf explored the idea that an individual's world view is influenced by the language he speaks. The first tenet of the Sapir-Whorf thesis claims that the world is differently experienced and conceived in different linguistic communities; "structural differences between language systems will, in general, be par-

alleled by nonlinguistic cognitive difference, of an unspecified sort, in the native speakers of the two languages". (Brown, 1976, p. 128). The second tenet states that language causes a particular cognitive structure, "the structure of anyone's native language strongly influences or fully determines the world-view he will acquire as he learns the language" (Brown, 1976, p. 128). The *strong* thesis is an extreme interpretation claiming that cognition is completely shaped by language and that semantic systems of language are not constrained by biology, as paraphrased Kay and Kempton (1984, p. 66), "The semantic systems of different languages vary without constraint". The *weak* Sapir-Whorf thesis claims that language has a causal influence on cognitive structures, but cognition is also embedded in a biological basis. Thus, non-linguistic cognitive judgements will be influenced by the individual's linguistic structure *and* by his or her biology. It is generally agreed that the strong thesis is difficult to hold, but still it is often the strong thesis that is attacked by proponents of universalism.

Brown and Lenneberg

Colour forms the ideal test bed for investigating linguistic relativity. It is a well defined domain and allows for uncomplicated experimental procedures to investigate the influence of language on the perception and division of colour. The colour space is divisible into millions of distinguishable differences, but only a handful of colour categories exist which are commonly used and named. As many non-Western cultures have an unexpected manner of categorising colour (as compared to Indo-European cultures), colour categorisation is ideal for investigating the relation between cognition and language. For example some languages, such as Tarahumara spoken in northern Mexico, have only one name for the colours GREEN and BLUE (dubbed GRUE in the literature), while English has two separate categories for GREEN and BLUE. English speakers when asked to name a green-blue colour would doubt, and would probably resort to a descriptive label as green-blue, greenish blue or turquoise, while language users having one single category and term for GRUE, would name this colour with ease.

Brown and Lenneberg (1954) implemented a very influential experiment that tested wether naming of colours could be related to non-linguistic properties of colour perception. Their main motivations were the theories of Sapir and Whorf[9]. Brown and Lenneberg tested English speaking students by letting them name colour stimuli. They devised a measure, the *codability*, based on the length of the given colour description, the response time and the agreement with other subjects and with themselves on later trials. A red stimulus for example has a higher codability than a

[9]For a historical perspective on Brown and Lenneberg's research the reader is referred to (Brown, 1976)

sample hesitantly described as "yellowish green". In their experiment they first measured the codability of selected number of colour samples. These colour samples were then used in a non-linguistic memory experiment, to investigate the correlation between the codability of colour samples and the ease of remembering colour samples. They indeed found a correlation and concluded that codability was related to the subjects' ability to recognize colours. This led to the conclusion that lexical differences were indicative for cognitive differences. Ironically, Brown later converted to universalism after being confronted by justified criticism on his work and by the strong research of Berlin and Kay and of Rosch (Brown, 1976).

Other naming and memory experiments

In the wake of Brown and Lenneberg's 1954 article other work was published which refined and extended the original experiments. Lantz and Stefflre (1964) refined the experiment by using *communication accuracy* instead of codability. Lenneberg (1961) himself had found some anomalies in his results compared to research by Burnham and Clark (1955), which could be explained by problems of the definition of codability. Lantz and Stefflre's communication accuracy is a measure of how accurately a colour can be communicated. If one person names a colour, and that name is given to another person, how accurately can he pick the same colour from an array of colours. The experiments showed that colours with a high communication accuracy we better remembered in memory experiments. The communication accuracy was even better at predicting the ease of remembering colours than the codability measure of Brown and Lenneberg. These results were later reconfirmed in a different language, namely Yucatan (Stefflre et al., 1966). Lantz and Stefflre conclude that the results of their study can be interpreted as "evidence of the influence of language on nonverbal behaviour". In the light of their work we should again mention Heider and Olivier (1972) who tried a very similar experiment with the Dani and found that even though the Dani only possessed two colour terms, they were much better at remembering focal colour stimuli; thus, according to Rosch, language does not influence memorability at all. But remember that Lucy and Shweder (1979) and Davidoff et al. (1999) were able to refute this.

Kay and Kempton (1984) demonstrated a clear Whorfian effect in an experiment where subjects had to judge perceptual distance between colour stimuli. They used speakers of two linguistic communities, one language (English) had two colour terms for GREEN and BLUE and the other language (Tarahumara) had only one colour term covering both GREEN and BLUE stimuli. The idea was that speakers with a separate category for GREEN and BLUE will perceive these stimuli as more different than speakers having only one term. They indeed found that English speakers exaggerated the subjective distance between green and blue colours, which

cross the English lexical category boundary. While Tarahumara speakers found the distance between green and blue colours to be not larger than for example between two different shades of green or blue. But when Kay and Kempton made the task more complex to try cancelling out the possibility that the subjects would use colour terms for judging the distance, they found that both English and Tarahumara speakers apply an identical subjective distance to colours. Nevertheless, in their conclusion they admit that a "modest Whorfian effect" is supported by their experiments.

Various evidence for relativism

Other direct and indirect evidence for relativism can be found in different disciplines. Some important work is mentioned here.

Sandell et al. (1979) investigated whether macaque monkeys partition the spectrum in the same way as humans do. They showed that monkeys could be trained to react to colour stimuli and that they generalised to other colours which resembles human categorisation: "monkeys appear to partition the spectrum into the same basic hue categories as do humans". They present their results as a demonstration of the innate presence of colour categories, but the very fact that the monkeys had to be trained before exhibiting this colour categorisation behaviour only proves that some mechanism for colour categorisation might indeed be present, but not that categories for blue, green, yellow and red are readily available in the brain.

Other evidence comes from genetics. The genes responsible for coding the opsins in the cone receptors have been isolated. The genes for the S-cone opsins, for picking up short-wavelength light, are very ancient. But the genes for the L and M-cone opsins have a common ancestor and split up only recently in evolutionary history (Nathans, 1989, 1999). This is reason why genetic defects often lead to anomalous colour vision: the maximal sensitivity of the L and M-cones is normally 30nm apart, but genetic defects often produce L and M-opsins of which the spectral sensitivity is much closer together, or sometimes even equal to each other. Nevertheless, people with these genetic defects still manage to exhibit, even although obviously limited, trichromatic colour vision (Neitz et al., 1999). This shows that human colour perception is much more plastic than was ever suspected.

Recently, evidence has been shown that subjects with anomalous colour vision judge colour settings for unique hues (red, green, blue and yellow) that seem to be largely unrelated to their anomalous vision (Webster et al., 2000), suggesting that colour categories are not innate but learned.

The reader might also be referred to work critically considering universalist evidence (van Brakel, 1993; Gellatly, 1995; Saunders and van Brakel, 1997; Lucy, 1997; Davidoff, 2001).

Empiricism: learning categories through biological and environmental constraints

If colour categories are indeed learned, there are still different influences under which these categories can be learned. The *empiricist* position admits that colour categories are indeed shared by humans, but does not suppose that the categories themselves are innate. The categories are learned individually by each human being, but through constraints on biology and the near-universal environmental conditions all individuals arrive at almost identical categories. This view is widespread among "empiricist psychologists" (Elman et al., 1996).

The problem now shifts to finding an appropriate category acquisition model. Various learning algorithms exist which can acquire colour categories from sensory stimuli, despite the poverty of stimulus and without direct linguistic or categorical feedback. The learning algorithms are *individual*, meaning that they are only based on the interaction between the learned and the environment, and *observational*, meaning that there is no feedback from other learners. A typical example of such a learning algorithm would be the k-nearest neighbour algorithm, an algorithm which clusters data in categories according to an instance-based learning strategy (Mitchell, 1997, p. 231). Another example is the work by Lammens (1994) who represents categories as Gaussian bells of which the parameters are optimised to fit to human colour categories.

The problem of *sharing* of colour categories (the remarkable agreement between the categories of individuals) is solved by assuming that everybody arrives at more or less the same categories because they are the natural outcome of induction over the types of stimuli present in a normal environment processed by normal physiological structures. Different cultures might still arrive at different colour categories, either because their environment has another distribution of colour stimuli, for example the individuals live in an environment without technical colour reproduction, or because the physiology of the individuals is different from normal human beings. Sacks (1996) for example describes a community on the Pingelap atoll[10] of which a large population is born colour blind; the achromatopic islanders have developed their own way of perceiving colour through its achromatic shade and texture.

Culturalism: learning categories through cultural constraints

The cultural hypothesis states that colour categories are learned with a strong causal influence of language and that they propagate in a cultural process (Davidoff, 2001). A position favoured by cultural psychologists

[10] Belonging to the Caroline Islands in the Western Pacific.

(Tomasello, 1999) and by those viewing language and its underlying conceptual framework as a complex adaptive system (Steels, 1997c).

Culturalism approaches the problem of category acquisition in a similar way as empiricists. Categories are acquired through a learning process that is identical for all human beings. This learning process is influenced by constraints on physiology and neurology as every individual shares the same physical and cognitive apparatus. It is also accepted that the environmental conditions are shared between individuals, meaning that all individuals in a community have access to the same stimuli. Also ecology, which influences the meaning and importance of the colour stimuli, is experienced identically for all individuals in a single culture. But culturalism argues that these constraints are not enough for the learner to arrive at shared concepts, and so an additional ingredient in the form of strong social interactions is needed. Social interactions assure that the learner gets feedback on what he learns. The social environment of a learner stimulates the learning of certain categories and discourages the learning of others. It provides motivation to learns and gives social and cultural feedback. Language learning forms a major part of this. Interactions in language learning provide the learner with feedback on the appropriateness of the categories and their labels. Steels (1996a) introduced a computational model in which *language games*, linguistic interactions between individuals to name objects in a shared context, are used to arrive at shared categories.

Some might argue that no explicit feedback is given during language learning and so categories cannot be learned by only observing the way in which other language users handle those linguistic categories. However, the feedback need not be explicit. Often feedback is hidden in social interactions (Tomasello, 1988; Baldwin, 1993; Tomasello and Barton, 1994) and other mechanisms can be used to deduce meaning from contextual setting (Smith, 2001). Thus feedback is pragmatic, and has been shown to occur during almost every utterance, even for adult conversation.

The cultural approach tackles the problem of sharing categories by adopting a notion of self-organisation (Steels, 1996b). Self-organisation is familiar concept in natural sciences, particularly in non-linear physics and biology (Stengers and Prigogine, 1986). An often-used example of self-organisation is the path-finding of ants. Ants deposit pheromone when travelling, while at the same time being attracted by pheromone left by other ants. If a path, for example between the nest and a food source, is often traversed the pheromone concentration will be higher on that path, attracting even more ants. This positive feedback loop will produce the most efficient paths between the nest and the food sources and will in the end be used by all ants (Camazine et al., 2001). In a similar way, the feedback loop between the use of a word and its success in conveying meaning to other language users will strengthen the position of the word and its meaning in the population of language users. The more speakers adopt

a word and the underlying meaning, the more successful communication with that word will be and the more other speakers will be attracted to that word and its meaning. It has been shown that a strong coupling of the processes whereby individuals acquire concepts and the process causing the self-organisation of a lexicon leads to a sharing of concepts (Hurford, 1989; Steels, 1996b). For this no central coordination or prior innate knowledge is required.

2.3 Summary

This chapter started with an overview of physiology, neurology and psychology involved in human visual perception. Human colour perception finds its roots in three chromatic receptors, each sensitive to a different part of the spectrum. Psychologically humans however react in an opponent fashion to colour, giving rise to the opponent-colour theories of perception.

The chapter then continues by giving an overview of different positions in colour categorisation and colour naming research. Everyone agrees that different languages use different colour words, and that colour categories have a focal point in perceptual space with a fuzzy boundary. But here consensus ends. The *universalist* position argues that colour categories are genetically determined and therefore identical for all human beings (this is analogous to *nativism*). Others believe that colour categories are learned, the *relativist* position argues that category learning is influenced by constraints on biology, the environment, the ecology or culture. Two tendencies can be distinguished: *empiricism* and *culturalism*. Empiricists claim that colour categories are learned individually but become shared among individuals through environmental, biological and ecological constraints. Culturalism extends this by claiming that social interactions are required to explain the apparent shared nature of colour categories. Table 2.1 summarises the different positions in the debate.

Position	Acquisition	Sharing
Universalism / Nativism	Genetic expression during development	Gene propagation
Empiricism	Individual learning	Similar environment, ecology and physiology
Culturalism	Social and cultural learning	Similar environment, ecology and physiology *with* cultural self-organisation

Table 2.1: Summary of positions in the origins of colour categories debate.

Chapter 3

Representing colour

The visual world arrives at our eyes as a spatial and temporal distribution of electromagnetic energy over wavelengths in the visual spectrum. In that mishmash, information on texture, form and colour of the world can be found. Colour is a perceptual phenomenon, not a physical phenomenon[1], which is formed by the complex interplay of light hitting the retina and being transformed into neural signals that are then conveyed to the visual processing regions of the human brain.

3.1 Describing light and its perception

The physical origin of colour perception is electromagnetic energy at visible wavelength from 380nm to 780 nm. Although other factors —such as spatial content, adaptation, colour constancy, etcetera— have an influence on colour perception, the spectral content forms the basis of all research on psychological representation of colour.

The spectral energy distribution that reaches the eye is represented by $E(\lambda)$, it is the power of the light reflected of or transmitted through the material of an object in function of the wavelength λ (expressed in nanometre). $E(\lambda)$ depends on the spectral energy distribution of the light source(s) falling on the object $S(\lambda)$, and the spectral reflectance characteristic of the object $R(\lambda)$. The spectral reflectance characteristic describes how light is reflected from an object, a tomato for example has a reflectance charactistic which reflects long wavelengths and absorbs all other wavelengths. When we suppose that the object reflects the light source equally in all directions (meaning that the object acts as a "diffuser", for example a matte painted wall), we can write

$$E\left(\lambda\right) = S\left(\lambda\right) R\left(\lambda\right)$$

[1] This is the physicists's position on colour, as expressed by Locke and Descartes.

When, for example, $S(\lambda)$ is a uniform light source (a spectrally flat white light) and the reflectance characteristic only reflects light of short wavelengths in the visual spectrum, then $E(\lambda)$ will give the impression of a purple-blue stimulus. The task of the brain is to retrieve the object colour from the stimulus. The human visual system is remarkably efficient at doing this; under a wide number of lighting conditions, the brain is able to retrieve the "original" colour of the object; the ability to cancel out the effect of the light source is called *colour constancy*. The brain not only uses information about the lighting conditions to get to the object colour, but it also uses cues given by object orientation, shading, object form, reflections, object recognition, colour memory and context to retrieve the object colour (Kaiser and Boynton, 1996; Gegenfurtner, 1999). In this thesis I want to steer clear from colour constancy and its intricacies. Colour constancy introduces complexities that are immaterial for the ideas presented in this thesis. This work is concerned with the categorisation of chromatic stimuli, the labelling of these categories and what influence linguistic interactions can have on the coherence of the labelling of the categories and on the categories themselves.

A spectral power distribution (SPD for short) is a function which plots a radiometric quantity against wavelength in the visual spectrum. Sometimes a spectral power distribution is expressed as spectral irradiance in W/m^2nm or W/m^3, but often the SPD is normalised to aid the comparison of colour properties. Sometimes the spectra are normalised so that the value at 560 nm is 100, making the spectra are unitless. Illuminants (i.e. light sources) are often represented this way. The spectral power distribution of the colour samples in this thesis are normalised according to the maximal value of the measuring equipment with which they were recorded, and are limited between 0.0 and 1.0. Figure 3.1 on the next page shows the relative spectral power distribution of CIE illuminant D65 (normalised to be 100 at 650 nm), which resembles average northern daylight, and a highly saturated red (normalised relative to a white test sample).

The visual stimulus $E(\lambda)$ is decoded by passing three different types of chromatic photoreceptors, these are called the cone receptor (see section 2.1). Each cone has a different spectral sensitivity, their peak responsiveness corresponding approximately to red, green and blue light. The cone types are respectively called the *long*, *medium* and *short* wavelength sensitive cones, or the L, M and S-cones. When we define the spectral sensitivity of the three cones type as $l(\lambda)$, $m(\lambda)$ and $s(\lambda)$, and convolute them with the spectral energy distribution[2] $E(\lambda)$ of the stimulus, we obtain tristimulus values as in (3.1 on the following page).

[2]The spectral power distribution and spectral energy distribution, when expressed in relative units, are equivalent when the light source is constant over time; as power is energy over time.

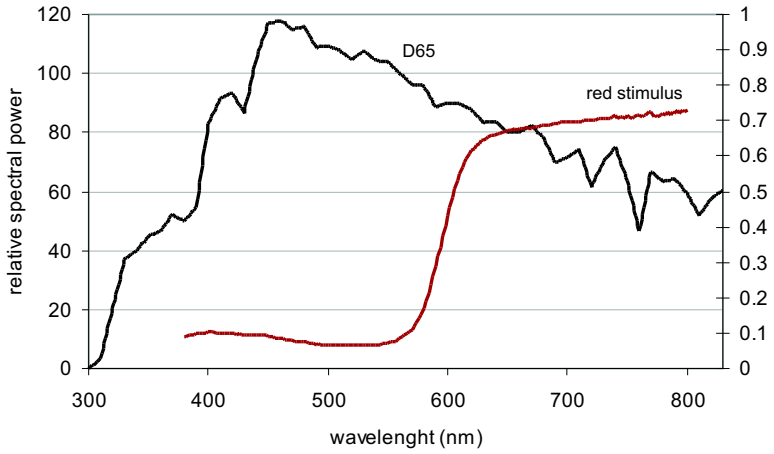


Figure 3.1: Relative spectral power distribution of CIE illuminant D65 and a saturated red colour (the Munsell colour 5R 5/14).

$$\begin{aligned} L &= \int E(\lambda)\ l(\lambda)\, d\lambda \\ M &= \int E(\lambda)\ m(\lambda)\, d\lambda \\ S &= \int E(\lambda)\ s(\lambda)\, d\lambda \end{aligned} \tag{3.1}$$

The mapping from a spectral power distribution to three values implies a large reduction of information; different distribution will lead to the same three excitations of the cones. Two stimuli that have a different distribution, but result in the same tristimulus values are called *metameric*.

Humans are thus a trichromatic species, this implies that any colour sensation can be reproduced by mixing a certain amount R of red light $\mathcal{R}$, an amount G of green light $\mathcal{G}$ and an amount of blue light $\mathcal{B}$[3]

$$C \equiv R(\mathcal{R}) + G(\mathcal{G}) + B(\mathcal{B})$$

These amounts are called the tristimulus values of a colour. Different

[3]This is Grassman's law of additive colour mixture. It is applied in all additive colour devices, such as television sets, LED billboard displays and computer monitors. A coloured image element is made by exciting three coloured elements (phosphors in the case of CRTs) on the screen. One phosphor is red, one green and one blue; when watching the screen from a distance, the three phosphors blend to one colour.

light primaries will require different tristimulus values to reach the same colour impression.

3.2 CIE tristimulus values

In the beginning of the century the CIE (Commission Internationale de l'Éclairage) laid the basis for modern colorimetry. According to Wyszecki and Stiles (1982) colorimetry is the quantification of the human visual responses on the earliest level of visual perception. Colorimetry forms the basis for numerous mathematical models describing psychological colour experience. The basis for colorimetry is the *colour matching experiment*. In a colour matching experiment an observer tries to match a reference stimulus with a second stimulus composed of three primaries (red, green and blue monochromatic light). The subject tries to make the second stimulus indistinguishable from the reference stimulus by changing the mixture of the three primaries. The amount of red, green and blue light needed for making a match leads to insights on the sensitivities of the cone receptors. The CIE did these colour matching experiments with a number of subjects, and distilled out of this the *standard observer colour matching functions*. The visual field used subtended to 2 degrees, meaning that the two stimuli to be matched were only projected on the foveal part of the subject's retina.

For the colour matching experiments the CIE used three monochromatic primaries at $\mathcal{B} = 435.6$nm, $\mathcal{G} = 546.1$nm and $\mathcal{R} = 700.0$nm. Figure 3.2 on the next page shows the resulting colour matching functions for the CIE 1931 2° standard observer, the graph shows what relative amounts of each primary are needed to match a monochromatic wavelength read from the horizontal axis. For a certain region in the spectrum, a negative amount of $\mathcal{R}$ is needed to have a match. Clearly you can not add a negative amount of light; the negative values mean that subjects needed to add $\mathcal{R}$-light to the reference stimuli to obtain a match, this desaturated the reference stimulus and made a match possible (in technical terms: the gamut of the three primaries is insufficient to additively match the whole spectrum).

The CIE then transformed these colour matching functions to colour matching functions for *virtual* primaries; primaries that are not physically realisable because they are more saturated than monochromatic light. The CIE colour matching functions are denoted as $\bar{x}(\lambda)$, $\bar{y}(\lambda)$ and $\bar{z}(\lambda)$. The transformation from $\bar{b}(\lambda)$, $\bar{g}(\lambda)$, $\bar{r}(\lambda)$ tot $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, $\bar{z}(\lambda)$ comes down to a linear transformation. The CIE chose to have the middle colour matching curve coincide with the photopic luminous efficiency functions $V(\lambda)$, a curve describing how luminosity is experienced in function of wavelength. The virtual primaries were chosen such that the two other colour matching curves had no negative values. Figure 3.3 on page 44 shows the resulting colour matching functions, known as the colour matching functions of CIE

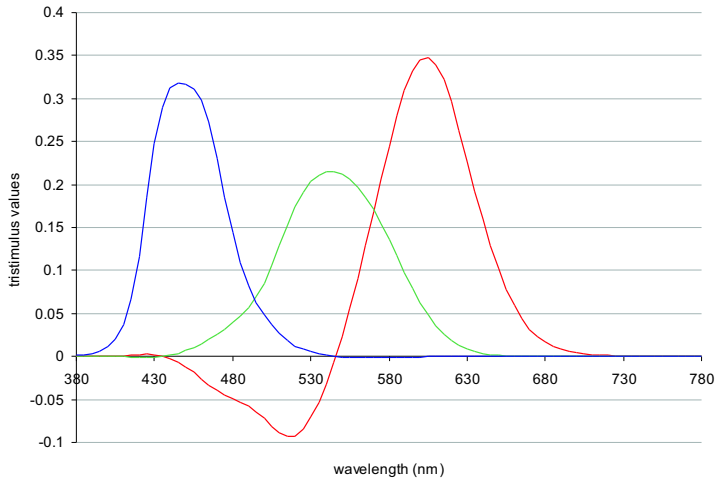


Figure 3.2: Colour matching functions of the CIE 1931 2° standard observer with respect to real primary stimuli at 700.0nm, 546.1nm and 435.8nm. Left to right: $\overline{b}(\lambda)$, $\overline{g}(\lambda)$ and $\overline{r}(\lambda)$. Data adapted from (Wyszecki and Stiles, 1982).

1931 2° standard colorimetric observer. When the spectral power distribution $E(\lambda)$ is known, the XYZ tristimulus values can be calculated using (3.2).

$$\begin{aligned} X &= k \int E(\lambda)\, \bar{x}(\lambda)\, d\lambda \\ Y &= k \int E(\lambda)\, \bar{y}(\lambda)\, d\lambda \\ Z &= k \int E(\lambda)\, \bar{z}(\lambda)\, d\lambda \end{aligned} \tag{3.2}$$

k is normalising constant and depends on whether you use absolute or relative (i.e. unitless) colorimetry. The colour spaces used in this thesis use relative colorimetry with $k = 0.00946300$, which corresponds with CIE illuminant D65. So, if the D65 illuminant is used as stimulus, the Y-value will be exactly 100.0. For other stimuli, this results in XYZ values between 0 and approximately 100. The 1931 standard observer is based in stimuli comprising only 2° of the visual field, meaning that the colours that are to be matched are only perceived using the fovea of the retina. In the 1950s experiments were done using 10° of the visual field, in an attempt to minimize the influence of macular absorption; this resulted in the CIE 1964 10° standard observer data. The experiments in this work use the 1931 2° standard observer colour matching functions.

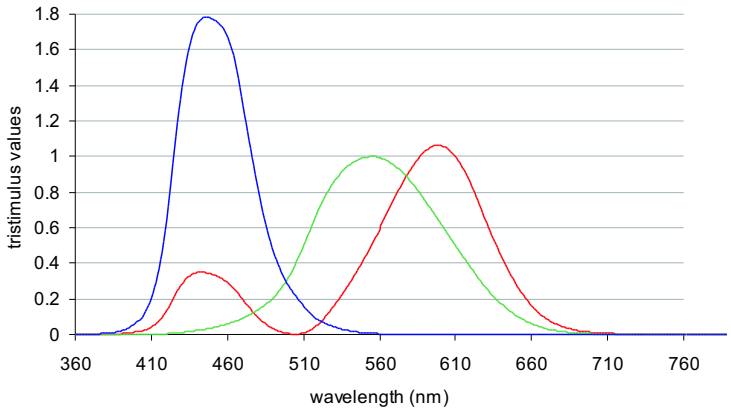


Figure 3.3: Spectral tristimulus values of the 1931 2° standard observer. Left to right: $\overline{z}(\lambda)$, $\overline{y}(\lambda)$ and $\overline{x}(\lambda)$.

Often the XYZ tristimulus values are normalised as in (3.3), x, y and z are called the chromaticity coordinates. The chromaticity coordinates do not contain any information on the intensity of the stimulus and $x+y+z = 1$.

$$\begin{aligned} x &= \frac{X}{X+Y+Z} \\ y &= \frac{Y}{X+Y+Z} \\ z &= \frac{Z}{X+Y+Z} \end{aligned} \tag{3.3}$$

Often only x and y are given, because z follows from

$$z = 1 - x - y$$

If only x and y are given, one has only information on the "hue" of the stimulus. For this reason the Y value, which represents luminance information, is mentioned together with the x and y chromaticity coordinates. The x and y values are often used to represent colour in a two-dimensional plot. Figure 3.4 on the next page shows such a chromaticity diagram for the CIE 1931 2° standard observer.

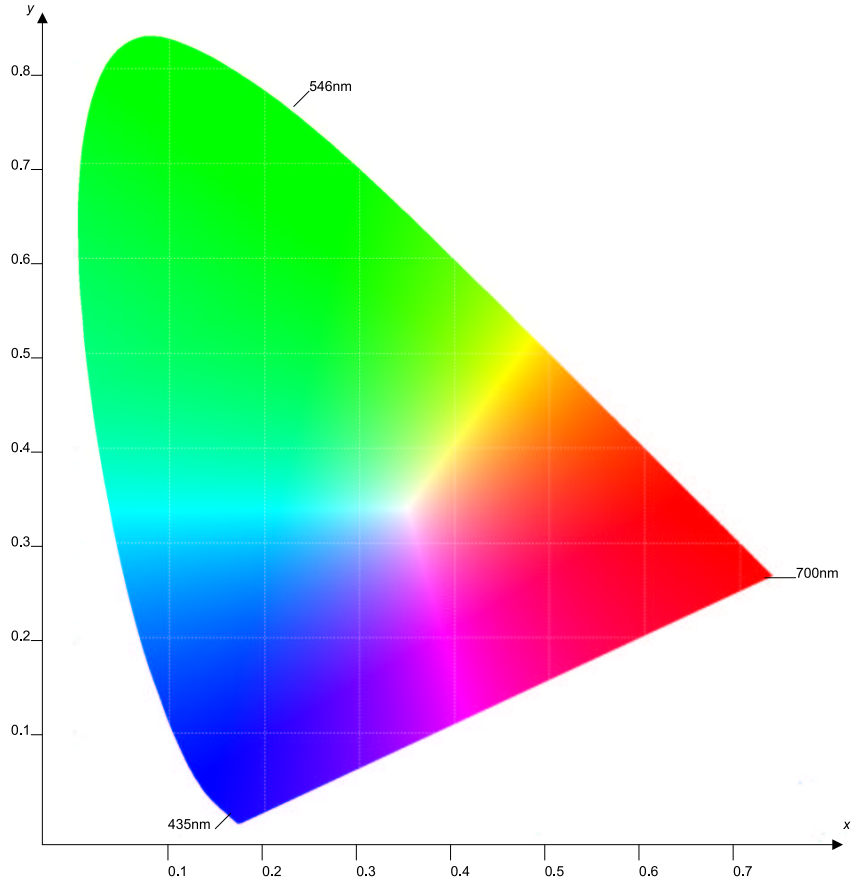


Figure 3.4: Chromaticity diagram for the CIE 1931 2° standard observer. On the upper rim of the chart lie all monochromatic colours (the "spectral" colours); the position of the CIE red (700nm), green (546nm) and blue (436nm) is marked. The straight bottom edge consists of extra-spectral purples. The rim delimits all physically realisable colours. Note that this is only a rendition; the paper or the screen on which you are viewing this chart has a gamut too limited to show all physically realisable colours.

3.3 CIE colour spaces

The CIE XYZ space is the "mother of all colour spaces", but it does not lend itself for practical use. CIE XYZ tristimulus values and Yxy chromaticity coordinates are perceptually non-linear spaces. Comparing two colour stimuli by computing a distance between the two tristimulus values is rather hard. As can be seen in Figure 3.4, green stimuli cover a large area of the chromaticity diagram, while for example yellow only covers a small part of the diagram. Due to this non-uniform mapping of colour stimuli, one cannot use the same distance measure for the whole chart: the distance measure will depend on which two stimuli you are computing the distance between[4]. Instead of using complex distance measures, the CIE in 1976 proposed two new colour spaces that try to get round this perceptual non-linearity by transforming the CIE XYZ space: the CIE $L^*a^*b^*$ space and the CIE $L^*u^*v^*$ space. As these names are bothersome to spell, they are often written as CIE LAB and CIE LUV. Here, only the CIE LAB space is explained.

3.3.1 The CIE LAB space

The CIE LAB space defines an approximately uniform colour space, meaning that the distance between two stimuli presented in the LAB space can be computed with one and the same distance function, no matter where the two stimuli are located in the LAB space (remember that this is not the case in the CIE XYZ space). The CIE LAB space has three quantities L^*, a^* and b^*, computed as in (3.4). The tristimulus values X_n, Y_n and Z_n are those of a reference white stimulus. Usually, this nominally white stimulus is defined by the light source under which the the colour samples are viewed[5]. Generally the light source is taken to be one of the CIE standard illuminants, such as illuminant D65 (which closely resembles northern sky daylight) or illuminant A (representing a Tungsten light source).

[4] In the 1970s the CIE had as many as 20 different formulas to compute the distance between two stimuli (Fairchild, 1998, p. 219).

[5] The technically correct way is to say that X_n, Y_n and Z_n are the tristimulus values of the white object-colour stimulus given by the spectral radiant power of the illuminant reflected into the observer's eye by the perfect reflecting diffuser (Wyszecki and Stiles, 1982, p.167)

$$
\begin{aligned}
L^* &= \begin{cases} 116\left(\frac{Y}{Y_n}\right)^{1/3} - 16 & \frac{Y}{Y_n} > 0.008856 \\ 903.3\left(\frac{Y}{Y_n}\right) & \frac{Y}{Y_n} \leq 0.008856 \end{cases} \\
a^* &= 500\left(\mathrm{f}\left(\frac{X}{X_n}\right) - \mathrm{f}\left(\frac{Y}{Y_n}\right)\right) \\
b^* &= 200\left(\mathrm{f}\left(\frac{X}{X_n}\right) - \mathrm{f}\left(\frac{Z}{Z_n}\right)\right) \\
\mathrm{f}(x) &= \begin{cases} x^{1/3} & x > 0.008856 \\ 7.787x + {}^{16}/_{116} & x \leq 0.008856 \end{cases}
\end{aligned}
\tag{3.4}
$$

The total colour difference ΔE^*_{ab} between two colour stimuli, each given in L^*, a^*, b^*-values, is calculated as in (3.5).

$$
\Delta E^*_{ab} = \sqrt{(\Delta L^*)^2 + (\Delta a^*)^2 + (\Delta b^*)^2} \tag{3.5}
$$

The X/X_n, Y/Y_n and Z/Z_n terms in (3.4) are a modified form of the von Kries (1902) chromatic-adaptation transform, providing a unsophisticated solution to colour constancy in the model. The L^* dimension correlates with perceived lightness and ranges from 0.0 for black to 100.0 for white stimuli. The a^* and b^* dimensions correspond approximately with red-green and yellow-blue chroma perceptions. For white, black and grey stimuli, both a^* and b^* are 0.0. Their range is only limited by the chromatic material being described, but for practical purposes it ranges between -60 and 60. Figure 3.5 shows the CIE LAB space in Cartesian projection, a mode of presentation often used for CIE LAB.

The CIE LAB colour space was designed to be perceptually uniform, and performs well on physical colour samples such as the colour samples of the Munsell colour book (Fairchild, 1998, p.221). For systems which can achieve higher chroma, such as CRT devices, the CIE LAB shows some deviation. Recently it has been superceded by other colour appearance models performing better at predicting high chromatic colour differences. However, as CIE LAB is a well established *de facto* standard and suffices for our purposes, namely investigating the influences on the formation of colour categories on natural colours, we will be using CIE LAB throughout this work to compute colour differences.

3.3.2 Other colour spaces

The CIE has designed other colour spaces as well. At the same time of CIE LAB, a second colour space, named CIE LUV, was introduced. The CIE

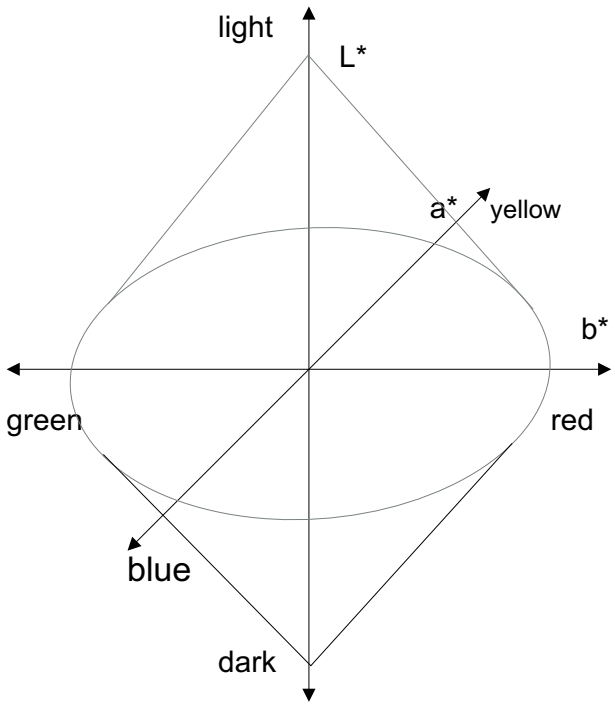


Figure 3.5: The CIE LAB space in a Cartesian projection.

LUV, also known as CIE $L^*u^*v^*$ 1976, is also a uniform colour space. It also has tristimulus values as input and three output predicting perceived lightness, hue and chroma. It applies a different chromatic-adaptation transformation than CIE LAB, instead of dividing the tristimulus values by a normalising white stimulus it uses subtractive normalisation (Wyszecki and Stiles, 1982). The resulting transformation is less physiologically correct than in CIE LAB, furthermore the CIE LUV space is extremely inaccurate with respect to predicting visual data (Fairchild, 1998, p. 230).

One other model is worth mentioning. In 1996 the CIE started an effort to design a new colour appearance model, called CIE CAM 1997. It is intended to contain all present knowledge on colour appearance, but as it is aimed at practical applications and industry it is also relatively uncompex (but still much more complex than the CIE LAB model) (Fairchild, 1998).

3.4 Displaying the colour stimuli

For displaying the colour stimuli on a screen or a printer a conversion is needed from the spectral power distributions to a RGB representation. The RGB colour space (where RGB stands for *red*, *green* and *blue*) is a technical colour representation where a colour is constructed by adding quantities of red, green and blue primaries. The RGB colour space finds its origin in cathode ray tube display devices, where a colour sensation is achieved by additive mixing of luminous phosphors. The phosphors each emit a different wavelength corresponding to red, green and blue (which approximates the maximal response of the human cone receptors). The RGB representation of a colour is a triplet; for example $\{0, 0, 0\}$ is black, $\{1, 1, 1\}$ is white and $\{1, 0, 0\}$ is the purest red the device can produce. RGB is merely a technical colour space and does not lend itself to be used as a colour appearance model. The RGB space is not equidistant, meaning that no simple distance metric can be used to quantify the psychological distance between two RGB triplets. Users also find it very counter-intuitive to work with colour using RGB triplets; that is the reason why other technical colour spaces have been devised such as HSI and HSV, which split the colour representation up into hue, saturation and intensity allowing for more intuitive handling by computer users.

To get from CIE XYZ values to RGB triplets we first normalise the XYZ values computed with (3.2), by the XYZ values for a maximum radiance flat spectrum (which corresponds to "whitest white" of all possible stimuli).

$$\begin{aligned} X' &= X/X_n \\ Y' &= Y/Y_n \\ Z' &= Z/Z_n \end{aligned} \tag{3.6}$$

This white point $X_nY_nZ_n$ is the CIE Illuminant D65, using the CIE 1931

colour matching functions, its XYZ values are $\{95.04682, 100.0, 108.88300\}$. To get the RGB coordinates a linear transform is needed which can be written in matrix notation.

$$\begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = M \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \tag{3.7}$$

The matrix M depends on the colour system of the output device we are using, a compromising set of values which performs quite well on a variety of typical colour devices is

$$M = \begin{bmatrix} 2.9459 & -1.3974 & -0.4532 \\ -0.9693 & 1.8760 & 0.0416 \\ 0.0556 & -0.2040 & 1.0574 \end{bmatrix} \tag{3.8}$$

The R', G' and B' values might be negative or greater than one, meaning that they lie outside the gamut of this particular transformation. We solve this by clipping R', G' and B' between 0 and 1. After this we apply a gamma correction of $\gamma = 1.7$, which is actually a nonlinear transformation stretching the RGB values to give better results on non-linear colour display devices.

$$\begin{aligned} R &= R'^{\frac{1}{\gamma}} \\ G &= G'^{\frac{1}{\gamma}} \\ B &= B'^{\frac{1}{\gamma}} \end{aligned} \tag{3.9}$$

3.5 Summary

This chapter provides a brief introduction to the colour representation. It explains how a light stimulus is converted in the human cone receptors to a tristimulus value, and how the CIE models this conversion by measuring the cone sensitivities indirectly through colour matching experiments. Finally, the CIE LAB colour appearance model is explained; as it will be used in the experiments described in the next chapter. For more information on this topic the reader is referred to (Wyszecki and Stiles, 1982; Kaiser and Boynton, 1996; Fairchild, 1998).

Chapter 4

Modelling perception, categorisation and lexicalisation

This chapter describes the models and representations as used in all simulations. To understand the functioning of the simulation and the results reported in later chapters, the reader is encouraged to familiarize himself with the structure of the representations.

The simulations try to mimic real-world behaviour using computational models. A simulation and the models upon which it is based can however never attain the full complexity of the real world. While building models it is (often) impossible to faithfully reconstruct real-world phenomena and their interactions. A good start to constructing a model is to split up the system that is to be modelled into components. Components are easier to handle, the implementation of components is clearer, and it improves the understanding of the overall functionality of the system. Of course, defining components requires the researcher to introduce divisions; divisions that might seem artificial and might sacrifice certain interactions between components that only arise when they act as one intricately intertwined system. Of some parts —for example the categorisation— the functional behaviour is well known, but the exact inner workings remain hidden in the neuronal interactions of our brain. One can only try to faithfully reconstruct the functionality and should choose a model that imitates the behaviour up to needs of the experimenter. This demands well-considered choices and compromises; often the researcher has to let go of some realism to achieve a model that is easy enough to understand, complex enough to faithfully mimic the desired functionally, and simple enough to allow for speedy simulations (Steels, 2001a).

The basic entities of the simulations are the *agents*. An agent is the equivalent of an individual in the simulation. Each agent has components

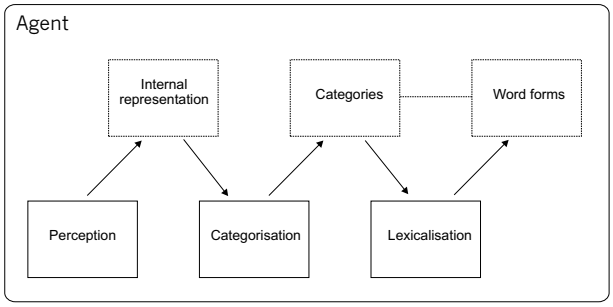


Figure 4.1: The conceptual structure of an agent.

to model perception, representation of perception, categorisation and lexicalisation. On top of that, the agent is able to interact with the world and with other agents. Figure 4.1 shows how an agent is organised internally; the sequential organisation of perception, internal representation, categorisation and lexicalisation does not tell anything about the order of processing; it rather illustrates the interdependence of each process. The internal representation uses the perception, the categorisation happens on the internal representation and results in categories which are then lexicalised.

The agents start off as *tabula rasae*: at the start of the simulation they have no representations, no categories and their lexicons are empty. Their perception is translated into a representation, this representation is subsequently used to define categories on. When the agents participate in interactions they need to lexicalise these categories in order to be able to communicate their meaning to the other agents. These lexicalisations are random for the individual agent, but nonverbal feedback on their effectiveness in communicating meaning is used to keep a "score" for the form-meaning association, which influences the form's future use. The selectionist pressure of the interactions decides on whether certain form-meaning pairs are to be used in the future: word forms that are not successful are likely to be less used and can eventually be forgotten. On the other hand, word forms that successfully convey meaning to other agents are strengthened and the form-meaning associations are apt to arise as stable concepts throughout the entire population. Additionally, agents are able to pick up new word forms and can associate them with new or existing categories. All this is fully explained in the next chapter.

The simulation contains two levels. First there is the individual level, at which a single agent is described; this is the topic of the current chapter. And next there is the population level, where the interactions between the agents are defined. The population dynamics are described in the next two chapters.

At the individual level we have to describe the internal components that make up an agent. Figure 4.1 shows how the functionality of an agent can be seen as four distinct parts: perception, categorization and lexicalisation. These functional parts depend on each other in a sequential manner: first there is perception, then there is the representation of the perception, followed by the categorization of the representation space. As a last step, the result of the categorization can be lexicalised. This chapter describes how these functional parts are implemented, and what choices have been made in doing so. While explaining how individual agents function, we also set the stage for explaining the next stage, which is the population level where agents interact with each other.

Section 4.1 explains the the conversion of colour stimuli to private representations. Section 4.2 explains the representation of categories, it also explains how category representation can be adapted. Finally 4.3 describes how the association between meaning and form is implemented.

4.1 The perception and its representation

Agents should be able to perceive their environment. The perception is needed to be able to interact with the world; and in our case to build a basic symbolic representation of the world, on which further cognitive processes work. It would be a daunting, if not impossible, task to fully model human perception (or the perception of any other animal). One has to make two important choices. First, what parts of perception will be modelled since modelling the whole is infeasible. Second, how will those parts be modelled, taking into account the simulation that one wishes to build and the phenomena that one tries to study.

For most perceptual capacities neuroscience has only uncovered the top of the gargantuan neurological iceberg involved with the senses and with their connection to other brain regions. The transduction process of the senses has been extensively studied; for an overview see (Arbib, 1995; Levine, 2000). Quite a lot is known about receptor cells, which are often located at easily accessible places on the body. But as we start following afferent nerve cells towards the brain and eventually towards the cerebral cortex, our understanding becomes unclear. In spite of in-vivo experiments or fMRI studies, nobody has been able to piece together an all-explaining theory of perception; even a neurological model for a single faculty currently still lies beyond science's reach. Fortunately enough is known about colour and colour perception to allow us to simulate the phenomena of categorization and lexicalisation and the dynamics needed for communicating meaning between agents.

The goal of the perception is to map multi-dimensional sensory stimulations to a lower-dimensional representation on which categories are de-

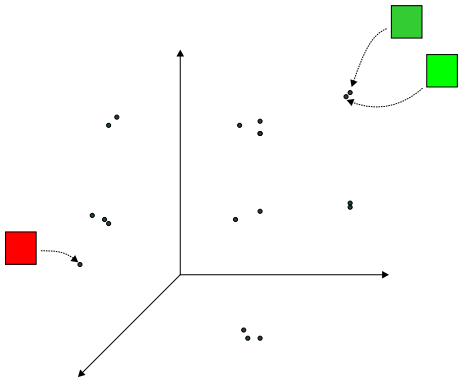


Figure 4.2: An illustration of a three-dimensional representation space. The axes correspond to representational dimensions (for example the three colour dimensions). Each point R_i is a mapping of a sensory sensation into the representation space. Points being close together correspond to similar sensations in the real world, while points being far apart represent dissimilar sensations.

fined. We need to find a transformation $f(S, R) : S \rightarrow R$ from a sensory stimulus $S \in \mathbf{S}$ to a representation $R \in \mathbf{R}$. S could be any sensory stimulus, such as a haptic sensation or an image projected onto a visual sensor. $\mathbf{S}$ then contains the set of all possible sensory impressions. R is a representation of a sensory impression and $\mathbf{R}$ contains all possible representations, thus forming the *representation space*. f is a psychophysical mapping from the physical space onto a psychological space (Shepard, 1987).

The representation should fulfil two requirements. First, it should make *discrimination* possible. Two sensory stimuli are discriminable if and only if they map onto two different points in representation space. Second, one should be able to define a *similarity measure* over the representation space. Sometimes uncomplicated solutions can be found for this, for example using the distance between two points in the representation space. In the set-up described here, a representation R is a set of N discrete or continuous values r_i, so that $R = \{r_1, ..., r_N\}$. It is this psychological space that will be used for further processing. Figure 4.2 shows an illustration of a possible space.

4.1.1 Representing colour perception

For the experiments in which colour samples are offered to the agents, a mapping is needed from a colour percept to an internal colour representation. This representation is a point in a certain colour space.
There exist numerous colour spaces, each mapping spectral power distributions to a lower-dimensional numerical space. Each of the colour spaces

has been constructed to aid some purpose, technical or not. Chromaticity diagrams for example, have been devised to allow a two-dimensional representation of colours. Other examples are the CIE LAB and CIE LUV-spaces, which are three-dimensional spaces correlating with perceived lightness, chroma and hue of a stimulus and which were constructed to be perceptually uniform, allowing for a straightforward similarity measure between two colours (see chapter 3).
In our experiments we stick to the CIE LAB space. It is a relatively simple space, requiring straightforward computation and producing good results. Lammens (1994) constructed a colour categorisation and naming mechanism, and compared the performance of three different colour spaces: CIE XYZ, CIE LAB and NPP[1] space. The CIE LAB space turned out to be the best space for representing colour categories in.

> The model works best on the $L^*a^*b^*$ space, which is meant to be a perceptually equidistant space (Lammens, 1994, p. 140).

The CIE LAB space, CIE 1976 $L^*a^*b^*$ for full, is explained in section 3.3.1. The equations are repeated here for convenience. In the equations X, Y and Z are the tristimulus values of the colour sample, and X_n, Y_n and Z_n are the tristimulus values of a reference white (the whitest stimulus in the environment). L^* represents lightness, a^* corresponds approximately to redness-greenness, b^* to yellowness-blueness.

$$\begin{aligned} L^* &= \begin{cases} 116\left(\frac{Y}{Y_n}\right) - 16 & \frac{Y}{Y_n} > 0.008856 \\ 903.3\left(\frac{Y}{Y_n}\right) & \frac{Y}{Y_n} \leq 0.008856 \end{cases} \\ a^* &= 500\left(\mathrm{f}\left(\frac{X}{X_n}\right) - \mathrm{f}\left(\frac{Y}{Y_n}\right)\right) \\ b^* &= 200\left(\mathrm{f}\left(\frac{X}{X_n}\right) - \mathrm{f}\left(\frac{Z}{Z_n}\right)\right) \\ \mathrm{f}(x) &= \begin{cases} x^{1/3} & x > 0.008856 \\ 7.787x + {}^{16}/_{116} & x \leq 0.008856 \end{cases} \end{aligned} \tag{4.1}$$

4.2 The categorisation

When an agent is to act upon the world, a certain form of processing is needed between perception and action. This processing can be very simple; behavioural robotics describes how direct coupling between sensing and acting can amount to complex behaviour (Braitenberg, 1984; Brooks, 1991). However, planning in the world and communicating about it require a symbolic representation of the world. This symbolic representation

[1] NPP space, or neuro-psycho-physical colour space, is a colour space constructed by Lammens and based on neurophysiological data by De Valois et al. (1966).

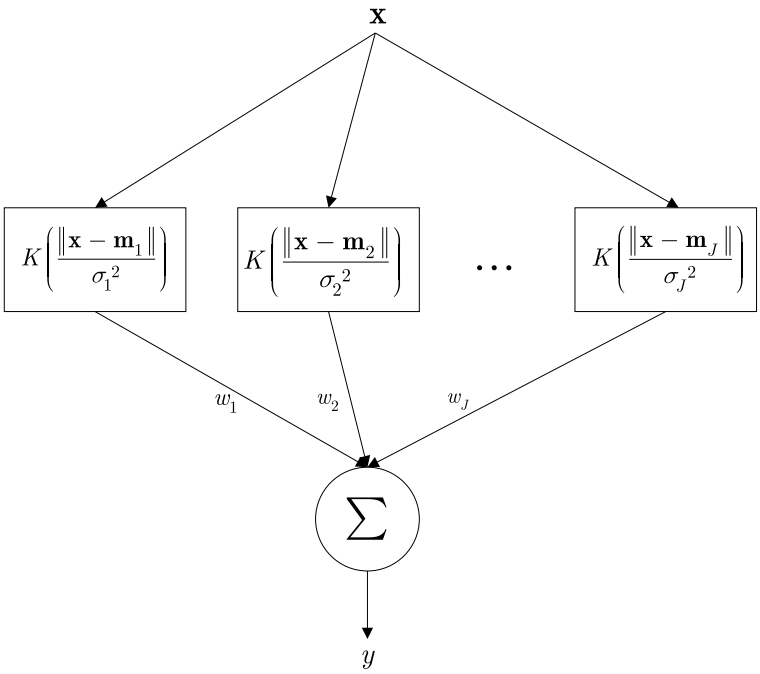


Figure 4.3: The adaptive network, it consists of one hidden layer of locally tuned units fully connected to a linear output unit. One adaptive network is used to represent one category.

is formed by structuring and cutting up real world sensations in space and time. In the simulations described here, categories are defined in the representation space (sometimes also called *feature space*). Categories delineate, albeit sometimes fuzzily, what features an object should have and in what quantity they should be present in order to belong to that category. Categories can be described using language, for example the category "big" could be described as "any object being larger than the average size of similar objects". However, in simulations a numerical definition is preferred, as this facilitates the computing of a membership function. If an object is perceived, it should be possible to decide to which degree a category is applicable to the object. It should also be able to learn categories and to adapt categories to reflect new experiences.

4.2.1 Representing categories with an adaptive network

In this work a category is represented by an *adaptive network*. The adaptive network used for representing categories is very much based on *radial basis function* (RBF) networks. Radial basis functions have locally tuned re-

sponse characteristics; such specific responses can be found in abundance in biological nervous systems. They exist in the form of nerve cells that are responsive only to a restricted range of input stimuli. The first work on radial basis functions dates from the early sixties (Medgassy, 1961). RBF networks in their current form differ from RBFNs from the 1960's in that they can be trained; see (Broomhead and Lowe, 1988; Lee and Kil, 1988; Moody and Darken, 1989), for an introduction see (Hassoun, 1995; Ghosh and Nag, 2000). The *adaptive network*[2] used in this work has a feed forward structure consisting of one hidden layer of J locally tuned units and one output layer, as shown in figure 4.3. The locally tuned units all receive the same n-dimensional vector $\mathbf{x}$ as input. The output y of the network is the sum of the outputs of the locally tuned units. The locally tuned units compute a measure for the "closeness" of the input vector $\mathbf{x}$ to an n-dimensional vector $\mathbf{m}_i$ associated with the i-th unit. For this, each locally tuned unit computes a function (4.2).

$$z_j\left(\mathbf{x}\right) = K\left(\frac{\left\|\mathbf{x}-\mathbf{m}_j\right\|}{\sigma_j^2}\right) \tag{4.2}$$

Where K is a strictly positive radially symmetric function, the *kernel* function. K is maximal at the centre $\mathbf{m}_j$, but drops towards zero when further from the centre. $\|\cdot\|$ is a distance metric. And the parameter σ_j determines the width of the locally tuned unit. We need a strictly positive function for $z_j\left(\mathbf{x}\right)$, which acts as a receptor that reacts maximally at $\mathbf{m}_j$ and decreases monotonically with the distance to $\mathbf{m}_j$. A Gaussian function (4.3) fulfils these requirements.

$$G\left(x\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{4.3}$$

A Gaussian has the property that $\int G(x)\,dx = 1$, and not only the width of the Gaussian but also the maximum depends on σ. For our purposes the factor $1/\sigma\sqrt{2\pi}$ is not needed (we do not use the function as a normal probability distribution) and we want the receptor function to have a maximal reaction of 1.0 at $\mathbf{m}_j$, independent of σ. Therefore we rewrite the Gaussian function as (4.4)[3].

$$G'\left(x\right) = e^{-\frac{1}{2}\left(\frac{\sqrt{\sum_{i=1}^{N}\left(x_i-\mathbf{m}_{ji}\right)^2}}{\sigma}\right)^2} \tag{4.4}$$

[2] I prefer to use the wording *adaptive network* instead of radial basis function networks, to denote the difference between RBFNs, which are trained to fit a function using a learning method, and adaptive networks, which are adapted.

[3] Courtesy of (Lammens, 1994)

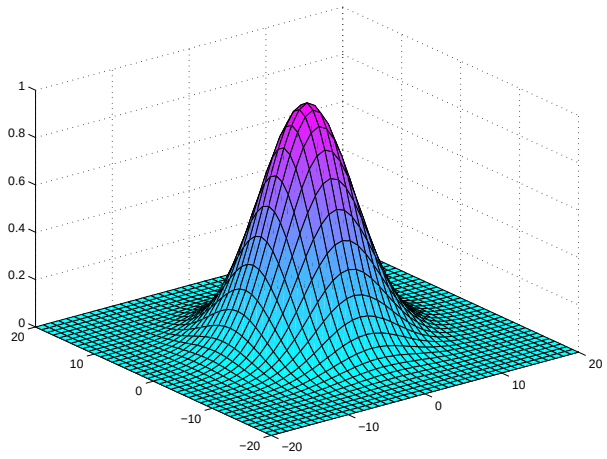


Figure 4.4: An illustrative plot of the output of $z_j(\mathbf{x})$ for a two-dimensional space, with $\mathbf{m} = (0, 0)$ and $\sigma_1 = \sigma_2 = 5$, the function has a maximum of 1 at $\mathbf{m}$.

When the "width" σ of the Gaussian is not equal for all dimensions, we can use a different σ_i for every dimensions i, as in (4.5).

$$G'(x) = e^{-\frac{1}{2}\sum_{i=1}^{N}\frac{\left(x_i - \mathbf{m}_{ji}\right)^2}{\sigma_i^2}} \tag{4.5}$$

This forms the basis for the reaction function of a locally tuned unit. The output of the j-th locally tuned unit $z_j(\mathbf{x})$ is computed as in (4.6). The locally tuned unit has a centre $\mathbf{m}_j$ and a receptor width σ_j. Figure 4.4 shows an illustration of the output of a locally tuned unit.

$$z_j(\mathbf{x}) = e^{-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i - \mathbf{m}_{ji}}{\sigma_i}\right)^2} \tag{4.6}$$

The output of the adaptive network is computed as the weighted sum of the outputs of the locally tuned units (4.7). Figure 4.5 gives an example of how the output of one adaptive network could look like.

$$y(\mathbf{x}) = \sum_{j=1}^{J} w_j z_j(\mathbf{x}) \tag{4.7}$$

Adaptive networks can be used to make a mapping $f : R^N \rightarrow R^M$, a continuous real-valued mapping from an N-dimensional space to an M-dimensional space ($M \leq N$). In our case $M = 1$ and the adaptive network is used to solve a classification task. Each network represents exactly one class. When an input is fed to the network, the output is the "degree of membership" of the input to that class.

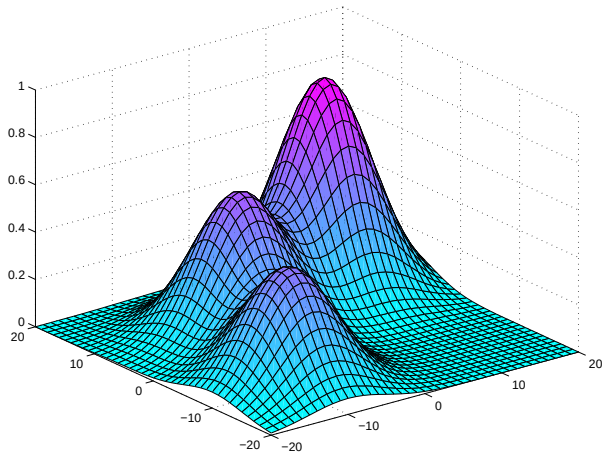


Figure 4.5: Illustration of the output of an adaptive network in a two dimensional space. The network has three locally tuned units with centres at $\mathbf{m}_1 = (-10, -10)$, $\mathbf{m}_2 = (10, 10)$ and $\mathbf{m}_3 = (-5, 5)$, with $\sigma_i = 5$ and with respective weights of 0.5, 0.9 and 0.6.

Of an adaptive network the following parameters can be modified to tune the reaction of the network:

- The number of locally tuned units J.
- The location $\mathbf{m}_j$ of the maxima of the locally tuned units.
- The width σ_j of the locally tuned units.
- The weights w_j of the locally tuned units.

Though several training methods exist, the networks are not trained according to some supervised, reinforcement or unsupervised learning scheme; they are adapted according to their performance during discrimination games and during interactions in the language games.
In brief, the adaptive network reacts to an input vector and returns a membership measure. When the input evokes a high weighted reaction from one or more of the locally tuned units, the network output will be high. When the input vector is unable to elicit a strong reaction from the units, the network output will be close to zero. In this way an adaptive network can represent a category[4]: the locally tuned units in the network represent regions, or instances, in the n-dimensional space that belong to the same category. When an unknown vector is presented, the network returns a value describing how much the vector fits the category.

[4]Machine learning adepts might think of a category as a class.

4.2.2 Adapting the adaptive network

As mentioned before, the response of the adaptive network can be altered by changing one or more of the following parameters: the number of locally tuned units, the location and width of each unit, and the weight of each unit. In the experiments described here, only the number of locally tuned units and their weights are modified; the centre and width of the locally tuned units remain unchanged. These are the following cases in which a locally tuned unit is adapted.

Adding a locally tuned unit. Adding a locally tuned unit can occur at two times: when initialising the network and when adjusting the network. An adaptive network is used to represent a category; when it is initially created, a first locally tuned unit is added to it. When, at a later stage, the network needs to be modified to better represent the category it represents, adding a new locally tuned unit might do this.

Removing a locally tuned unit. When the weight of the locally tuned unit is lower than a certain threshold, the locally tuned unit is removed from the adaptive network. When no more locally tuned units are associated with the network, the network becomes invalid for representing a category; and is subsequently removed.

Increasing the weight of a locally tuned unit. This happens when the locally tuned unit contributed to a successful classification. However, the weight is clipped at a maximum value of 1.

Decreasing the weight of a locally tuned unit. As the simulation progresses the weights of all locally tuned units of all adaptive networks are decreased by a small number. This weight decay takes care of "forgetting": locally tuned units that do not contribute to the classification will eventually end up having a zero weight, making them eligible for removal.

These four ways of changing the adaptive network take care of moulding it to the category it represents. The receptor width of the locally tuned units, represented by σ_j, is set to a default value and is not changed. The reader is referred to section 5.1 on page 68, which describes the dynamics of the simulations, for more information on how the adaptive networks are adapted.

Observe that the ways in which the adaptive networks are changed is deliberately sampling dependent. If stimuli belonging to one category are often presented this category will become stronger, meaning that the recurring presentation of a stimulus will elicit a higher reaction from the adaptive network sensitive to that particular stimulus. On the other hand, when a category is not stimulated weight decay will make its reaction less

prominent. In summary, the nature of the categories is dependent on the environment[5].

4.2.3 Why adaptive networks?

The implementation of a category is twofold; (1) we need a *representation* of the categories and (2) we need a *matching operation* on the representation. Several alternatives exist for representing categories; actually almost any classification algorithm would more or less qualify. Category representations come in many forms, some examples are lists of features, structural descriptions or templates. Two often used representation in language games are discrimination trees (Steels, 1996a; Vogt, 2000; Steels and Kaplan, 1999a) or adaptive subspaces (de Jong and Vogt, 1998; de Jong, 1999); in which each input dimension is divided into a number of ranges, a category is then represented by a conjunctive set of these ranges. Discrimination trees or adaptive subspaces are however not an appropriate choice for our purpose, as they divide a continuous space into discrete regions. Colour categories, and most –if not all natural– categories have fuzzy boundaries (Rosch, 1978). Another alternative could have been multilayer perceptrons[6], but neural networks require off-line training, which makes them less suited for online adaptation. Yet another alternative would be a k-nearest neighbour approach (Mitchell, 1997) or a point representation of which the membership function would be computed by some distance function.

The adaptive networks presented here belong to the instance-based learning class of learning algorithms, for an introduction see (Mitchell, 1997, p. 230-248). It has several advantages over other alternatives, summed up in the following list.

- Adaptive networks have a multi-point representation. When using a single point representation of a category in feature space, one supposes that the category can be represented with a radially symmetrical kernel, this does not allow for subtleties such as asymmetrical or non-convex category shapes.
- Adaptive networks are an implementation of *prototype theory* as proposed by Rosch (Rosch et al., 1976; Rosch, 1978). In this theory, categories are defined in terms of prototypes. A prototype contains the

[5] An alternative, in which categories are not dependent on the perceived world, is described in chapter 6. There the nature of the categories depends on Darwinian evolution.

[6] Laakso and Cottrell (2000) trained neural networks to map colour stimuli to colour names. Their experiments are however only used for demonstrating how different external and internal network representation can arrive at the same categorisation, and are not trained for faithful colour categorisation.

attributes most representative of items inside and least representative of items outside the category. It is not restricted to one locus or one category member, instead a prototype is a collection of attributes and category members with high "prototypicality". Prototypical categories are categories[7] that are used often, named quickly and recognised by all members of a culture. DOG and CAT are examples of prototypical categories; CHIHUAHUA and SYLVESTER are not.

- Adaptive networks are easy to analyse, as opposed to multilayer perceptrons. One can define a measure to calculate the similarity between two adaptive networks by comparing the locally tuned units. This is not straightforward with artificial neural networks.

- When the basis functions used in the locally tuned units are infinitely narrow, the adaptive network becomes a lookup table: it only reacts to the value associated with one of the training vectors, the network will not respond to other values. When the width of the locally tuned units is increased, the network interpolates between points on which it was trained. As a result is can be made to cover an extensive region of the input space while relying on a finite number of memorized examples (Poggio and Hurlbert, 1994; Edelman, 1999).

- Adaptive networks are speedier than neural networks. Neural networks require retraining every time a new instance needs to be added to a category represented by the neural net.

An adaptive network can also represent categories that are not connected. Which means that a category can have several hidden units, and in between the units of this category other hidden units of other categories could appear. In this way, although not used in this work, adaptive networks are able to represent different levels of abstraction (Rosch, 1978). For example a category might exist for representing the basic level BIRD, but in the membership space of this category other, more specific, "superordinate" categories such as ROBIN or PENGUIN might exist. To facilitate this the membership model, which is now only fit for judging the membership of basic level categories, should be adapted.

[7]Some might ask how the representation with adaptive networks relates to MacLaury's vantage theory (MacLaury, 1992). Categorisation in the vantage theory allows for universal constraints and models the method of how a person employs, constructs, use, changes and recalls categories from memory. The distinguishing property of vantage theory is that categories have infinitely variable gradation of categorical membership values. As vantage theory is not a computational theory but a descriptive theory drawn from MacLaury's experience as an anthropologist, the comparison is not straightforward and will be left for another time.

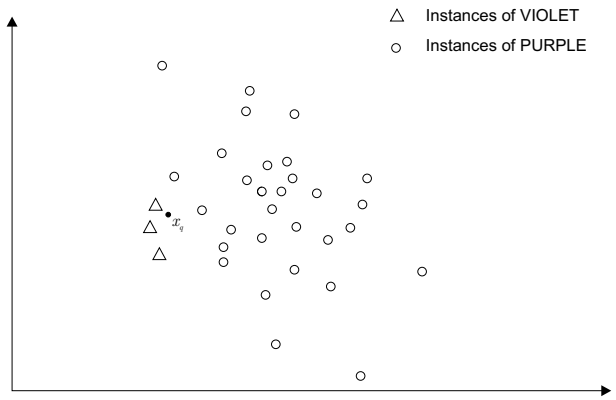


Figure 4.6: Comparing k-nearest neighbour and adaptive networks.

4.2.4 How do adaptive networks compare to k-nearest neighbour?

There are many similarities between the adaptive network approach and the k-nearest neighbour algorithms (KNN for short). Both use exemplars to represent categories (or "classes" in the KNN terminology), both can add weight to exemplars to tweak their influence on the decision process. Both rely on a distance measures to compute the membership of a unknown point. So, where exactly lays the difference? Why choose for adaptive networks, and not for a KNN approach (which is much simpler and more established)?

Let's first recapitulate how k-nearest neighbour works (Mitchell, 1997). There is a training set $< x, f(x) >$ with x being a point in a space $\Re^N$ and $f(x)$ is the class the point belongs to, such as PURPLE or BLUE. Let $V = \{v_1, \ldots, v_s\}$ be the set of all classes. Let x_q be a query instance that needs to be classified. Now $x_1, \ldots, x_k$ denote the k instances from the training set that are nearest to x_q. Now,

$$f\left(x_q\right) \leftarrow \arg\max_{v \in V} \sum_{i=1}^{k} \delta\left(v, f(x_i)\right) \tag{4.8}$$

with $\delta(a, b) = 1$ if $a = b$ and where $\delta(a, b) = 0$ otherwise.

In plain words: the algorithm assigns x_q to the class of which most instances are nearest to x_q. And exactly in this behaviour do we find the difference with adaptive networks. Adaptive networks assign an unknown instance x_q to the class of which the network gives to highest output when x_q is applied as input to the network.

An example is used to illustrate how KNN can produce undesired and

wrong classifications. Consider figure 4.6 where we have two categories, one for PURPLE and one for VIOLET. Now VIOLET is a very specific category, with only few instances; while PURPLE is a category covering a large part of the space, with many instances. Suppose a query instance x_q, which we know belongs to the class VIOLET, is to be classified. k-nearest neighbour will classify it as belonging to PURPLE, while adaptive networks would, correctly, classify it as VIOLET. In general, adaptive networks are able to classify instances if there are classes with few, or only one, instance. While with KNN, classes with many instances quickly dominate the classification procedure, making subtle classification impossible[8].

4.2.5 The significance of the width σ of a locally tuned unit

Every locally tuned unit has a width σ, which defines how broad the selectivity of the unit is. σ therefore seems to be yet another variable in the already variable-burdened adaptive network model. However, due to the functionality of the network σ is not a critical parameter. The only condition for the categorisation to work properly is that the σ values of all locally tuned units of all adaptive networks of an agent are equal and in the interval $]0, +\infty[$.

Consider the example in figure 4.7. There are two categories, each having two locally tuned units. Also, there is an unknown stimulus marked x that has to be classified as either belonging to category A or category B. The locally tuned units have a width σ defining the reach of their sensitivity. If σ would be 0, a locally tuned unit would only react if the unknown stimulus was dead on the centre of that unit; which is not very useful considering that one would like the categories to interpolate between their locally tuned units. That is why σ should be greater than 0. Now the reaction of the category to the unknown stimulus is the summation of the reaction of its locally tuned units, and the category having the highest reaction wins. This behaviour makes the classification independent of the value of σ, as long as σ is high enough to allow interpolation between the units. I have experimented with different values of σ, and for practical and presentational reasons all simulations take as default $\sigma = 5$.

4.3 The lexicalisation

Categories can be associated with one or more labels, called *forms* in the linguistics field. Human colour categories, if labelled, are always associated with a word form; however forms are not restricted to words, they can be any sign (written, visual, auditory, etcetera).

[8] Although it must be granted that weighting the KNN decision process with the inverse of the distance to the instances might provide a solution (Mitchell, 1997, p. 234).

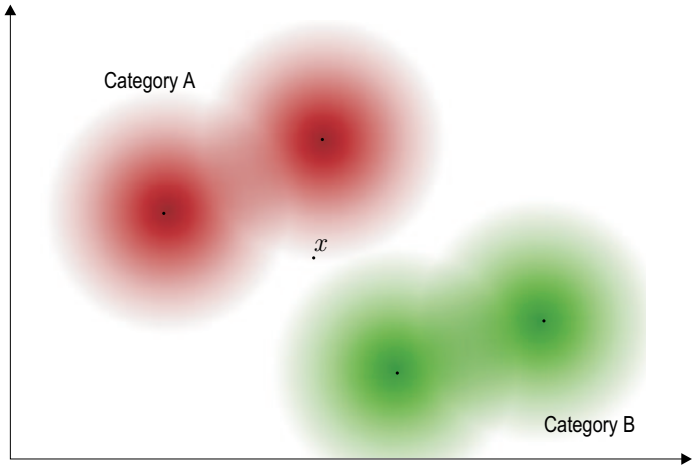


Figure 4.7: Illustration of two categories each containing two locally hidden units. Each locally tuned unit has a radially decaying reactiveness, illustrated by the shaded region around the centre of the unit. The percept x has to be classified.

In this work we will suppose that a category c is associated with a set of forms F, this gives a form-meaning association $< c, F >$. The set F can be empty ($F = \emptyset$), it can contain one or more forms ($F = \{f_1, \ldots\}$). The strength of the association between a word form and its category is represented by a *score*, this a scalar value between 0 and 1. If the score is 1, the strength is maximal. This strength is set under influence of linguistic interactions (see 5.1.3). For the sake of familiarity, the forms connected to the colour categories will be considered to be word forms, constructed from a finite alphabet. This is as far as the analogy goes with human colour terms; it is, for example, not the case that often-used word forms are shorter (Zipf, 1929; Brown and Lenneberg, 1954). The word forms are only used to noiselessly convey labels for categories between agents.

4.4 Populations of agents

The experiments used in this thesis are multi-agent experiments. Multi-agent research is a broad field that has won recognition in artificial intelligence, artificial life, robotics, and other scientific and engineering fields. The strength of multi-agent systems lies in allowing single autonomous functioning entities to work together and to communicate with each other in order to obtain a more complex functional level; relying on the old adagio that the whole is more than the sum of the individuals. For an introduction see (Wooldridge and Jennings, 1995; Ferber, 1998; Jennings and

Wooldridge, 1998). This paradigm fits very well with the idea of language and shared meaning being a phenomenon that can only arise in communities.
Often an agent is seen as an autonomous entity. Each agent has only limited knowledge; it can only access its own internal state and does not have the possibility to access the internal state of other agents. The interactions between the agents are local; this means that no agent has an overview of the system, the sharing of ontologies only happens through local interactions. The agents are autonomous; their view of the world, their acting on it and communicating about it is independent from other agents. Finally, there is no global synchronisation; no clock drives the simulation, the multi-agent system is distributed and parallel. This approach has several important features (Steels, 1998a).

- There is no central controlling agency. Coherence arises in a bottom-up and self-organised fashion. Stable states of the system emerge from the simple rules and interactions of the individual agents.
- The system is open. Agents can enter and leave the community without seriously disturbing the equilibrium of the system. All agents will adapt their internal representations to accommodate the change in the system[9].
- The agents are adaptive. When the world the agents observe changes, the agents will create new meanings and lexicalisations to describe the world. When new word forms are introduced the agents will adapt their internal representation to absorb the meaning of these new word forms.

We already defined the abstract entity of an agent. A set of agents is called a *population*, formalised as $\mathcal{A} = \{A_1, \ldots, A_N\}$. In this work only small populations are used (typically not larger than 20 agents). The mechanisms used can scale up to populations of thousands of agents (Steels et al., 2002).

4.5 Summary

This chapter explained the internal organisation of the an agent. Colour perception is modelled using a conversion from spectral measurements to

[9]Kaplan (2000, p. 45-84) has investigated language games in changing populations. He made a thorough analysis of the influence of the flux of agents on the stability of the population's communicative success (see section 5.2). He describes how a large influx of agents can disturb the system, as the population is not allowed enough time to reach a stable state. When however the influx is decreased the population will again build up coherence and reach a stable state.

an internal representation space. Colour categories are represented using adaptive networks, these compute a membership value based on summing the reaction of many locally tuned receptors. The lexicalisation is modelled by an association between word forms and colour categories. The internal representation of an agent is identical for both the adaptive and evolutionary approach.

Chapter 5

The adaptive model

This chapter describes the first variant of the simulation where colour categories are *learned* according to environmental, ecological and linguistic pressure.

5.1 Language games

The perception converts physical signals to a psychophysical representation, the categorization partitions the psychophysical space, and form-meaning associations couple a category to one or more lexicalisations. All this, however, only describes the static internals of the agent. The dynamics are determined by the way the agents use these internal representation, and this happens during interactions with other agents. The interactions are defined in the form of *language games* (Steels, 1996a,b, 1997a,c). Language games are played in a population of homogeneous agents. The games follow strict rules and the outcome of them steers the adaptation of the internal representations of the agents. During the language games, the agents have no access to the internal representations of the other agents in the population.

5.1.1 The discrimination game

The *discrimination game* is a simulation involving individual agents, and is an essential component of a language game. Unlike the guessing game (see 5.1.2), the discrimination game lacks a social and cultural dimension since agents do not communicate[1]. In fact the discrimination game can be played with one single agent, but a larger population of agents is preferred in order to obtain statistically significant results and to be able to compare results between populations. The discrimination game is the first

[1]Strictly speaking, the discrimination game is not a *language* game as it does not involve the use of linguistic communication.

step needed for other language games; language games are agent interactions in which aspects of language formation and language dynamics are investigated (Steels, 1996b, 1997b,a, 1998a).

The goal of the discrimination game (DG) is to construct conceptualisations in order to successfully discriminate a stimulus from a set of surrounding stimuli. The DG has been defined by Steels (1996a, 1997c), and several variations of the discrimination game have been investigated and experimented with, see for example (de Jong and Vogt, 1998; Steels and Kaplan, 1998; Kaplan, 2000).

The discrimination game is a selectionist and adaptive mechanism. The agent's internal representation undergoes selective pressure from the environment and from the dynamics of the discrimination game. The game is also adaptive in the sense that it adapts the internal representation of agents to be more successful in future discrimination games. When the agents are confronted with an open environment, the agent's internal representation will mould itself onto the changing conditions. The important issue here is that with an open environment there is no equilibrium state to be reached, there is a constant evolution towards fulfilling the task imposed by the environment.

The discrimination game mechanism

The discrimination game follows a fairly simple scenario, involving a single agent. The game needs a context $O = \langle o_1, ..., o_N \rangle$ containing N objects o_i. In the games reported here, the objects are colour stimuli. The context is randomly chosen by the simulation and presented to the agent. The goal of the discrimination game is to divide the sensory space such as to allow the agent to successfully discriminate objects in the world. So the sensory space is divided into a set of possibly overlapping[2] categories $C = \{c_1, ..., c_M\}$. An agent can perceive objects in the world and for every external object o it has an internal sensory response s. The discrimination game described here is not restricted to colour categories, any continuous perceptual space can be used for playing the discrimination game on. A discrimination game runs through the following steps.

- A context $O = \langle o_1, ..., o_N \rangle$ is presented to the agent. Out of the context O a topic o_t is chosen. The topic is the object that has to be discriminated from the other objects in the context[3].

[2]The fact that the categories can overlap differs from Steels's approach, where categories are clear-cut divisions of the agent's sensory channels that can not overlap and have no fuzzy boundaries.

[3]A possible variation on the discrimination game would be that the goal of the game is to discriminate all objects in the context from each other, instead of only discriminating the topic from the other objects in the context. Discriminating all objects would result in a more complex game, with slower convergence.

- The agent perceives each object o_i, and returns a sensory representation for each object s_i: $\langle o_1, \ldots, o_N \rangle \rightarrow \langle s_1, \ldots, s_N \rangle$.
- For all sensory representations s_i a closest matching category $c_{s_i} \in C$ is found[4], according to,

  $$\forall c \in C : y_c(s_i) \leq \hat{y}(s_i)$$

  y_c is the output of the adaptive network belonging to category c, and $\hat{y}$ is the output of the category c_{s_i} reacting best to the sensory experience s_i. See (4.7).
- If the topic can be matched by a category which is not used for matching the other objects in the context, the discrimination game is successful. If not, the game has failed. In other words, the category matching the topic should be unique,

  $$\text{count}\left(\left\langle c'_{s_1}, ..., c'_{s_N} \right\rangle, c'_{s_t}\right) = 1$$

  c'_{s_t} is the category matching the topic stimulus s_t.

This scenario can fail in two ways: the agent has no categories yet ($C = \emptyset$), or the category describing the topic also matches one or more other objects of the context. Now two actions can be taken according to the situation.

1. If the agent has no categories yet, a new category will be created describing the topic. In concreto, this is done by creating a new adaptive network with one locally tuned unit with its centre $\mathbf{m}$ on the representation s_t of the topic o_t.
2. If no discriminating category could be found, there can be two possible actions:
   (a) Either a new category is created to represent the topic or,
   (b) The best matching category is adapted to better represent the topic.

   The choice to add a new category or to adapt an existing one depends on a threshold τ_{adapt}. If the discriminative success of the agent (see 5.2 on page 79) is lower than τ_{adapt} a new category will be created, if not the closest matching category will be adapted.

[4]Note that more than one category could be the best matching, for the sake of clarity let's assume that only one category is found as being the closest match. If two or more best matching categories are found, only one is selected. This could be random selection, or a selection according to some measure, e.g. the successfulness of the category in previous games.

Discrimination games and colour discrimination

The mechanism described above is not dedicated to any specific implementation as it does not give details on the kind of sensory information the agent receives, it is independent of the implementation of the categories, and can be used for simulated as well as real autonomous agents. This paragraph elaborates on how the discrimination game is used in this work to discriminate colour sensations.
Each agent has a category set C containing categories represented by adaptive networks. The adaptive networks have a numerical colour representation as input, in the experiments this will be a three-dimensional vector containing a L^*, a^* and b^*-value from the CIE LAB space. The centres of the locally tuned units are then points in that colour space.

Creating a new colour category When creating a new category (this happens when the agent has no categories yet, or when no discriminating category can be found and the discriminative success of the agent is lower than the threshold τ_{adapt}, a new adaptive network is created with only one locally tuned unit centred on the internal representation of the topic s_t.

Adapting an existing colour category Categories can also be adapted, this happens when no discriminating category can be found for the topic. Adaptation is done by adding a new locally tuned unit to the adaptive network; the new locally tuned unit has the sensory representation s_t of the topic as its centre.

Next to the mechanisms of creating and adapting categories, there is also the changing of the weights of the locally tuned units. When a discrimination game is successful, the weights of the discriminating category (and only the discriminating category) are increased as in (5.1); the term added consists of a non-negative learning rate α weighted by the activation of the locally tuned unit as in (4.6).

$$w_i \leftarrow w_i + \alpha z_i(s_t) \tag{5.1}$$

After every discrimination game, the weights of all the locally tuned units of all categories of an agent are decreased with a non-negative decay $\beta \leq 1$, as in (5.2).

$$w_i \leftarrow \beta w_i \tag{5.2}$$

The weights are limited between $[0, 1]$. When the weight of a locally tuned unit is zero, it is removed from the adaptive network. More details on the consequences of this can be found in section 5.1.4.

Weight decay is standard in the neural network literature (Rumelhart and McClelland, 1986; Krogh and Hertz, 1995) and is used here to keep unused locally tuned units from cluttering the category. It has as effect that the output of the categories becomes less strong; without increasing the weights again, the output of the categories would become insignificant after a number of these steps. But the weight increase in 5.1 prevents this from happening. This increase and decay of weights makes the nature of the categories dependant on the environment. If a certain category is not used for discrimination for some time, its weights will have decreased. If the parameter α, which controls the rate of weight decay, is set too low, the categories will "fade out" before their weights can be increased again. α is a bit of an arbitrary parameter, even when set to 1 the discrimination will not be influenced. It is a "household" parameter of the simulation, meant to keep the categories free from unreactive locally tuned units. Experience learns that setting α in range of $[0.95, 1]$ delivers good results.

Why discrimination?

To conclude this section on the discrimination game, I would like to elaborate on the reason for chosing discrimination. As was mentioned before, the visual world arrives at our eyes as a blur of electromagnetic energy in which we detect regularities such as shape, texture and colour. Our communication on the other hand is discrete[5], its main constituents being vocalised, written and visual symbols. In between the continuous perception and the discrete communication, there has to be a step that carves up the perception into categories. These categories can and will not only be used to communicate: the power to symbolise did most probably not evolve for the benefit of language[6], but for other cognitive faculties (such as recognition, learning, etcetera) (Deacon, 1997). Language only draws upon these symbolic representations.

To drive the symbolisation of perception, a task is needed. In a natural environment this would be a task related to survival or to social interactions; such as distinguishing a predator from a rock, food from foliage, or a possible mate from a competing male. In this work, the task to create symbols takes the form of the discrimination of objects. A set of objects is presented to an observer and the task consists of discriminating one object from the others. This requires the creation of appropriate internal representations and ontologies. Needless to say that other tasks might also arrive at a symbolisation of the world; for example detecting similarities, rather than differences, between objects needs a classification of perception as well.

[5]Communication does not need to be discrete, for example one can imagine how the volume of an auditory signal is related to the hungriness one feels.

[6]Though Davidoff (2001) puts forth the suggestion that language might have originated to aid categorisation.

5.1.2 The guessing game

The guessing game[7] is played at the population level; as opposed to the discrimination game, which is played at the individual level. In the guessing game two agents are randomly chosen from the population; one acts as the *speaker*, the other as the *hearer*. A common context is presented to both agents and the goal is to communicate about a topic chosen from the context. In the guessing game only the speaker knows the topic, and communicates a word form for it to the hearer. The hearer then interprets the word form and tries to identify the topic in the context.

The following narrative might clarify the idea of a guessing game. Suppose that an Englishman and a Frenchman are sitting at dinner. The Frenchman is the speaker, while the Englishman is the hearer. On the table several kinds of dishes are placed; the dishes form the context. Now imagine that the topic is chosen to be the potatoes. The Frenchman perceives the potatoes, and finds the associated word for it. He says "pommes de terre". The Englishman, acting as the hearer, interprets this. If he doesn't know the word, the game fails. If he does know "pommes de terre", he looks at the table and points out what corresponds best to his knowledge of "pommes de terre". There are two possibilities: he points at the potatoes and the game succeeds, or he points at the wrong dish and the game fails. If the game fails, the Frenchman shows him topic, which are the potatoes, and the Englishman can try to associate his internal representation of the potatoes with the word "pommes de terre".

The guessing game is one of many variations of a language game. Variants include the ostensive game (Vogt, 2000) in which both agents know the topic before the conversation starts, the hearer still receives feedback from the speaker. In the observational game (Vogt, 2000) both agents know the topic, but no feedback is given on the correctness of the word forms; according to some learning parameter the agents pick up word forms from each other. In the imitation game (de Boer, 2001) the agents try to imitate forms, shifting their internal representations to better resemble others and to maximise understandability; this has been used to simulate the formation of vowel systems.

The guessing game mechanism

Out of a population of agents $\mathcal{A}$ two agents A_s and A_h are randomly picked ($A_s \neq A_h$). The guessing game proceeds along the following scenario:

- A context $O = \langle o_1, ..., o_N \rangle$ is presented to both agents. Out of the context a topic $o_t \in O$ is chosen; only the speaker A_s knows the topic.

[7] The guessing game is not a game in the sense that there is a winner and loser, it is called a "game" because the interactions follow strict rules and of course because of the similarity to the Wittgensteinian notion of language games (Steels, 1999, p. 24).

- The speaker A_s tries to discriminate the topic by playing a discrimination game. If the discrimination game fails, then the guessing game fails as well. If the speaker finds a discriminating category c', the game continues.
- The speaker looks in its lexicon if any word forms are associated with c'. If not, with a certain probability p_{cf} a new word form f is created and associated with c'. If only one word form is associated with c', this is the word form f that will be communicated. If however two or more word forms are associated with c', then one word form f is selected with the highest score for category c'.
- The speaker A_s conveys the word form f to the hearer A_h.
- The hearer checks if it has word form f in its associative memory, if not the game fails: the hearer is shown the topic o_t and it learns the speaker's word (see below). If the hearer does have the word form f in its lexicon, it finds the associated category c" and matches this category to all the objects in the context O. The hearer points at the best matching object o_m. The best matching object is the object to which c" has the highest reaction,

$$o_m = \arg\max_i \left(y_{c"} \left(o_i \right) \right)$$

- If the hearer succeeds in pointing out the topic, the game is successful. If the hearer points out the wrong object, the speaker identifies the topic and the hearer adapts its category c" to better match the topic in future games (see below).

Learning and adapting categories During a guessing game the internal representation of the agents can be changed on two occasions: when the hearer does not know a word form used by the speaker, and when the hearer incorrectly identifies the topic. The hearer respectively learns a new form-meaning association or adapts an existing meaning.
When the hearer does not know the word form, the speaker points out the topic and the hearer tries to find a discriminating category for the topic. If a discriminating category is found, f is associated with it. If not, the hearer creates a new form-meaning association. This association is a pair containing the word form f and a new category centred on the representation s_t of the topic.
When the hearer knows the word form, but points at another object instead of at the topic, the game fails. In this case the meaning needs to be adapted. This happens by adding a locally tuned unit centred on the sensory representation of the topic s_{o_t}. In a nutshell: the topic is used to refine the

category, so that the category represents the topic better in the future.

5.1.3 Form-meaning associations

The association between form and meaning is an important one. The *score*, a scalar between zero and unity signifying the strength of the association, determines the association. Every form has a score through which it is bound with its meaning. During a guessing game the score of the uttered form is increased or decreased depending on the outcome of the game. This happens only for the form-meaning association of the hearer.

- If the communication was successful, the hearer increases the weight of the score between f and the category c'' which matched to topic by δ_{increase}, as in (5.3).

$$score \leftarrow score + \delta_{\text{increase}} \tag{5.3}$$

  The score of the other word forms associated with category c'' are decreased[8] by δ_{decrease}, as in 5.4.

$$score \leftarrow score - \delta_{\text{decrease}} \tag{5.4}$$

- The score of the hearer is decreased when, during a guessing game, it knows the communicated form but fails to point out the topic. The score between f and its meaning is decreased by δ_{decrease} as in 5.4.

δ_{increase} and δ_{decrease} are small values, in the simulation $\delta_{\text{increase}} = \delta_{\text{decrease}} = 0.1$. When a new word form is created for a category or when a word form is learned, its score is initially set to $score_{\text{default}} = 0.8$. There is no real rationale behind this number, except that setting the initial score to 1 would make new word forms immediately dominant, which not natural; and setting the default score too low (< 0.5) would make it difficult for new word forms to become accepted in the population.

5.1.4 Removing word forms and meanings

There is also a mechanism for "forgetting" underused or unsuccessful categories and word forms. If this would not happen, the agent's repertoire would get cluttered with insignificant and therefore unused categories and

[8]This *lateral inhibition* is also used in the training artificial neural networks and is used in a similar manner by Oliphant and Batali (1997) in experiments on the emergence of communication.

unused word forms. The removal of categories and words serves an aesthetical purpose, for the dynamics of the simulation it is not necessary to remove unused and unreactive structures. There is the removing of word forms and the removing of categories, each has its specific conditions.

Removing categories

A category is removed when the weight of the locally tuned units of its category has decreased beyond a small value $\epsilon = 0.01$, indicating that the category hardly reacts anymore to input; or when it has not been used for a certain period of time and has not been lexicalised yet. (5.5) shows the condition for removing a category; W is the set containing all the weights of all the locally tuned units of the category, $\theta'_{\text{last used}}$ is a threshold and F is the set containing all word forms associated to the category.

$$(\forall w \in W : w < \epsilon) \vee \Big((last\ used > N.\theta'_{\text{last used}}) \wedge (F = \emptyset) \Big) \qquad (5.5)$$

The conditions described here are not supported by any psychological observations. They just prove to work well in the simulation and serve their purpose of keeping the agents' memory free of unused form-meaning associations.

Removing word forms

For each word form four scores are kept: *age*, *last used*, *use* and *success*. The age is initialised to zero, and is increased with one every time the agent is involved in a guessing game. The last used score is reset to zero when the word is used during a guessing game, it is increased with one when the agent is involved as speaker in a guessing game. The use is initialised to zero at creation of the word form, and is increased with one when the word form is used by the agent in the role of speaker. The success is initialised to zero, and is increased with one when the word form is successful in a guessing game with the agent in the role of speaker. The success score of a word is not to be confused with the success score of an agent, which measures the success of an agent during guessing games.

A word form will be deleted from the repertoire of the agent when it is not successful enough, or when it has not been used for a certain number of games; (5.6) shows the condition for deciding whether a word form will be deleted from the agent's repertoire. θ_{wfage} is the age threshold (word forms have to have a certain age before they are considered for removal), θ_{score} is the use score threshold, and $\theta_{\text{last used}}$ is the last used threshold. Some thresholds are dependant on the size of the population and are therefore defined for a population of one agent and made relative to the population size by multiplying by N, the number of agents in the population.

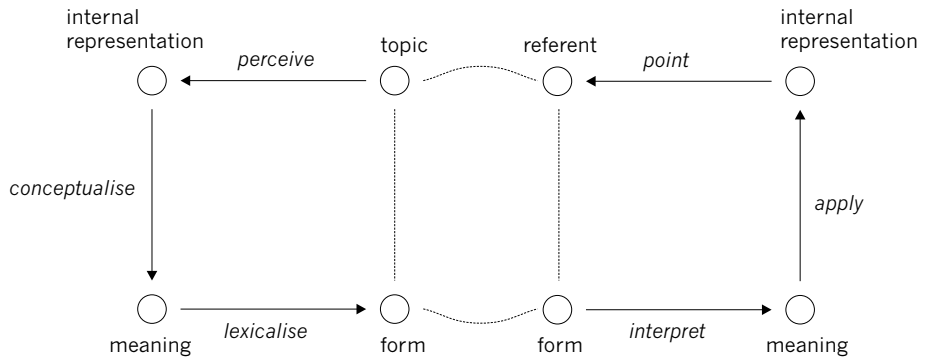


Figure 5.1: The semiotic square of the speaker (left) and hearer (right). The speaker perceives the topic as point in its internal representation space, the point is mapped onto a meaning (or category), and this meaning is associated with a word form. The hearer interprets the word form, finds the associated meaning and applies the meaning to the internal representation of the world to find the referent meant by the speaker. The referents and word forms are coupled indirectly (the dotted line) trough the internal representation and meaning.

$$(age > N.\theta_{\text{wfage}}) \wedge ((score < \theta_{\text{score}}) \vee (last\ used > N.\theta_{\text{last used}})) \qquad (5.6)$$

Default values for the thresholds can be found in the chapters reporting on the results of the simulations.

5.1.5 The semiotic square

When speaking of external stimuli, internal representations, meaning and word forms; we can not avoid venturing into the field of semiotics (Chandler, 2001). Semiotics is the study of signs. Peirce defines a sign as "something which stands to somebody for something in some respect or capacity"; this is not really instructive, but rather shows that a simple definition is not straightforward. Two influential theories exist. de Saussure (1974) views a sign as the dyadic association of a signifier with the signified. The signifier is the form that the sign takes (for example the word "open"), while the signified is the represented concept (the shop that is open for business). Peirce (1958) defines a sign as a triadic association of a representamen, interpretant and object. The representamen is the form that the sign takes (for example "open"), the interpretant is the sense made of the sign (the idea that the shop is open) and the object is the thing to which the sign refers (the fact that one can enter the shop to buy something).

The field of semiotics knows a wide tradition of interpreting and refining each others definitions, and it is with a certain reserve that I present an

additional interpretation of the semiotic relation proposed by Steels (1999), in which the semiotic triangle of Peirce is made into a semiotic square. The semiotic square has four constituents: the referent, the internal representation, the meaning and the form. The referent is the object or stimulus as it occurs in the world; for example, a flame or a greenish stimulus. The internal representation is the result of the mapping of the sensory enervation onto an internal sensory-map; the pain caused by the flame on a somatosensory map or the opponent-channel representation of the greenish light. The meaning is the concept or category to which the internal representation matches; for example, [PAIN-HAND] or [GREEN]. And the form is the utterance associated with the meaning. The relation between the referent and the form is indirect, and comes through the coupling of the four constituents.

Figure 5.1 uses the semiotic square to illustrate the relation between the speaker and hearer in a language game. The speaker and hearer only have access to the object or stimulus in the world, and to the form. The meanings of both the speaker and the hearer are coupled through social interactions.

5.2 Analysing simulation results

When constructing simulations, one should also spend time constructing sensible ways to objectively report the results of the simulations. While informal presentation of the results, such as a plot of the categories in colour space or a plot of the categories on a Munsell chart can be instructive, rigorous measures are needed to be able to compare and analyse results. This section describes statistical measures used in the various experiments. Each measure is explained, a formal definition or algorithm is given, the motivation for choosing the measure is given, the limitations are sketched and where appropriate, bounds and typical values are given.

Number of categories

The *number of categories* of an agent, whether or not lexicalised, can be informative for two reasons: when plotted against time it shows if the simulation can reach a stable category set. Secondly the number of categories is a gauge for the complexity of the environment: a simple environment with few colour samples provides less pressure to generate large category sets, since the agents do not need many categories to discriminate the simple contexts. Large category sets can be attributed to a complex context, which exerts a higher environmental pressure.

The number of categories the agents need to successfully discriminate the colour context and to communicate about colour samples can be specifically interesting in the light of observations of human colour lexicons.

Some students of colour naming have suggested that there is a boundary to the number of colour terms that are successfully and frequently used in a language community, a boundary emanating from the make-up of the human colour perception, see for example (Berlin and Kay, 1969).

Number of word forms

The *number of word forms* is different from the number of categories an agent possesses. A category can be associated with none, one or more word forms; and the same word form can be associated with more than one category. This allows for *synonymy* and *homonymy*. The number of different word forms circulating the population is indicative for the quality of the dynamics of the simulation. Many different word forms only used and understood by a limited number of agents will have a negative influence on the overall communicative success of the agents (see p. 79), the opposite is true for a set of word forms shared by a large number of agents.

Discriminative success

The *discriminative success* shows how well an agent is doing at discriminating the topic from the context; the discriminative success is the running average over T_a games and has a maximum of 1. At the end of a discrimination game i an agent A is given a discrimination score ds_i^A, $ds_i^A = 1$ if the agent could discriminate the topic from the context, otherwise $ds_i^A = 0$. The discriminative success at game i is computed as in (5.7), for $j < 1$ we consider ds_j^A to be 0. It is the number of successful discrimination games during the last T_a games, divided by T_a.

$$DS_i^A = \frac{\sum_{j=i-T_a}^{T_a} ds_j^A}{T_a} \tag{5.7}$$

The average discriminative success is the discriminative success averaged over all N agents in the population $\mathcal{A}$.

$$\overline{DS}^{\mathcal{A}}_i = \frac{\sum DS_i^A}{N} \tag{5.8}$$

Communicative success

The *communicative success* is the number of successful guessing games during T_a games. If agent A during game i is able to communicate a meaning to another agent, its communication score is $cs_i^A = 1$, otherwise $cs_i^A = 0$. The communicate success CS_i^A of an agent at game i is defined as in (5.9),

for $j < 0$ we consider cs_j^A to be 0. It is the number of successful guessing games for the speaker during the last T_a games, divided by T_a. The communicative success has a minimum of 0 and a maximum of 1.

$$CS_i^A = \frac{\sum_{j=i-T_a}^{T_a} cs_j^A}{T_a} \tag{5.9}$$

The *average communicative success* is the communicative success averaged over all N agents in a population $\mathcal{A}$.

$$\overline{CS}^{\mathcal{A}} = \frac{\sum CS_i^A}{N} \tag{5.10}$$

Distance between two sets

For the next measure a distance metric between two sets of points in N-dimensional space will be needed. Our distance metric D_{set} needs to fulfil the following requirements. $D_{\text{set}} : S \times S \to \Re$ is metric on a nonempty set S of sets, if for all $A, B, C \in S$ we have:

- The distance between two identical sets is zero, $D_{\text{set}}(A, A) = 0$.
- The distance is symmetrical, $D_{\text{set}}(A, B) = D_{\text{set}}(B, A)$.
- The distance should be non-negative, $D_{\text{set}}(A, B) \geq 0$.
- The distance metric should satisfy the triangular inequality, meaning that two sets that are highly dissimilar cannot be both similar to some third set: $D_{\text{set}}(A, B) + D_{\text{set}}(B, C) \geq D_{\text{set}}(A, C)$.

Furthermore, as the sets can contain a different number of points, $|A| \lesseqgtr |B|$, we require a distance metric capable of coping with this. Therefore, the measure should not require explicit pairing of points. This excludes some well-known distance metrics, such as the Euclidean distance or the Manhattan distance. One possibility[9] is to take the *Hausdorff metric*, defined as

$$H(A, B) = max(h(A, B), h(B, A)) \tag{5.11}$$

with

[9]There exist more distance metrics fulfilling some or all of our requirements (Eiter and Mannila, 1997; Ramon and Bruynooghe, 2001), but they all rely on computing a matching between point sets. As they require a matching, defined by the experimenter, the metrics are not objective. Another drawback is that although most of these algorithms complete in polynomial time, they are still too slow to allow extensive simulations.

$$h(A,B) = \max_{a \in A}(\min_{b \in B} \|a - b\|) \tag{5.12}$$

h is called the *directed Hausdorff metric*, it is the Euclidian distance between the point of set A farthest from set B and the point of B closest to A. The Hausdorff metric does not live up to the expectations one has of a psychological distance metric and is therefore less suited for our purposes.

Another possible metric, which does not require an equal number of elements in set A and B, is the *sum of minimum distances* function. It is defined as

$$d_{md}(A,B) = \frac{1}{2}\left(\sum_{a \in A} \min_{b \in B} \|a - b\| + \sum_{b \in B} \min_{a \in A} \|a - b\|\right) \tag{5.13}$$

It sums the minimum distances between each element and the elements in the other set. The distance metric $\|a - b\|$ can be any distance metric; but unless mentioned, we will consider the metric to be Euclidean distance between point a and point b. Unfortunately it is not weighted by the number of elements in each set, making it less useful for our purposes. However, it does provide the basis for the distance metric I would like to propose in (5.14).

Just as the metric defined in (5.13) it sums the minimum distance between each element to the other set, but now the sum is weighted by the number of elements in set A and B. We will call this metric the *weighted sum of minimum distances*.

$$D_{\text{set}}(A,B) = \frac{\sum\limits_{a \in A} \min\limits_{b \in B} \|a - b\| + \sum\limits_{b \in B} \min\limits_{a \in A} \|a - b\|}{|A| \,.\, |B|} \tag{5.14}$$

Table 5.1 illustrates how the Hausdorff metric, the sum of minimum distances function and the weighted sum of minimum distances metric behave (for $\|\cdot\|$ the Euclidean distance is taken). The dataset used is shown in figure 5.2, $A = \{a, b, c, d\}$, $B = \{e, f, g\}$, $C = \{a\}$, $D = \{b, c\}$, $E = \{f, g\}$, $F = \{b, c, d\}$ and $G = \{f\}$. The first two examples show that all three metrics are symmetrical ($d(A,B) = d(B,A)$), the third example shows that all three metrics return a distance of zero for identical sets, the other examples give an idea of how the metrics behave. The Hausdorff metric sees no difference between set (D,E) and (B,F) as well as between (A,G) and (C,G); while the other two metrics, as one should expect, do. In the last four examples one can also[10] see that the weighted sum of distances has a psychologically more faithful behaviour than the sum of distances metric; justifying our choice for the weighted sum of distances metric.

[10] Rather subjectively.

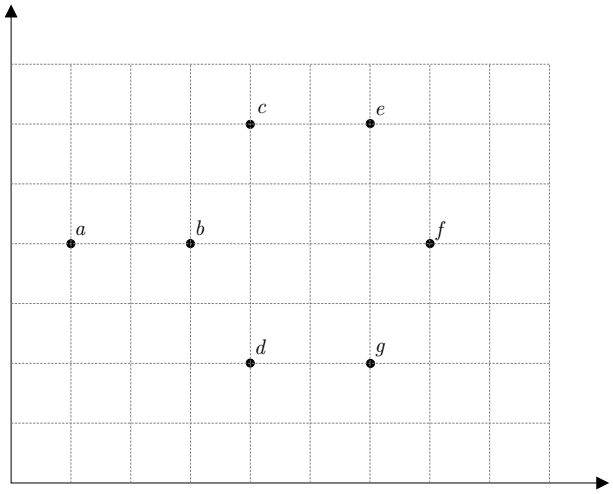


Figure 5.2: Data for illustrating three distance metrics; $A = \{a, b, c, d\}$, $B = \{e, f, g\}$, $C = \{a\}$, $D = \{b, c\}$, $E = \{f, g\}$, $F = \{b, c, d\}$ and $G = \{f\}$ (see text).

	Hausdorff metric	Sum of minimum distances	Weighted sum of minimum distances
$d(A, B)$	5.39	10.30	1.72
$d(B, A)$	5.39	10.30	1.72
$d(A, A)$	0	0	0
$d(B, C)$	6.00	11.08	7.39
$d(D, E)$	3.61	7.21	3.61
$d(B, F)$	3.61	7.61	1.69
$d(A, G)$	6.00	10.41	5.20
$d(C, G)$	6.00	6.00	12.00

Table 5.1: Illustration of three distance metrics.

Interpretation variance

When hearing a word form, do all the agents relate that word form to the same meaning? In other words, is there coherence in their interpretation of word forms? This is what the *interpretation variance* measures.
The interpretation variance of a word form f is defined as (5.15), C is the set of categories associated with f over the whole population, $|C|$ is the number of elements in C, and $\mathbf{m}_c$ is the set containing the centres of the locally tuned units of the category c. D_{set} is a distance measure between two sets of points, as defined in 5.14.

$$IC_f = \frac{1}{\frac{1}{2}\ |C|\ (|C|-1)} \sum_{i=1}^{|C|} \sum_{j=i+1}^{|C|} D_{\text{set}}\left(\mathbf{m}_{c_i}, \mathbf{m}_{c_j}\right) \tag{5.15}$$

The variance computed in (5.15) computes the interpretation variance for one word form only. It is more informative to compute the interpretation variance over the entire population, this done with (5.16); it is the sum of the variance for every word form in the population, weighted by the number of agents having the respective word form. $F^{'}$ is the set of all unique word forms present in the population, a_f is the number of agents having word form f.

$$\overline{IC} = \frac{\sum_{F^{'}} a_f\ IC_f}{\sum_{F^{'}} a_f} \tag{5.16}$$

If $IC = 0$ all agents have identical meanings for all word forms.

Category variance

The interpretation coherence gives a measure for how good the agents are at associating word forms with the same referent. But it is also interesting to analyse the opposite: if a stimulus is presented to the agents, will they coherently react to it? In other words: if two agents see the same colour stimulus, will they interpret it as belonging to the same colour category? When there is a finite number of meanings and a finite number of referents, it is easy to define a production coherence measure since it is possible to build a coding matrix relating a discrete set of word forms to a discrete set of meanings; see for example (Steels, 1996b; Oliphant and Batali, 1997; Steels and Kaplan, 1998; Kaplan, 2000; Steels and Kaplan, 1999a). In the system described here, where there is an unlimited number of non-discrete referents and meanings, this is impossible to do. Figure 5.3 illustrates our situation. There are two agents, each with a number of categories in their internal representation space. Each category consists of a number of instances, or points in the representation space, organised in an adaptive network. The category variance measures how much the internal categories of the agents agree, without taking the labels of the categories into account.

For the *category variance*, the distance between two categories of two agents needs to be computed; for this we use the weighted sum of minimum distances function from (5.14). As this function only gives the distance between two categories, and our goal is to get a distance metric between two category *sets* we use the weighted sum of minimum distances metric a second time, but this time to compute the distance between category sets. We now get

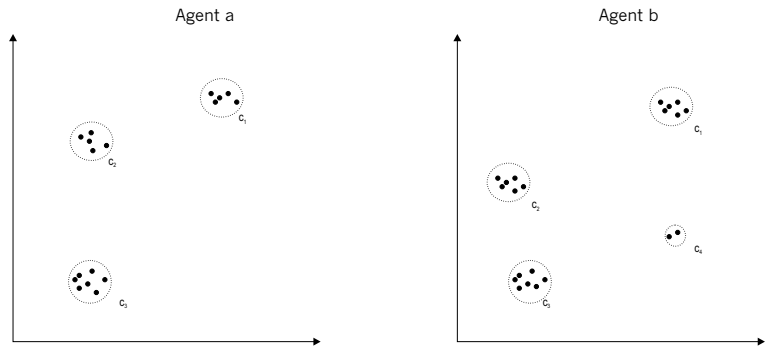


Figure 5.3: Illustration of a two-dimensional representation space of two agents a and b. Each agent can have a different number of categories. The categories consist of instances, or points in the representation space, organised in an adaptive network (not shown on the illustration). If the categories are placed on similar locations in the space, we expect a low category variance.

$$D_{\text{category set}}\left(A, A'\right) = \frac{\sum\limits_{c \in A} \min\limits_{c' \in A'} D_{\text{set}}\left(\mathbf{m}_c, \mathbf{m}_{c'}\right) + \sum\limits_{c' \in A'} \min\limits_{c \in A} D_{\text{set}}\left(\mathbf{m}_c, \mathbf{m}_{c'}\right)}{|A| \,.\, |A'|} \tag{5.17}$$

In (5.14) we implement the distance function $\|\cdot\|$ as

$$\left\|\mathbf{m} - \mathbf{m}'\right\| = \frac{1}{w.w'} \sqrt{\sum \left(\mathbf{m} - \mathbf{m}'\right)^2} \tag{5.18}$$

$D_{\text{category set}}$ is the weighted distance between the category sets of agent A and A', $|A|$ and $|A'|$ are the number of categories of each agent. The sets need not have the same number of categories. The more similar the sets are, the lower the $D_{\text{category set}}$ will be. If a category set of an agent would be compared with itself, $D_{\text{category set}}(A, A)$ would be zero.

This however only compares the categories of *two* agents. We would like to measure how well the agents of an entire population agree. This is done by computing the *category variance* CV as in (5.19). It is the sum of all pairwise computed distances between the category sets of all agents, weighted by the number of agents N. As usual with distances, the smaller, the more the category sets resemble each other. For two identical agents, the category variance is 0: $CV(A, A) = 0$

$$CV(A, A') = \frac{1}{2N\left(N-1\right)} \sum_{i=2}^{N} \sum_{j=i-1}^{N} D_{\text{category set}}\left(\mathcal{A}_i, \mathcal{A}_j\right) \tag{5.19}$$

Category variance between two populations

The *category variance across populations* CV' is used if category sets of two populations are to be compared. It is the average of the category variance computed between all agents of two populations $\mathcal{A}$ and $\mathcal{A}'$, and is defined as

$$CV'(\mathcal{A}, \mathcal{A}') = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} CV\left(A_i, A'_j\right) \tag{5.20}$$

N is the number of agents in a population, we assume the number of agents to be equal in both populations. The more similar the category sets are in two populations, the lower CV' will be. If all agents in two populations have identical category sets, then $CV' = 0$.

5.3 Summary

This chapter explains the adaptive approach towards learning categories and their labels. For this two learning paradigms are introduced: the discrimination game and the guessing game. Discrimination games are individual interactions with the environment and serve to create categories with which stimuli in the environment can be categorically distinguished. Guessing games model linguistic interactions between two agents and serve to acquire a shared lexicon in a population of agents. During the guessing game the internal representations of the agents are adapted to allow for more successful communication.

Chapter 6

The evolutionary model

The previous chapter was dedicated to the learning of colour categories. Opposite to individual or culturally learned categories we find *genetically evolved* categories. The simulations for exploring the genetic evolution of colour categories are the topic of this chapter. A combination of both, in which innate structures are fine-tuned through adaptation are also possible, but will be not considered here as we are only interested in implementing approaches at the dividing lines of the literature concerning the "origins" of colour categories.

Evolution implies that the categories are transmitted from generation to generation in the form of a genotype, with evolutionary pressure acting on the phenotype and deciding which individuals are apt for reproduction. The reproduction can be sexual or asexual, for convenience we will limit our evolutionary algorithm to asexual reproduction. As evolution can only function when there is diversity in the population, the asexual reproduction step has to introduce diversity through mutation. During mutation the genes of the offspring are a copy of the parent's genes, but with the introduction of slight random variations. As evolution is undirected (evolution is not purposely evolving *useful* colour categories) these mutations can produce a category set that leaves the offspring performing worse than its parent. On the other hand, if the offspring performs better at the task at hand, it will have a higher fitness and will be more likely to be selected for reproduction. In this way, over generations agents will become more successful at the task(s) imposed by the environment.

The evolutionary approach has been successfully used in computer science to solve optimization problems in a Darwinian manner. Several implementations of evolutionary computation[1] exist which, apart from some details, differ only in the representation of the genotype and the expression of the phenotype. All share the same basic idea of a population of individuals in which variation is introduced and where a task provides se-

[1] An excellent anthology on evolutionary computation can be found in (Fogel, 1998).

lectionist pressure. Every generation new individuals are created which are offspring of the fittest individuals from the previous generation. In *evolutionary programming* the genotype consists of parameters which need to be optimized (Fogel, 1999), *genetic algorithms* (Holland, 1975) encode the genotype in a bitstring and *genetic programming* (Koza, 1992)Koza, J. implements the genotype in a computer program, directly coding the phenotype in the genotype.

6.1 Evolution of colour categories

Each agent has a set of "colour genes", these are actually the categorical networks, so we shortcut the problem of modelling gene expression. The evolutionary algorithm thus functions on the structures explicated in section 4 on page 52. Remember that a colour category is represented by an adaptive network, consisting of a set of locally tuned units connected to a summing output unit (see figure 4.3). These networks do not change during the lifetime of the agent, networks only change by means of inter-generation mutation.

The evolutionary algorithm starts with an initial population of agents. In each generation all agents are tested on a task, the performance of each agent on that task is called the *fitness*. The fitness is used to select the agents that will be copied and mutated for the next generation.

	Variant 1: Genetic evolution without language	Variant 2: Genetic evolution with language
Task	Agents are tested on the discriminative potential of their categories	Agents are tested on a their potential to communicate colour meaning
Fitness measure	Discriminative success	Communicative success
Age structure	Non-overlapping population	Overlapping population
Reproduction	Mutation of category set of one single parent	Mutation of category set of one single parent

Table 6.1: Main properties of the two variants of the evolutionary simulation.

Two variations on the evolutionary model are used. The main properties of both simulation variants are summarised in table 6.1. Both variants cover both ends of the evolutionary spectrum: one the one hand we have genetic evolution of conceptual structures without communication, on the other hand we have genetic evolution complemented with cultural evolution. Investigating both variants should give insights into the dynamics and diachronic behaviour of genetically evolved grounded concepts, and should allow a fair comparison with the adaptive model. The task on which

the agents are judged is different for the two variants of the evolutionary model. The first variant evolves categories and the fitness is only judged on the discriminative power of the evolved category sets. The second variant evolves categories as well, but the fitness is instead judged on a linguistic task. The difference being that in the first variant the evolutionary pressure comes from a non-linguistic task, as opposed to the second variant where the evolutionary pressure solely comes from agent's potential to communicate the meanings of its categories. The fitness measure in the former is the discriminative success, in the latter is it the communicative success. Note that the second variant evaluates agents on their communicative behaviour, so the evolution of categories is under influence of language. As selectionist pressure for creating categories is now indirect, it might be interesting to see which categories are created of any at all.

6.1.1 Mutation of colour categories

The mutation of the colour categories is identical for both variants of the simulation: offspring has a mutated set of categories inherited from one parent. The category sets can be mutated in four ways: by adding or removing a category or by extending or restricting a category. The probability of one of these mutations occurring is defined by $p_{\text{mut}} \in [0, 1]$. Only one mutation operation of four is executed, so the probability of a mutation operator being chosen is actually $p_{\text{mut}}/4$.

An agent A has a set of categories $C = \{c_1, \ldots, c_{|C|}\}$, each category is represented by a network of locally tuned units with centres $\mathbf{m_i}$ and weights w_i: $c = \{\{\mathbf{m_1}, \mathbf{w_1}\}, \ldots, \{\mathbf{m}_{|\mathbf{c}|}, \mathbf{w}_{|\mathbf{c}|}\}\}$. The four mutation operators are:

Adding a category Adding a category c' to the category set is done by adding an adaptive network with one locally tuned unit $\{\mathbf{m}', \sigma_{\text{default}}\}$ at a random location $\mathbf{m}'$ in the $L^*a^*b^*$ representation space[2].

$$C' = C \cup c'$$

σ_{default} is set to 5.0, as in 4.2.1.

Removing a category A random category c' is removed from the category set.

$$C' = C \setminus c'$$

[2]The location of $\mathbf{m}'$ is limited by the space defined by the Munsell solid.

Extending a category First a category c is randomly selected. It is extended by adding a locally tuned unit, the centre of the locally tuned unit $\mathbf{m}'$ being a normal random deviation from the centroid $\mathbf{c}$ of category c.

$$c = c \cup \left\{\mathbf{m}', \sigma_{\text{default}}\right\}$$

The centroid $\mathbf{c}$ of the category is computed as (6.1). The centre $\mathbf{m}'$ of the added locally tuned unit is now randomly chosen from a normal distribution with mean $\mathbf{c}$ and standard deviation $\sigma_{\text{mut}} = 5.0$.

$$\mathbf{c} = \frac{\sum w_i \mathbf{m}_{c,i}}{\sum w_i} \tag{6.1}$$

Restricting a category A category c is randomly chosen, and from that category one locally tuned unit j is randomly chosen for removal.

$$c = c \setminus \{\mathbf{m_j}, \sigma_\mathbf{j}\}$$

If c contains only one locally tuned unit, removing this locally tuned unit would leave the category without any units, rendering it useless for further classification. That is why c is then removed from the agent's category set.

The agents initially start at generation 0 with no colour categories, implicating that they need to create at least two categories before discrimination is possible. If an agent has only one category, there is no selective advantage for that agent. In order to "jump start" the evolution a fifth mutation operator is introduced:

Adding two categories Adding two categories to the agent's category set by creating two adaptive networks each with one locally tuned unit at a random location in the representation space.

$$C' = C \cup c' \cup c''$$

This operator is basically identical to executing the first operator twice. The probability of this occurring is $p_{\text{mut}}/4$, but only when the agent has less than two categories in its repertoire. As soon as the agent has two or more categories, this operator is not considered during the mutation step.

6.1.2 Genetic evolution without language

The fitness measure

Each agent in the population plays a series of discrimination games; on the understanding that each agent plays the same series of discrimination games to allow comparison of fitness. The fitness of each agent A is defined by the discriminative success DS_i^A of playing a series of T_f discrimination games (see 5.2). At the start of each generation the discriminative success is reset, to ensure that the success of the agents is only determined by the current batch of discrimination games.

The selection process

Of the population the M fittest individuals are chosen for reproduction (in our case M is 50% of the population), and the $(|\mathcal{A}| - M)$ worst individuals are replaced by a mutated copy of the M fittest individuals. In this manner, the size of the population remains constant. Copying and mutating the fittest individuals to the next generation to replace the least fit individuals, and at the same time keeping those fittest individuals from the previous generation ensures that the evolution cannot run backwards (this is not really reminiscent of natural evolution, where each individual, even if very fit, eventually leaves the population by dying of old age).

6.1.3 Genetic evolution with language

The fitness measure

The agents are judged on a series of guessing games. The fitness of an agent A is defined by the communicative success CS_i^A. During the fitness evaluation if a population $\mathcal{A}$, a total of $|\mathcal{A}|(|\mathcal{A}| - 1)$ guessing games are played: each agent plays $|\mathcal{A}| - 1$ games in the role of speaker, and $(|\mathcal{A}| - 1)^2$ games in the role of hearer. At the start of each generation the discriminative and communicative success of all agents is reset.

The selection process

As we like to keep language in the population, and not throw it out each time we move to a new generation, we now introduce an age structure. At the end of a generation, the M_{die} oldest agents are removed from the population. M_{die} typically is between 1 and 10% of the population size, higher rates would prohibit the cultural transmission of category labels. These oldest agents, are replaced with mutated copies of the M_{die} fittest individuals. At the start of each generation, all agents's measures (discriminative and communicative success) are reset, but the repertoire of word forms is

not changed. Agents are thus allowed to take their lexical repertoire to the next generation, allowing for transmission of conventionalised labels for colour categories of older agents to younger agents. In this way, the second variant emulates the universalist account of the nature of colour categories (Berlin and Kay, 1969; Durham, 1991).

Language and evolved categories

The language aspect is implemented as in the adaptive approach. The agents play guessing games, exactly as described in 5.1.2. Of course, the categories are not adapted according to the success or failure of the guessing games. Changing the categorical repertoire is the only done through genetic mutation. All the same measure defined for the language games apply here. Only word forms are learned and are under the same selective pressure as in the adaptive approach: forms and categories have an association strength, represented by the form-meaning score, which is updated according to the outcome of the game. Word forms are also liable the same dynamics for forgetting unsuccessful and unused forms as described in 5.1.4.

6.2 Measures

The measures for the adaptive approach also apply to genetic evolution: we can again compute discriminative and communicative success for the agents, and the coherence of form and meaning can be computed with the interpretation variance and category variance.

The running average T_a over which discriminative and communicative success is taken is now equal to the number of games T_f over which the agents evaluated.

6.3 Why evolve colour categories?

The previous chapter describes the learning and adapting of colour categories. However, many students of colour categories are convinced that colour categories, or structures ontogenetically leading to colour categories, are innate and are thus shared by all humans over all cultures. To verify the plausibility of this hypothesis and to have a system against which the learned categories can be compared, we need a system that forms colour categories through genetic evolution. The basis for both systems, adaptive and evolutionary, is identical. There is (i) a task formed by the discrimination or communication, there is (ii) perception and there is (iii) an internal representation space in which categories are represented. The difference is

in the way categories are formed. In the adaptive approach, an individual learns an appropriate category set. In the evolutionary approach, a fit category set is obtained through selecting on the discriminative or communicative performance of diverse individuals. The time scale on which both approaches operate is poles apart: the former creates categories during ontogeny, while the latter operates over generations.

The evolutionary system allows us to see if evolving categories is *a priori* possible. If it is, it might be interesting to see where the differences might be with an adaptive system. How coherent are the evolved categories? How many categories will be evolved? Should communication of meaning be introduced to obtain "cognitive economy" (Rosch, 1978)? Is the evolutionary approach flexible to environmental changes? Results and an interpretation can be found in chapters 7 and 8.

The system described here does not make any claims on the nature of the evolution of categories in humans and other species. If categories evolve, the route followed could be radically different than the one described here. One can imagine the sensing of spectral content starting with the sensing of continuous gradations of wavelengths, much as the certain flatworms sense gradations in light intensity (it is hard to suspect these worms of having categories for light intensity), and this evolving into categorical perception. This is rather different than the evolutionary approach taken here, where full-grown categories are placed in the representation space and evaluated. Our approach can however validate claims concerning the evolution of colour categories, such as: What is the influence of a non-linguistic or a linguistic task on genetic evolution? Will evolution arrive at a sufficient and adequate categorical repertoire? Will populations that separated before the evolution of colour categories still arrive at the same repertoire?

6.4 Summary

This chapter introduced a genetic operator for colour categories. The simulations on genetic evolution will be used to investigate claims on the innate character of colour categories.

Chapter 7

Results on learning and evolution without language

This chapter and the following present results obtained by running the simulations described in the chapter 5 and 6. The different simulation models are summarised here

Learning without language The experiments in chapter 7 examine in how far a population of individuals exposed to a set of situations containing colour stimuli can acquire a repertoire of colour categories and in how far this repertoire is shared between the individuals. There are two alternative approaches.

1. **Individual learning**
   Each agent individually creates and adapts its categorical repertoire according to a learning mechanism. The resulting category set depends on ecological and perceptual biases.
2. **Genetic evolution**
   An individual's colour categories are encoded as colour genes and are changed through genetic mutation of offspring of the fittest individuals. The fitness is measured on a discrimination task.

Learning with language The experiments in chapter 8 study the impact of language (and thus of culture) on the formation of colour categories. For this a linguistic task is forced upon the agents. Again, two alternative approaches exist.

1. **Cultural learning**
   Learning is used both for the formation of categories and for the words associated to categories. Both are shown to have an influence on each other .

2. **Genetic evolution**
   The categories are still genetically evolved, but now the individuals are judged on their capability to communicate colour meaning using words associated with categories. Words are learned during the life time of the individual, and are culturally passed on between generations.

The intentions of this chapter are manifold. The first results presented here have an illustrative character and serve to make the reader familiar with the chromatic data serving as input for the agents and with the dynamics of the simulations. In addition, the different statistical measures are explained with practical examples. Section 7.2 shows how individual learning is able to arrive at a categorical repertoire that adequately discriminates colour. It is shown how the algorithms can cope with change in the environment and what the nature is of the learning categories. Also the agreement between the categories of agents is studied, and we conclude that the learning approach does not contain enough bias for coherence to emerge between the categories of individuals. Section 7.3 discusses the same topics for the genetic evolution of colour categories. In contrast to the learning approach, the categories do become shared between the individuals when using genetic evolution. Section 7.4 compares both approaches, and demonstrates how learning and evolution operate on different timescales. The chapter concludes with a discussion of the results.

7.1 Setting the stage

The simulations are built around the discrimination game paradigm (Steels, 1996a,b). In such as game, a context of objects —in our case colour stimuli— is offered to a single agent; one of the colour stimuli is selected as the topic. The discrimination game requires the agent to identify the topic with a category not used to identify the other stimuli in the context. An agent starts with an empty colour category repertoire. During the games, the agent extends and adapts its category set, either through learning or through genetic evolution. The examples shown in this section are aimed at making the reader familiar with the dynamics of the simulations and with the mode of presentation.

The visual stimuli presented to the agents are colour samples provided in aperture mode; that is, the colour samples are undone from any contextual information. The colours do not have any texture, there is no surrounding colour, they are undone from any cultural significance and they do not belong to any object: they are presented as if viewed through a small aperture in a neutral gray coloured sheet of paper, and we assume a constant chromatic adaptation. Phenomena such as spatial adaptation, local interactions in colour perception, colour adaptation, colour information from

edges, Mach bands and induced colours are ignored[1]. This implies that the categorisation of the colour samples is only affected by the achromatic and chromatic content and by nothing else.

Spectral power distributions as stimuli

In the experiments the colour samples are presented as spectral power distributions. The SPDs give the relative energy for wavelengths in the human visible spectrum. The SPDs used here have been measured using a spectrophotometer, and are sampled in discrete wavelength steps. Three data sets are used:

1. A set containing 1269 samples of the matte finished Munsell chart (Munsell, 1976). Measured with a spectrophotometer, for wavelengths from 380 to 800 nm in 1 nm steps (Parkkinen et al., 1989).

2. A set containing 218 spectral samples of plants and flowers[2] (Parkkinen et al., 1988). The samples have been measured from 400 tot 700 nm in steps of 5 nm.

3. A subset of the Munsell chart measured above, containing the 330 colour chips. 320 of these are the same as the chromatic colour chips used by Lenneberg and Roberts (1956) in their anthropological and linguistic research. These stimuli set consists of 40 equally spaced hue and eight degrees of brightness, all at maximum saturation[3]. The other 10 chips are achromatic, ranged from white, over intermediate greys, to black. Their SPD was not available, but their CIE XYZ values are obtained from (Newhall et al., 1943). This set is only used for displaying category sets, and is therefore not used as input to the agents.

The strength of using SPDs is that no information is lost by a preceding colour conversion. There is still the opportunity to introduce effects such as variation in the spectral sensitivity of individuals or other colour vision deficiencies. An illustration of two spectral power distributions can be seen in figure 7.1.

[1] There are numerous temporal and spatial factors that influence the way we see colour, none of them are taken into account in these experiments. More can be read in (Kaiser and Boynton, 1996, p. 408ff)

[2] These and other spectral power distribution databases can be downloaded from http://cs.joensuu.fi/~spectral/. For more information on the databases used in these specific experimental runs see (Parkkinen et al., 1988, 1989; Lenz et al., 1996).

[3] The exact codes of the chips used by Berlin and Kay (1969) are very hard to come by, if ever needed they can be found in (Hardin and Maffi, 1997) on the first page which comments on the cover illustration.

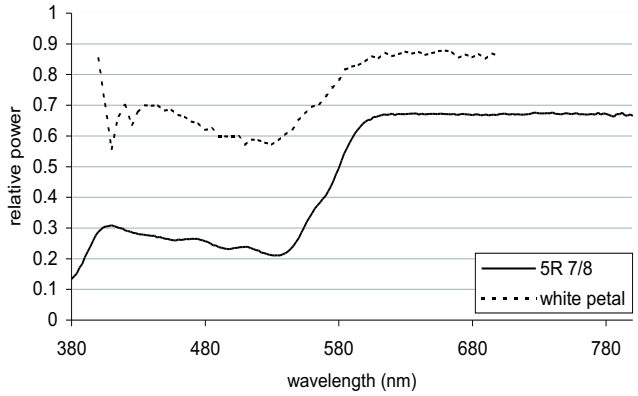


Figure 7.1: Two spectral power distributions of a Munsell chip 5R 7/8 and a whitish flower petal.

The sets contain only a limited sample of all colours distinguishable by humans. The Munsell database for example only contains about 1300 stimuli, while Kaiser and Boynton (1996) estimate the number of distinguishable colours to be about one million[4]. In some experiments we extend the database by subsampling the stimuli. It will be mentioned alongside the results when this procedure is used.

7.2 Individual learning of colour categories

This section describes how colour categories are learned through an adaptive mechanism.

7.2.1 An illustrative experiment

This first experiment illustrates how agents handle a simple context. The context contains only two different shades of saturated colour selected from a set of colours containing eleven chips from the Munsell solid: eight chromatic chips (from each hue symbol the chip with value 5 and maximum chroma) and three non-chromatic chips[5]. These chips correspond to

[4]Kaiser and Boynton arrive at this number by counting the just notable differences between colours. Bornstein (1975, p. 783) mentions that humans can distinguish 150 different *hues* among the visible wavelengths from 400 tot 700nm.

[5]The chips used here are 5 R 5/5, 5 YR 7/10, 5 Y 8.5/10, 5 GY 8/10, 5 BG 7/8, 5 B 5/8, 5 PB 5/10 and 5 P 5/8. For the black, grey and white chip we took 5 R 9/1, 5 R 5/2 and 5 R 2.5/1. These formally belong to the red Munsell hue, but as we have no spectral data

colours English speakers might call red, orange, yellow, greenish-yellow, turquoise, blue, bluish purple, purple, white, grey and black. Figure 7.2 shows the chips plotted in the CIE LAB space.

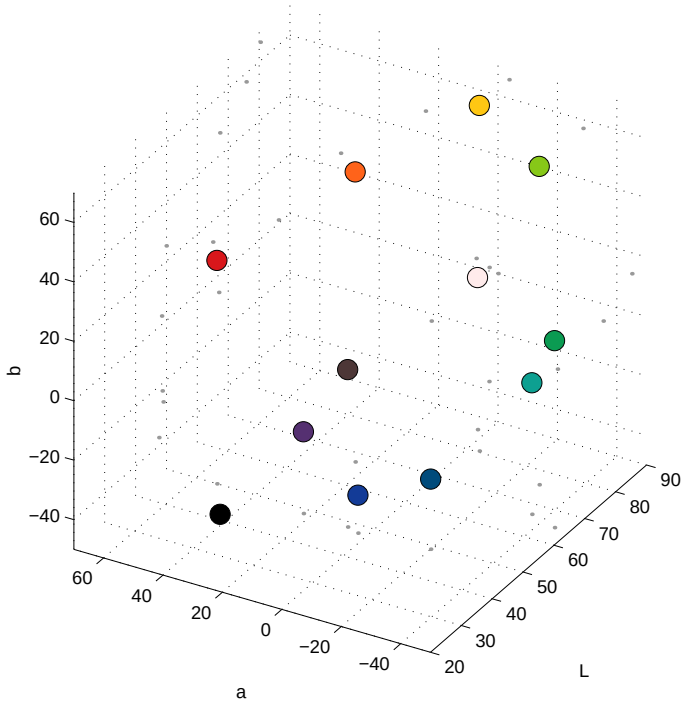


Figure 7.2: The set of Munsell chips used for the illustrative experiment. The chips are given as spectral power distributions, but are plotted here in $L^*a^*b^*$ space. The colour of each dot gives a rendition of how the Munsell chip looks under normal viewing conditions.

The population contains ten agents ($N = 10$) and the context contains two colour stimuli ($|O| = 2$) selected from the fore mentioned eleven possible stimuli. The two stimuli in the context are at a minimal distance of $D = 50$ in the CIE $L^*a^*b^*$ space; this for example means that the context will never contain two greenish colours, but will always contain stimuli that are chromatically well distinct, e.g. a red stimulus and a blue stimulus. Figure 7.3 illustrates how the hearer and speaker might perceive a context. Thousand games are played ($T = 1000$) and no more categories

of non-chromatic Munsell chips we select these since they well resemble white, grey and black. For more details see (Parkkinen et al., 1989).

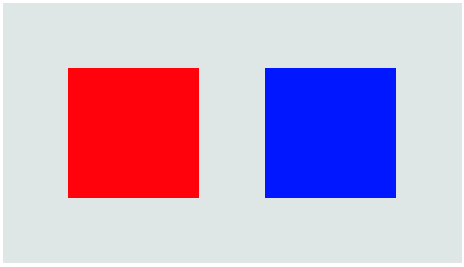


Figure 7.3: An illustration of a context with two stimuli (respectively 5 R 5/14 and 5 PB 5/10) as presented to the speaker and the hearer. One of these two stimuli will be chosen as the topic.

are created if the discriminative success is 95% or higher. The parameters are summarised in table 7.1. All parameters are by default set to these values, if they are different for a particular simulation this will be mentioned alongside the results.

Please note that the results shown here have in effect only an illustrative character. For comprehensive conclusions on the dynamics of the system larger input sets and several runs, initialized with different random seeds, will be needed.

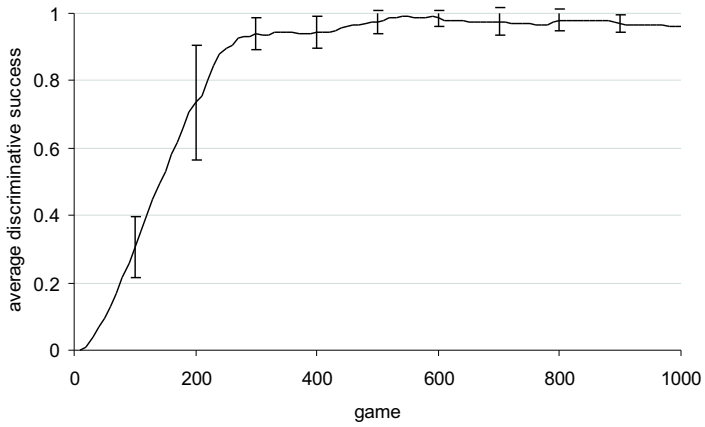


Figure 7.4: Average discriminative success $\overline{DS}$ (individual learning, simple stimuli set, $N = 10$, $|O| = 2$, $D = 50$).

Figure 7.4 shows the average discriminative success $\overline{DS}$ of the population, the discriminative success DS is the running average over the last $T_a = 20$ games of the ratio between the games where the agent was success-

Parameter	Value	Explanation
N	10	Number of agents.
$\|O\|$	2	Number of stimuli in the context.
D	50	Distance between the stimuli of the context.
T	1000	Number of games played.
T_a	20	Number of games over which the running average is taken for computing the DS.
τ_{adapt}	0.95	Threshold for the discriminative success, if the DS is above this threshold, no categories are added anymore, only existing categories are adapted.
σ	5.0	Width of the Gaussian of the locally tuned units.
w_{default}	1.0	Default weight of a locally tuned unit, set when creating a new locally tuned unit.
α	1.0	Learning rate for the adaptive network.
β	0.95	Weight decay for the adaptive network.

Table 7.1: Parameter settings for individual learning.

ful at discriminating the context and the total number of games played. The discriminative success rises quickly, meaning that all agents have created categories which are sufficient to discriminate the topic from the other stimuli in the context; in this specific case, where the context contains only two stimuli, this means that the agent has qualitatively and quantitatively sufficient categories to distinguish one stimulus from another. The task of creating discriminating categories is an individual one, and does not rely on linguistic interaction between agents. Figure 7.5 shows the average number of categories for the population. It rises quickly and stabilises at about five categories per agent; at game 1000 the average number of categories in the population is 5.1 ± 0.3. Note that although the context contains a possible eleven different colour stimuli, only five or six categories are needed to successfully discriminate the context.

The category variance CV illustrates how well the categories of the agents agree. It is based on computing the distance between all categories of all agents; $CV = 0$ if the category sets of all agents in the population are identical[6]. Figure 7.6 shows the category variance CV. The graph shows

[6]Note that it does not make sense to speak of the population average of the category

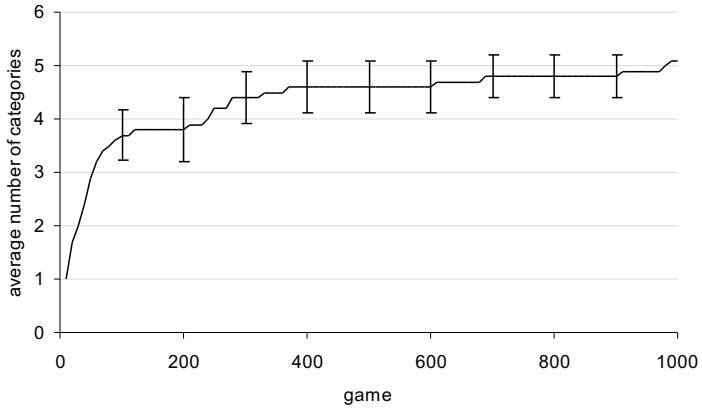


Figure 7.5: Average number of categories (individual learning, simple stimuli set, $N = 10$, $|O| = 2$, $D = 50$).

how categories evolve to resemble each other, this can be explained by the environment and ecology the agents share: all agents are offered the colour stimuli from the same set of 11 stimuli. It should however be noted that the agreement in categories is not absolute, the category variance hovers around 10.2 ± 4.1 at game 1000, meaning that the colour categories of the agents are not at all equal. This is illustrated in figure 7.7, where the location of the maximal reaction of all categories of two agents is plotted. Even though both agents share the same ecology and have near perfect discrimination, their category repertoires are not equal.

7.2.2 Experiment with full Munsell stimuli set

As opposed to the previous experiment, which used a limited set of input stimuli, the next experiment uses the full set of Munsell chips as stimuli: a context now is selected from a possible 1269 stimuli. When the set of possible stimuli is small, the chance of coherent categories arising in the population is intuitively large. For example, if the agents would only be given two stimuli, red and green, each agent would quickly converge on having a category for red and green. However, when the possible set of stimuli is large, it is not apparent which categories the agents will create. This second experiment illustrates the dynamics of such an experiment and explains the relevant measures and results.

The experiment has a population containing 10 agents, the context con-

variance, as the category variance is already taken over the entire population.

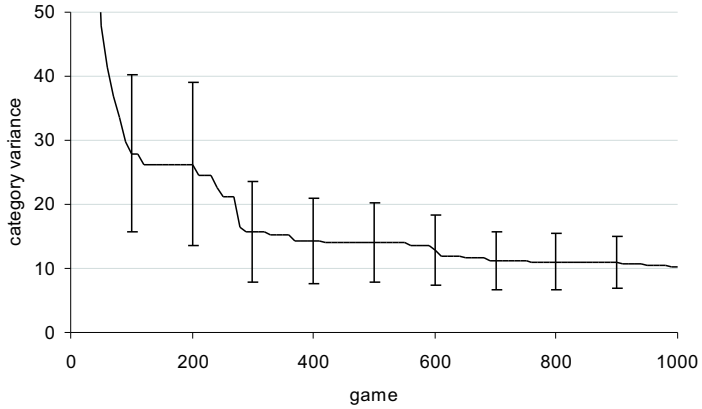
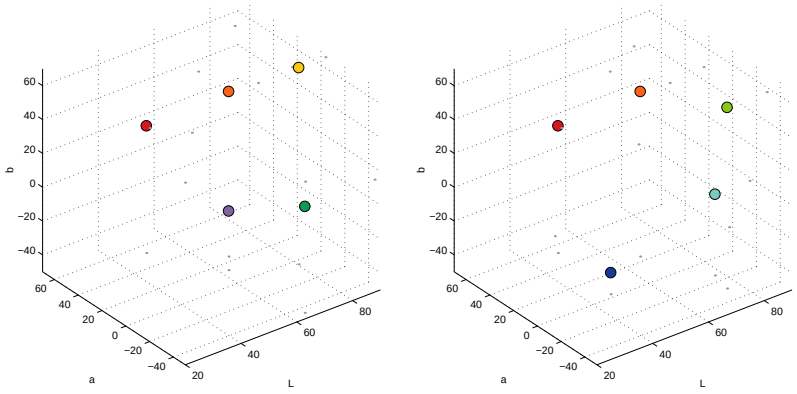


Figure 7.6: The category variance CV (individual learning, simple stimuli set, $N = 10$, $|O| = 2$, $D = 50$).



(a) (b)

Figure 7.7: The maxima of the categories of two agents plotted in $L^*a^*b^*$-space. Even though the agents share the same ecology and have near perfect discrimination scores, their categorical repertoires are not identical.

sists of three colour stimuli ($|O| = 3$) which are at a minimum distance of $D = 50$ in $L^*a^*b^*$-space. The running average for the measures is taken over $T_a = 20$ games. The parameters concerning the adaptive network are identical to the ones introduced in 7.2.1.

The Munsell chip set

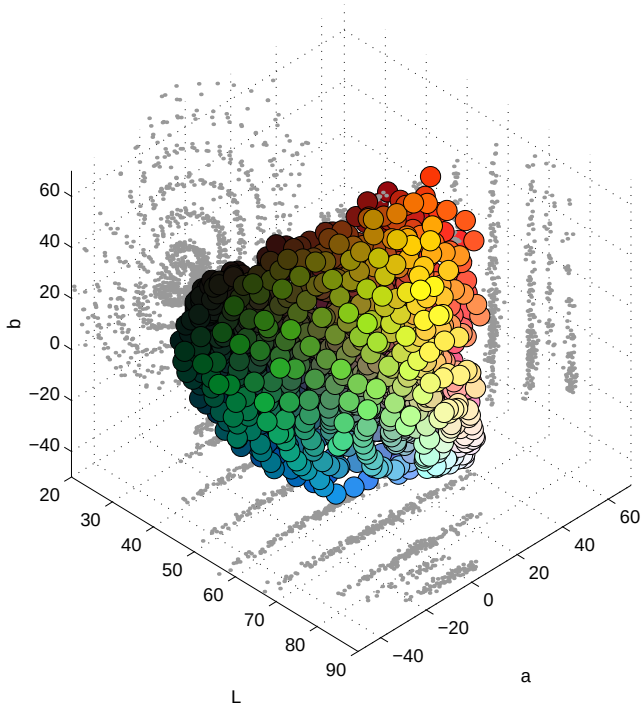
In the experiments the input set for the agents is chosen from a collection of 1269 Munsell chips (Munsell, 1976) measured using a photospectrometer (Parkkinen et al., 1988, 1989). The Munsell chips are measured from the matte collection of the Munsell company. Figure 7.8 shows these chips plotted in $L^*a^*b^*$-space[7]. The Munsell solid is designed to be a systematically organised colour space based on subjective judgements on the distance between the colour chips, no controlled perceptual experiments were used to compile the Munsell colour space (Nickerson, 1940). At the same time the $L^*a^*b^*$-space has been devised by the CIE to be perceptually equidistant. Although the $L^*a^*b^*$-space has been known to behave better in certain regions of the gamut than in others (the figure clearly shows how the blue chips appear to be closer together, while the red and green chips are more spaced out), considering the categorisation mechanism the agents use (categories are represented by a network of weighted exemplars on which this effect has little influence) it is an adequate space for our purposes. The limits of the Munsell solid as measured by Parkkinen et al. (1988) in the $L^*a^*b^*$-space are $L^*_{min} = 24.91$, $L^*_{max} = 87.68$, $a^*_{min} = -52.63$, $a^*_{max} = 53.77$, $b^*_{min} = -42.23$ and $b^*_{max} = 78.28$.

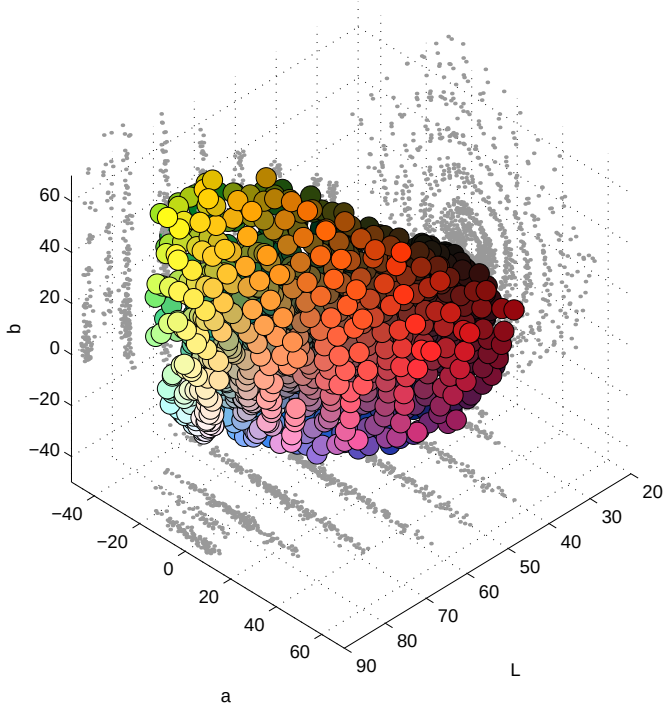
Results

Figure 7.9 shows the average discriminative success $\overline{DS}$. The agents rapidly achieve near perfect discrimination, showing that each has sufficient and adequate categories to discriminate the topic from the context. At 1000 games the population has a discriminative success of 0.96 ± 0.04. Figure 7.10 shows the average number of categories for the population, at game 1000 the agents have on average 9.7 ± 1.4 categories. Again note how the agents need only a limited set of categories to discriminate about 1300 different stimuli.

The category coherence CV, shown in figure 7.11 shows how well the categories of the agents agree, it is based on the distance between all the categories of all agents. If categories are more similar, CV will decrease. The category variance decreases exponentially, and stabilises at 8.3 ± 1.4. If all the agents would have identical categories, the category variance would be zero. The fact that the category variance is still quite high (compared to later experiments) is the result of the discrimination game being an individual learning method: all agents arrive at a categorical repertoire in their own way and there no pressure to have shared category sets. The only

[7]The colours shown in the figures are a rendition of the Munsell chips obtained by converting the spectral power distribution to RGB values for display on an average device. The colour impression might differ depending on the device or paper and the ambient conditions under which you are watching the figures at this very moment.


b
60
40
20
0
−20
−40
20
30
40
50
60
70
80
90
L
60
40
20
0
−20
−40
a


b
60
40
20
0
−20
−40
−40
−20
0
20
40
60
a
20
30
40
50
60
70
80
90
L

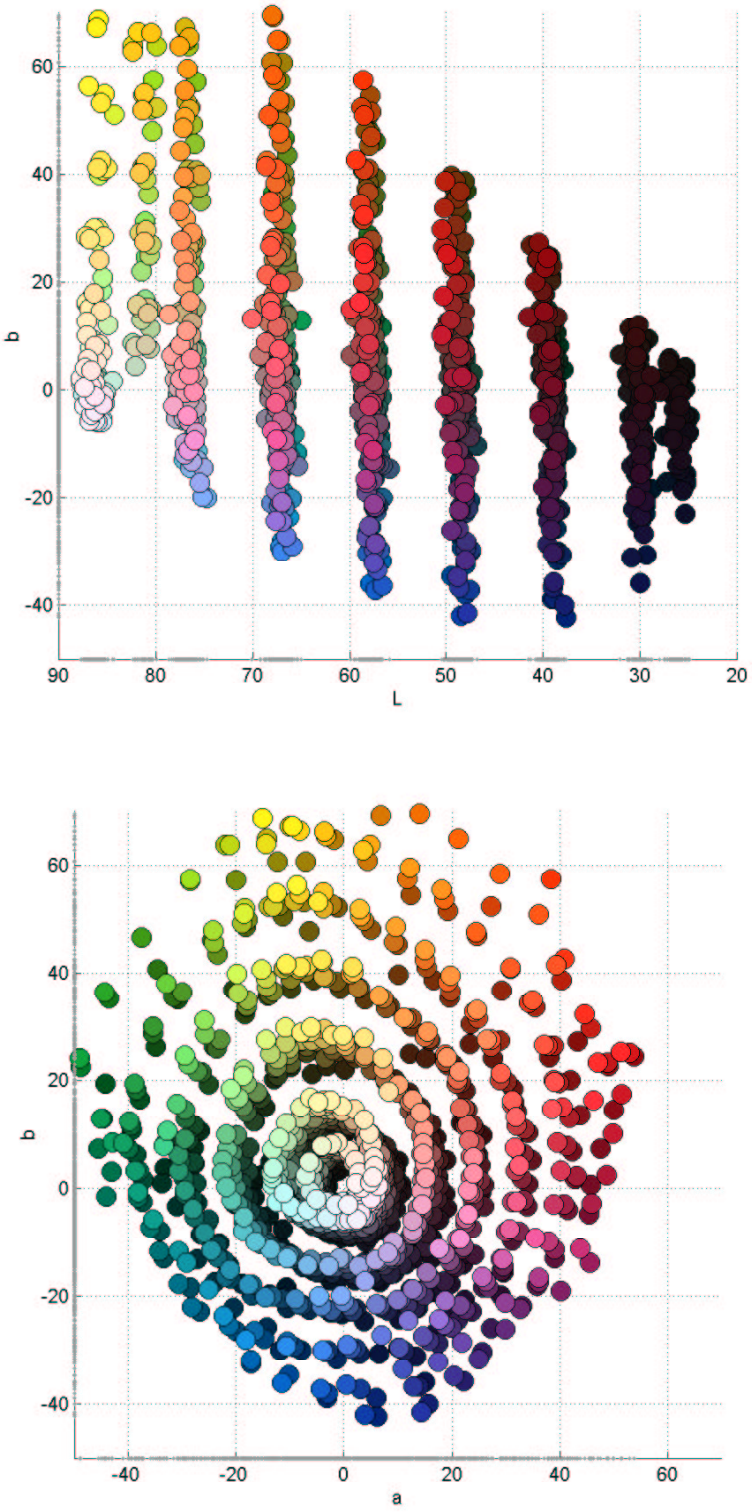


Figure 7.8: The 1269 Munsell chips from the Munsell matte collection plotted in CIE $L^*a^*b^*$-space

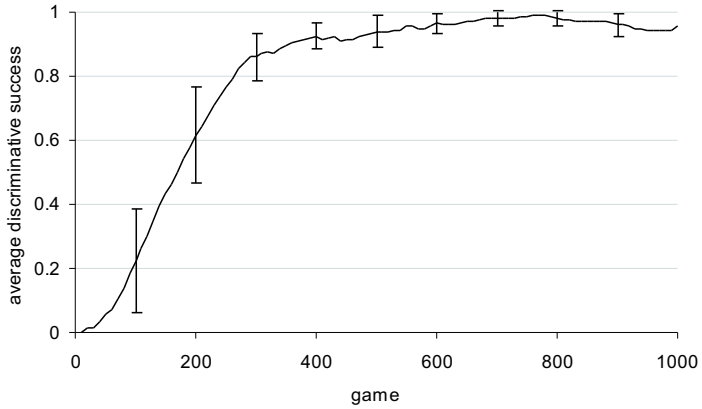


Figure 7.9: Average discriminative success $\overline{DS}$ (individual learning, full Munsel stimuli set, $N = 10$, $|O| = 3$, $D = 50$).

pressure present is ecological pressure, which forces the agents's categories to have a least a certain degree of similarity.

7.2.3 Changing the environment

It needs to be shown how the adaptive model copes with changes in the environment. A population of 10 agents is given a context of four stimuli chosen from a total of seven stimuli, corresponding to red, yellow, green, blue, purple, black and white[8]. The context contains $|O| = 3$ stimuli with a minimum distance of $D = 50$ from each other. After a repertoire has been established, the environment is changed by adding four more colours to choose a context from, these colours are composite (in between the previously mentioned fundamental colours). They are orange, yellowish green (or lime), turquoise and bluish purple[9]. As the added stimuli are mixtures of the colours offered in the first series of stimuli, the agents's established repertoire will not be able to discriminate these colours and will be forced to create new categories or to adapt their existing categories. Figure 7.12 shows what happens: at game 50 the environment is changed, the average discriminative success drops, but the agents react by creating new categories and by adapting their existing categories. At game 50 the agents have on average 6.2 ± 0.5 categories each; at game 100, after re-establishing a discriminative success of 100%, they have on average 7.1 ± 0.4 categories.

[8]The Munsell codes of the stimuli are 5 R 5/14, 5 Y 8.5/10, 5 G 7/10, 5 B 5/8, 5 P 5/8, 5 R 9/15, R 5/2.

[9]The four added stimuli are 5 YR 7/10, 5 GY 8/10, 5 BG 7/8 and 5 PB 5/10.

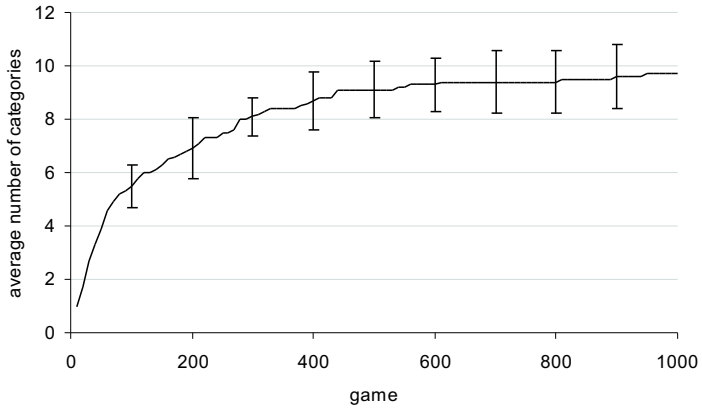


Figure 7.10: Average number of categories (individual learning, full Munsel stimuli set, $N = 10$, $|O| = 3$, $D = 50$).

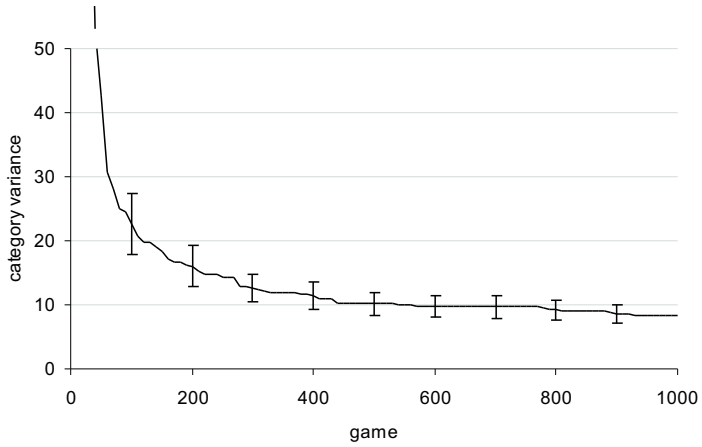


Figure 7.11: The category variance CV (individual learning, full Munsel stimuli set, $N = 10$, $|O| = 3$, $D = 50$).

Note that the population is not changed, and that the agents do not interact with each other: the resulting behaviour is purely due to individual learning.

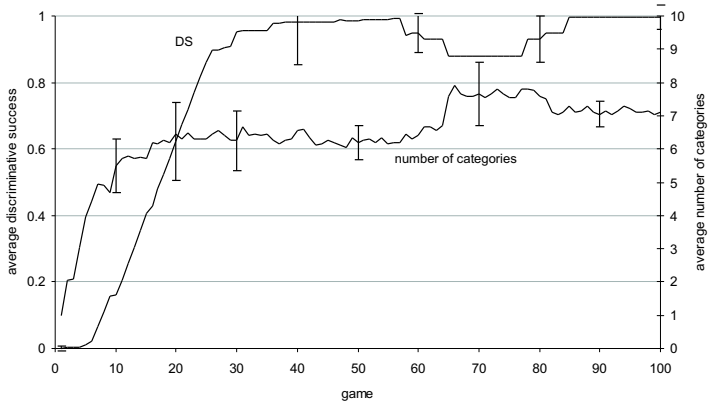
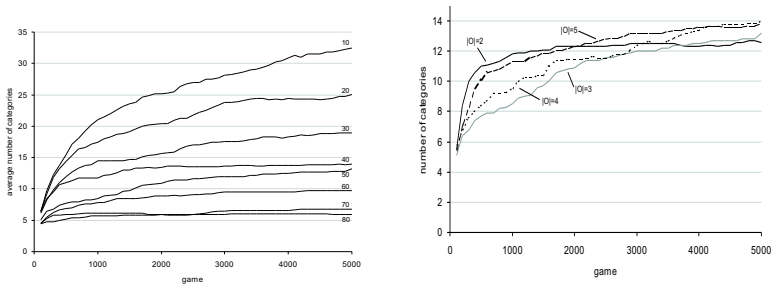


Figure 7.12: The average discriminative success and number of categories for a population of 10 agents with *adapted* categories experiencing a change in the environment.

7.2.4 Pressure for creating categories

The size of the categorical repertoire of the agents is influenced by the task and the environment. If only red and blue have to be discriminated from each other, one can do so with only two categories. But if we need to discriminate more colours, we need more categories to discriminate one specific stimulus from the others. Also, if the stimuli are more resembling, for example we have to discriminate ruby red from blood red, we will be forced to create more categories. These two factors, the number of stimuli present in one context and the perceptual distance between the stimuli, make up the *complexity* of the environment. Figure 7.13a shows the influence of the perceptual distance on the number of categories created by the agents. The context consists of $|O| = 3$ stimuli selected from the full Munsell set. Each plot shows the average number of categories of 10 agents for different distances between the stimuli, ranging from $D = 10$ to $D = 80$. The plot shows that stimuli that are perceptually closer force the agents to create more categories to discriminate the stimuli. Figure 7.13b shows the average number of categories, for the same experimental conditions, but now the distance between the stimuli in the context is kept constant at $D = 50$ and the size of the context is varied from 2 to 5 stimuli. The plot shows that the size of the context does not relevantly influences the num-

ber of categories the agents need to create. For 2 stimuli the agents have at game 5000 on average 12.6 ± 2.2 categories, for 3 stimuli 13.2 ± 1.0, for 4 stimuli 14.0 ± 1.5 and for 5 stimuli 13.8 ± 2.2. So no matter how many stimuli are shown simultaneously, it is the resemblance between the stimuli that forces the agents to create more categories. Category creation is driven by richness of the environmental stimulation.



(a) (b)

Figure 7.13: (a) The average number of categories for a context with three stimuli at different perceptual distances ranging from $D = 10$ to $D = 80$. (b) The average number of categories for a context with distance $D = 40$ between the stimuli at different sizes of the ranging from $|O| = 2$ to $|O| = 5$. The categories are individually learned.

7.2.5 The nature of the categories

It is interesting to look at the nature of the colour categories created by the agents. As discrimination is an individual task, there is no reason to expect that the category sets of two agents will be identical. The only source of a possible similarity is the environment, and as the environment can be highly complex, consisting of about 1300 different stimuli, there exist numerous category constellations with equal discriminative power. Figure 7.14 shows two category sets of two agents picked from the simulation in section 7.2.2 at game 1000. The categories are plotted on a two-dimensional chart consisting of 320 saturated Munsell chips and 10 non-chromatic chips ranging from white, over grey to black. The same chart has been used in various anthropological studies and was introduced by Lenneberg and Roberts (1956) and extended with non-chromatic chips by Berlin and Kay (1969).

The maximum of a category is represented by a white circle, and the extent of the category is plotted with colour coded smaller circles (the size

of the circles denotes the strength of the category at that location). The plot is produced by offering each of the 330 colour chips to all categories, the category with the highest reaction is the one for which a circle is plotted on the particular chip, the colour of the circle is the colour to which the category reacts best. The extents of the categories fill up the entire Munsell plot, meaning that every chip will be classified as belonging to a category: no chip will be left unclassified. This is an property of the classification algorithm, which uses a winner-takes all strategy.

As the plot is actually a projection of categories from a three-dimensional space onto a two-dimensional visualisation, some information is lost. It might not truthfully represent the actual category set, as categories for desaturated colours are forced onto the two-dimensional projection containing only saturated colours. This is also a point of criticism on anthropological studies using the Munsell chart for obtaining information on the nature of colour categories (Saunders and van Brakel, 1997, p. 175). Nevertheless, this representation of colour categories still has value in the context of our account, as it is only meant to illustrate how category sets of different agents relate to each other. As can be seen, both agents arrive at category constellations that are different from each other, as was expected.

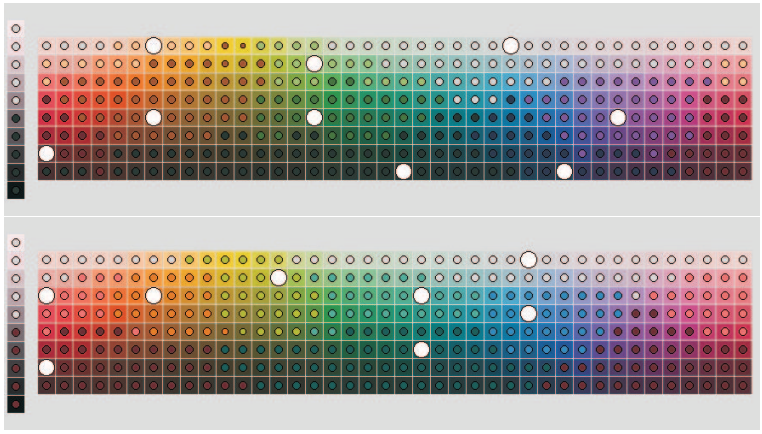


Figure 7.14: The maxima (white circle) and extent (colour coding) of categories of two agents picked from a population of 10. The chart consists of 320 saturated Munsell chips and ten achromatic chips.

7.2.6 Agreement across populations

As might be expected from the results discussed in figure 7.11, where the agents in a single population do not reach coherent colour categories, the agents from different populations will not reach coherent colour categories

as well. The inter-population agreement between agents's category sets can be computed with the inter-population category variance measure, see (5.20). Five simulations have been run with each $N = 10$ agents and $T = 1000$ discrimination games, but each time with a different random seed. The random seed influences the selection of stimuli for the context, thus influencing the environment to which the agent are exposed. The context is in each of the five simulation selected from the same set of about 1300 Munsell chips. Table 7.2 shows the resulting intra-population variance. Note that within a population the variance still is somewhat lower than across populations, this is due to the shared ecological circumstances in a population.

CV'	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}_3$	$\mathcal{A}_4$	$\mathcal{A}_5$
$\mathcal{A}_1$	9.29				
$\mathcal{A}_2$	10.14	9.38			
$\mathcal{A}_3$	10.62	10.51	9.62		
$\mathcal{A}_4$	10.84	11.25	10.94	9.22	
$\mathcal{A}_5$	10.89	11.14	10.31	11.21	9.83

Table 7.2: Intra-population category variance CV' of 5 populations of which the categories have been learned under identical experimental settings, only the random seed is different between the populations.

7.3 Evolving colour categories

This section describes results obtained with the evolutionary approach described in chapter 6. The task on which the agents are evaluated is again a discrimination task; linguistic interaction is not involved at this moment as we are only interested in the ability of the agents to create sufficient colour categories for discriminating spectral content in their environment.

7.3.1 An illustrative experiment

In a first experiment a population of agents is evolved using the same environment as in 7.2.1: for each discrimination game a context of two stimuli is compiled from an environment of eleven colour stimuli. The two stimuli in the context $|O| = 2$ are at a minimum Euclidean distance of $D = 50$ in the $L^*a^*b^*$-space. The population contains $N = 20$ agents. The probability of selecting a mutation operator is $p_{\text{mut}} = 0.1$ (see section 6.1.1). The fitness of the agents is computed by playing a series of $T_f = 50$ discrimination games. The parameter settings are summarised in table 7.3.

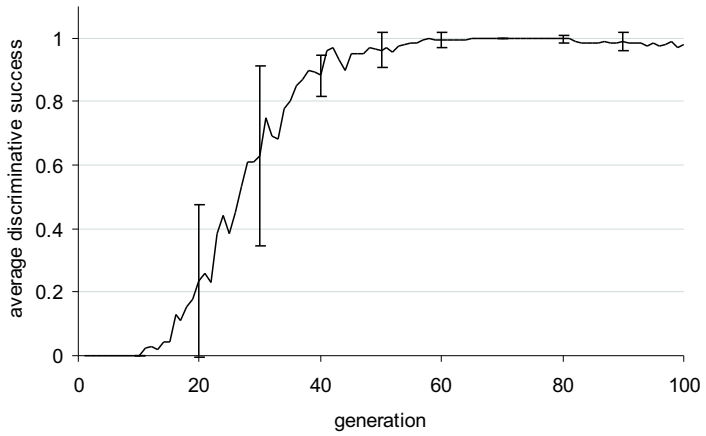


Figure 7.15: Average discriminative success $\overline{DS}$ (genetic evolution, simple stimuli set, $N = 10$, $|O| = 2$, $D = 50$).

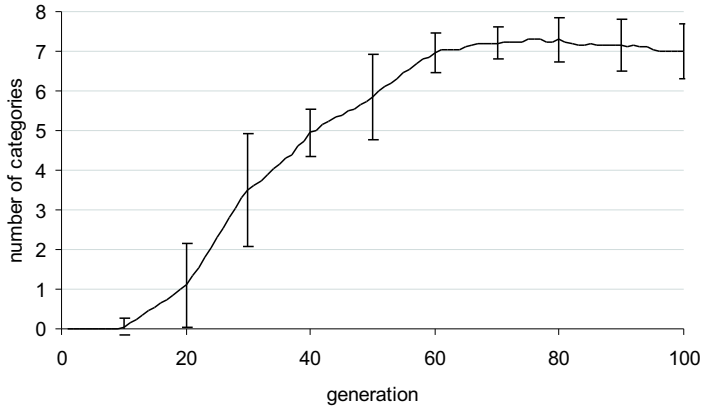


Figure 7.16: Average number of categories (genetic evolution, simple stimuli set, $N = 10$, $|O| = 2$, $D = 50$).

Parameter	Value	Explanation
N	20	Number of agents.
$\|O\|$	2	Number of stimuli in the context.
D	50	Distance between the stimuli of the context.
T	100	Number of generations.
T_f	50	Number of discrimination games on which the fitness if calculated.
σ	5.0	Width of the Gaussian of the locally tuned units.
w_{default}	1.0	Default weight of locally tuned unit, set when creating a new locally tuned unit.
p_{mut}	0.1	Probability of mutation, see 6.1.1.
σ_{mut}	5.0	Standard deviation of Gaussian random distribution, see 6.1.1.

Table 7.3: Parameter settings for genetic evolution.

The average discriminative success $\overline{DS}$ is shown in figure 7.15. The evolutionary strategy is clearly able to come up with a sufficient and adequate set of categories for discriminating the context. Figure 7.16 shows the average number of categories for the population, at generation 100 this is 7.0 ± 0.7. Around generation 70, the number of categories stabilises, individuals are selected for having about 7 categories for a context chosen from a total of 11 different stimuli. The category variance CV is shown in figure 7.17, it shows how much categories vary between the agents in the population. Compared to the category variance of a similar experiment where the agents use a learning strategy (figure 7.6), the category variance is much lower for the evolutionary strategy. This is due to the nature of evolution: successful individuals are allowed to spread their "colour genes" through the population, which reduces variance. Eventually all individuals in the population will be the offspring of a few highly successful individuals. The little variance still present in the population is due to mutations.

7.3.2 Experiment with full Munsell stimuli set

We now demonstrate the genetic evolution of colour categories when the context is chosen from the full set of 1269 Munsell chips. The context contains $|O| = 3$ stimuli with a distance of $D = 50$ in between the stimuli. The population contains $N = 20$ agents, and runs for $T = 200$ generations. At the end of each generation each agent is evaluated by playing $T_f = 50$ discrimination games.

Figure 7.18 shows the average discriminative success $\overline{DS}$ of the popula-

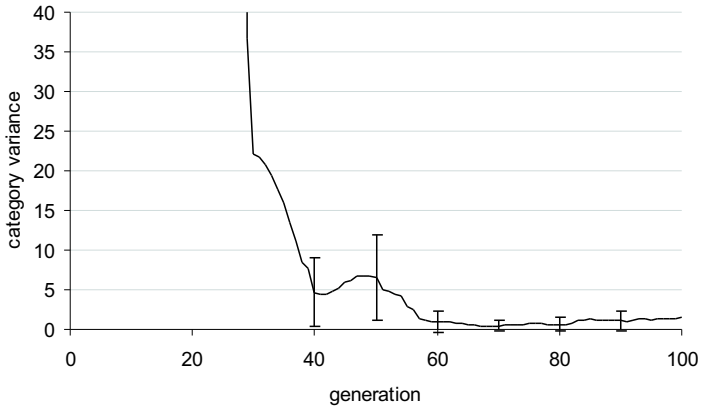


Figure 7.17: Category variance CV (genetic evolution, simple stimuli set, $N = 10$, $|O| = 2$, $D = 50$).

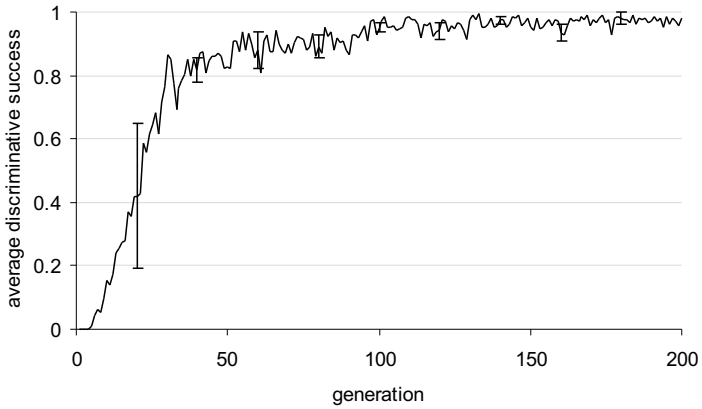


Figure 7.18: Average discriminative success $\overline{DS}$ (genetic evolution, full Munsell stimuli set, $N = 10$, $|O| = 3$, $D = 50$).

tion. The population has no problem to evolve a categorical repertoire near perfect for discrimination. Figure 7.19 shows the average number of categories, one can see that the plot does not level off, there seems to be nothing to stop the agents from evolving more categories with each next generation. The why and how of this is explored in section 7.4.1 on page 120. The category variance is shown in 7.19, again we see that genetic evolution yields

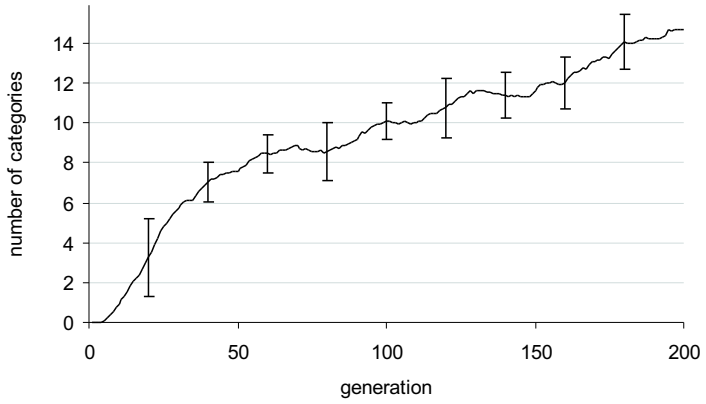


Figure 7.19: Average number of categories (genetic evolution, full Munsell stimuli set, $N = 10$, $|O| = 3$, $D = 50$).

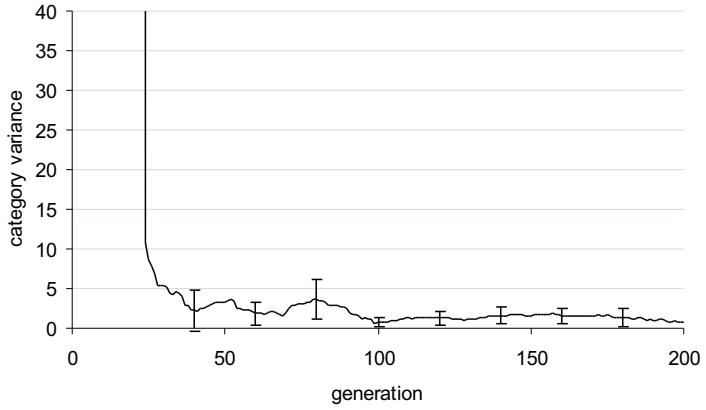


Figure 7.20: Category variance CV (genetic evolution, full Munsell stimuli set, $N = 10$, $|O| = 3$, $D = 50$).

a lower variance compared to the adaptive approach (see figure 7.11). As we have mentioned before, this is due to the nature of genetic evolution: fit individuals tend to have more offspring in the population at the expense of underperforming individuals, thus spreading their colour genes in the population. After a while, the entire population will have more or less identical colour categories, besides from some variation introduced by the

mutation operator.

7.3.3 Changing the environment

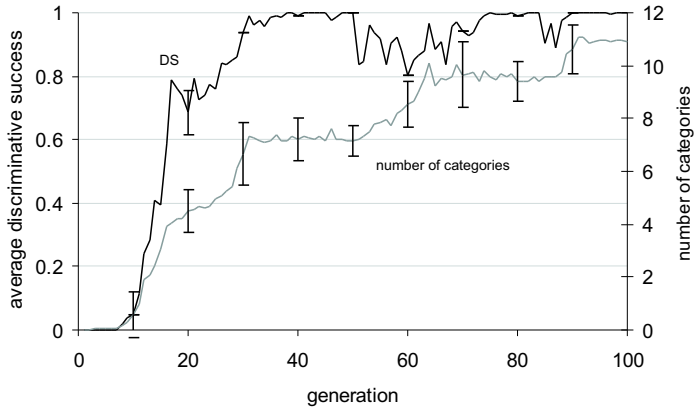


Figure 7.21: The average discriminative success and number of categories for a population of 20 agents with *evolved* categories experiencing a change in the environment.

Similar to 7.2.3 for the adaptive approach, we need to show how genetic evolution copes with a changing environment. Figure 7.21 shows the average discriminative success $\overline{DS}$ and the average number of categories for a population of 20 agents. The context contains $|O| = 3$ stimuli with a minimum distance of $D = 50$ from each other. The fitness of each agent is again evaluated by playing $T_f = 50$ discrimination games. The simulation runs for 100 generations. Before the fiftieth generation the context is selected from a set of seven stimuli, after that the context is selected from a set of eleven stimuli. This increases the environmental complexity and poses a more difficult discrimination task for the agents. As the plot shows, the discriminative success rises and after a while the agents reach perfect discrimination (1.0 ± 0.0 at generation 50), after generation 50 the complexity is increased and the agents's discriminative success is destabilised. However, after about 25 generations the category sets are changed and extended so that the discriminative success again reaches 100%.

The same graph demonstrates a noteworthy property of innate functionality. In generation 50 the environment suddenly changes, as new stimuli are introduced causing a temporary decrease in the fitness of the population. But again, in generation 87, the discriminative performance slightly decreases. This time because the evaluation task for which 50 discrimina-

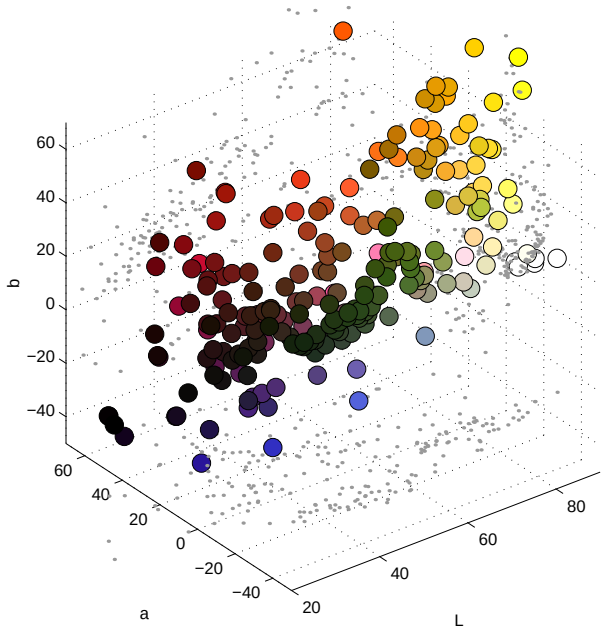
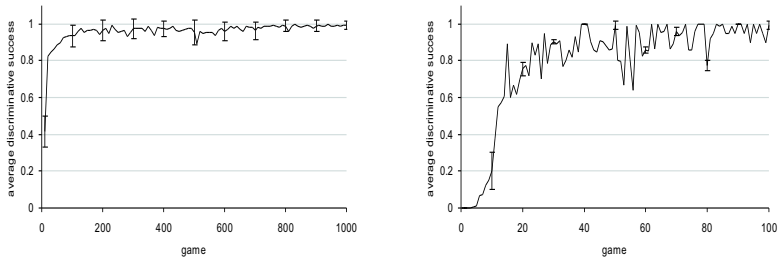


Figure 7.22: Natural colour stimuli plotted in the CIE LAB space.

tion games are played, is played on randomly selected contexts. In this particular generation the random construction of the context has repeatedly delivered a context that could not be discriminated by the agents's current category set. This shows that innate functionality is "brittle". It can only cope with unexpected changing environmental conditions by genetically adapting the functionality over generations. If the environmental change is too abrupt and significantly influences the individual's reproductive chances, genetic evolution will not be able to cope at all and the entire population might die out. As Darwin elegantly phrased: "A grain in the balance will determine which individual shall live and which shall die…" (1859, p. 352).

To again illustrate how both agents can adapt from one environment to another we first feed to agents stimuli from the natural colour samples (see 7.1) of which the $L^*a^*b^*$-values are plotted in figure 7.22. In the middle of the simulation we switch the environment to the full Munsell set, which contains more colour stimuli. Neither individual learning ($N = 10$, $|O| = 4$, $D = 40$) nor genetic evolution ($N = 20$, $|O| = 4$, $D = 40$) seems to have trouble with quickly adapting to the changing environment (figure 7.23).



(a) (b)

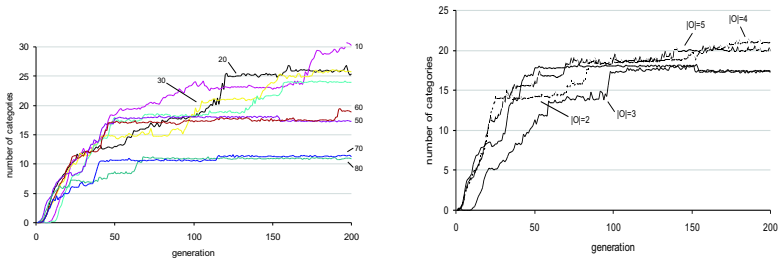
Figure 7.23: The average discriminative success if the context is first selected from the natural stimuli set, and after $T/2$ games or generations is selected from the full Munsell stimuli set. For (a) Individual learning (b) genetic evolution.

7.3.4 Pressure for creating categories

The influence of the environment on the size of the categorical repertoire can be seen in figure 7.24. In the first plot the context consists of 3 stimuli chosen from the full Munsell set, but the distance between the stimuli is swept from $D = 10$ to $D = 80$. At generation 200 all agents have an average discriminative success higher than 99%. The plot shows that the agents need to create more categories if the distances are closer together. There are two important differences when comparing with the same experiment for learned categories. First, the genetic evolution creates more categories than learning does. On average the evolutionary approach has 48% more categories at the end of the simulation. Second, the influence of the distance between the stimuli is not as clear cut as with the learning approach. Both observations can be explained by taking a closer look at the conditions under which categories are added. With the evolutionary mechanism, a category is added during the mutation step. As an additional category can only improve the discriminative performance and never hamper it, there is no inhibition on the creation of new categories. If the discriminative success is near perfect, the creation and deletion of categories will form an equilibrium: the rate at which categories are added will be approximately equal to rate at which they are deleted. So, there is no disadvantage to having a large set of categories. This is further explored in 7.4.1.

7.3.5 Nature of the categories

In the case where categories are learned, there is no reason to expect that categorical repertoires of two agents will be identical. In the case of ge-



(a) (b)

Figure 7.24: (a) The average number of categories for a context with three stimuli at different perceptual distances ranging from $D = 10$ to $D = 80$. (b) The average number of categories for a context with distance $D = 40$ between the stimuli at different sizes of the ranging from $|O| = 2$ to $|O| = 5$. The categories are genetically evolved.

netic evolution of categories, the repertoires will be as good as identical as successful repertoires are passed on from generation to generation, with only slight variations due to mutation. Figure 7.25 illustrates the categorical repertoire of two agents at generation 200 of a population of 20 agents from the experiment in 7.3.2. There a strong resemblance between the two repertoires (which was already noticeable from the low category variance in figure 7.20), the two agents have had a common ancestor and only differ because their genetic information has undergone some mutations since they both split off.

7.3.6 Agreement across populations

Within a population the colour categories are very similar, as demonstrated by the category variance measure in 7.3.2. But can the evolutionary model achieve coherence *between* different populations, something which the adaptive model could not? Would the evolutionary manner of creating category sets provide some new constraint leading different populations to similar category sets. The answer is no.

The inter-population agreement between agents's category sets can be computed with (5.20). Five simulations have been run with each 20 agents for 100 generations, each population with a different random seed. As mentioned before, the random seed influences the selection of stimuli for the context, thus influencing the environment to which the agents are exposed. The context is selected from the full set of Munsell chips. Table 7.4 shows the resulting intra-population category variance. As expected the category

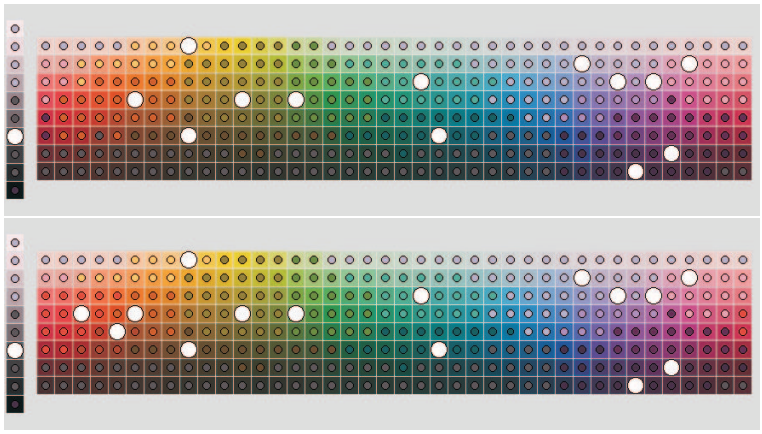


Figure 7.25: The maxima (white circle) and extent (colour coding) of categories of two agents picked from a population of 10. The chart consists of 320 saturated Munsell chips and ten achromatic chips.

variance within a population is low. But the variance between two different populations is much higher, from which we can conclude that the evolutionary approach does not provide constraints leading to high coherence *between* populations.

When comparing the category variance between populations for the adaptive approach and the evolutionary approach (table 7.2 and 7.4), one will notice that populations in the evolutionary approach have a smaller variance between their category sets. This is due to the nature of the mutations implemented in the model: remember that a category is represented by an adaptive network which consists of a collection of hidden units. The mutation operator for creating new hidden units creates these units in the proximity of the centroid of the category (see 6.1.1 on page 88), therefore the "spread" of a category is lower than with the adaptive approach, where hidden units of one category can be located anywhere in the representation space. This phenomena causes the category variance between populations to be somewhat lower for the evolutionary model. As the location where new hidden units are created during the mutation step is merely depending on the parameter σ_{mut} of the model (by default set to $\sigma_{\mathrm{mut}} = 5.0$), increasing σ_{mut} will result in a higher inter-population category variance.

7.4 Comparing adaptation and evolution

Figure 7.26 shows the discriminative success of five different simulation runs for an agent for which the categories are *adapted*. The context of the

CV'	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}_3$	$\mathcal{A}_4$	$\mathcal{A}_5$
$\mathcal{A}_1$	0.40				
$\mathcal{A}_2$	4.91	0.40			
$\mathcal{A}_3$	3.98	5.75	0.05		
$\mathcal{A}_4$	3.67	4.54	4.64	0.20	
$\mathcal{A}_5$	5.60	6.26	6.10	5.55	0.27

Table 7.4: Intra-population category variance CV' of 5 populations of which the categories have been *evolved* under identical experimental settings, only the random seed is different between the populations.

agents is chosen from the complete Munsell set as described in 7.2.2. The context contains four stimuli ($|O| = 4$) at a minimum Euclidean distance of $D = 40$, all other parameters are identical to the settings used in 7.2.2. The only difference is the way in which the discriminative success is computed: after each adaptation (when creating or adapting a colour category), the discriminative success is computed on a series of 50 discrimination games. This to allow a faithful comparison with the evolutionary experiment described below. As demonstrated before, the agents reach a high discriminative success in a short number of adaptations. As shown in figure 7.26, the number of categories levels off as the discriminative success is high enough (same figure).

Figure 7.27 shows the discriminative success of 50 agents evolved over 50 generations; the results are averaged over five different simulations. During each generation the fitness of all the agents is evaluated by playing $T_f = 50$ discrimination games. The games are identical for each agent, to allow fair selection. The evaluation step is followed by a selection and mutation step: the fittest agents (50% of the population) can continue to the next generation, while the other agents are removed from the population and are replaced by mutations of the fittest agents.

The number of categories created by the evolutionary approach is unbounded (see figure 7.27). The reason for this is that there are no restraining dynamics for the number of categories: the mutation step is free to generate new categories at any time, and the selection step does not select for agents showing "categorical economy".

7.4.1 Unbounded number of categories

As shown in figure 7.27, the number of categories is unbounded in the evolutionary approach due to the fact that the evolution only selects on the discriminative capacities of the category set and does not punish the size of the set. The fact that there is no limit to the number of categories is of course unrealistic and several solutions can be found to limit the adding of

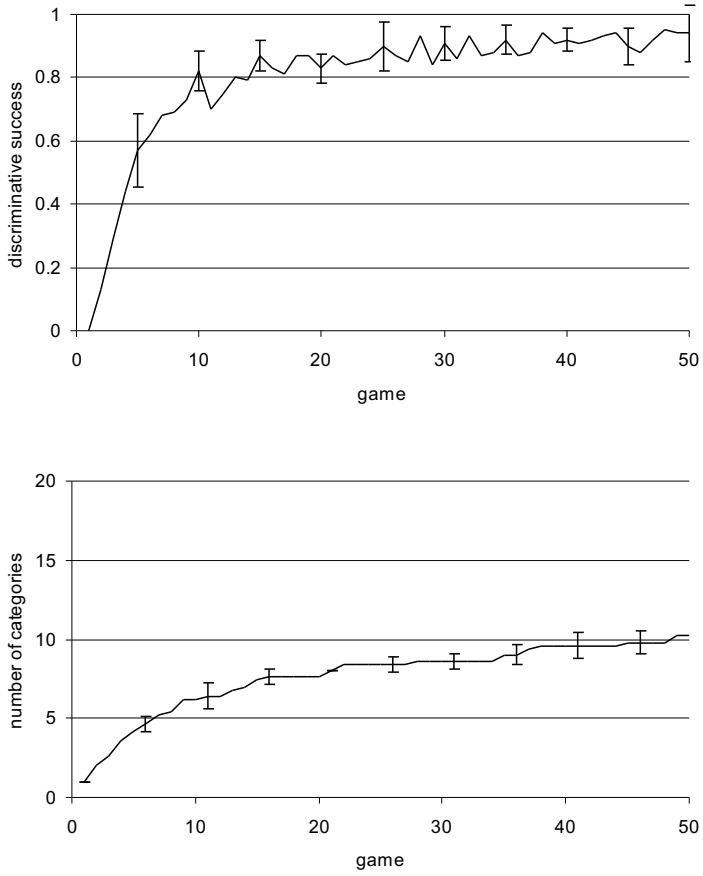


Figure 7.26: The discriminative success of one agent (for five different simulation runs). After each adaptation, the discriminative success is computed on a series of 50 discrimination games; so during each game, the agent's category set is adapted and then evaluated by playing 50 discrimination games (top). The bottom figure shows the average number of categories.

categories. One solution would be to impose an upper limit on the number of categories; unfortunately this upper limit would be an arbitrary number and it seems unlikely that there is a hard boundary to the number of colour categories one can have. A better solution might be to have an evolutionary cost proportional to the number of colour categories, a high number of colour categories comes at an associated cost, and the evolutionary benefit of having many colour categories will eventually be less than the costs. When the cost of creating a category and the benefit of owning an additional category is known, the upper limit of the number of categories could be calculated. But the benefit of having an additional category depends on the complexity and ecology of the environment, and is not easily

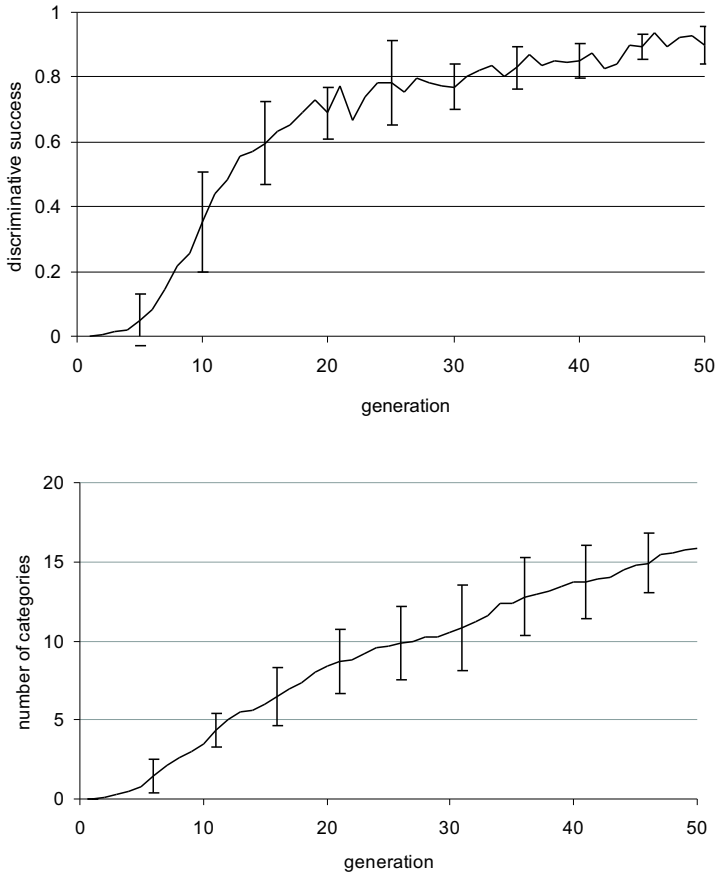


Figure 7.27: The averaged discriminative success of 50 agents for five different simulation runs (top) and the number of categories (bottom).

calculated. Yet another solution would be to make the mutation rate of the individuals a function of their fitness[10]. In this way the individuals would have a high probability of mutation in their offspring when their fitness is low, and would have a low probability when their fitness is high or even perfect. Figure 7.28 illustrates this by plotting the average number of categories for two genetic evolution *with* the mutation probability depending on the fitness and *without*. The population contains 50 agents and runs for 50 generations, the individuals are evaluated by playing 50 discrimination games. The mutation probability p_{mut} is made relative to the fitness by the following equation, in which f is the fitness of the individual and ϵ is a small value, set to 0.1, to ensure that even 100% fit individuals

[10]Granted, this mechanism seems doubtful when applied to the evolution of innate colour categories, nevertheless it has been observed in cell mutations (Cairns et al., 1988).

have a chance at mutation (this is important to keep some variation in the population). As can be seen in the plot, making the mutation probability dependant on the fitness keeps the growth of categories under control.

$$p_{\mathrm{mut}} = (1 - f) + \epsilon$$

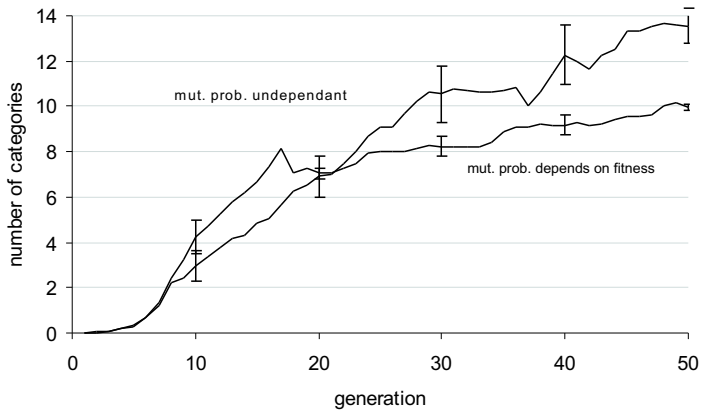


Figure 7.28: The number of categories of a population of which the categories have been genetically evolved. If the mutation probability of an individual is inverse proportional to the fitness of the individual, the number of categories is lower than when the mutation rate is not influenced by the fitness.

In the next chapter we will demonstrate that resorting to such a drastic measures is unnecessary, as the need to communicate provides enough inhibition on the growth of the number of categories.

7.5 Discussion

The important difference between the adaptive approach and the evolutionary approach lies in the fact that in the former, the agents develop categories during their "lifetime", whereas the latter evolve categories through a generational selection and differentiation. Though very much a platitude, the nature-nurture dichotomy is plainly demonstrated here. We should point out the limited validity of the simulations to make any conclusions on the origins of human colour categories. The models and simulations used here suffer from reductionism and are isolated from interactions which might have an influence on the nature of the colour categories, such as contextual information (e.g. shape, texture), lexicalisation, communication of colour meaning, etcetera. However, the results might tell something about the viability of various positions held in the colour category discussion.

The first thing to note is that both approaches have no problem in constructing a set of categories sufficient to discriminate any context thrown at the agents. This is not surprising, as both are hill-climbing approaches. A hill-climbing algorithm has a goal (think about it as the highest peak of a landscape) and a continuous evaluation function (the height already reached). The algorithm tries to reach the highest peak in the fitness landscape through some process of changing the representation. Two of many possible approaches are demonstrated here: one adaptive approach, which uses a learning mechanism to reach its goal, and one evolutionary approach, which uses random mutation of the representation to search an optimal solution.

The number of categories generated by both approaches is very different. The adaptive approach creates just enough categories to discriminate the environment. This is a property of the algorithm, and can be contributed the threshold τ_{adapt}, if the discriminative success is equal or greater than τ_{adapt} no categories are created anymore. The evolutionary approach creates categories without bound, since a high number of categories has no influence on the discriminative performance. The evolutionary approach, as implemented here, keeps on creating categories and may evolve more categories than there are possible stimuli in the environment. This very unnatural behaviour is due to the absence of any inhibitory force on the creation of colour categories. Other influences, such as a probability of mutation correlated to the fitness, or the communication of category meaning, or constraints on neural correlates might provide an inhibitory influence on the number of categories.

In the evolutionary approach, the coherence between the categories of different agents is markedly good. This is not unexpected, as the complete population is the offspring of only a few fit agents, which are again the offspring of only a few fit agents, etcetera. The coherence between the categories is genetically determined: all agents share the same genotype (besides some variation due to random mutation), and consequently all have identical colour categories. This account is often given when explaining the (much debated) universality of human colour categories, see for example (Durham, 1991).

The only disconcerting result is shown when the agents need to adapt to a changing chromatic environment. Both approaches can cope with change, but the timescale on which they do is fundamentally different. The adaptive approach is able to adjust the categorical representation within one individual during its life cycle, while the evolutionary approach needs several generations before it can cope with changing external conditions. The simulation is certainly not a good model of how genetic change occurs in nature: it uses asexual reproduction relying only on mutation, and the mutated parameters are very likely not faithful to the ones evolution operates on in biological species. However, it gives a relative idea

of the timescale needed by evolutionary dynamics to adapt to changes. Humankind often needs to adapt to radically changing chromatic environments (Casson, 1997; Zollinger, 1988) and its seems likely that these representational changes have come about in an adaptive manner and not through evolutionary adaptation. This is an argument often given by opponents of radical nativism (Worden, 1995; Elman et al., 1996).

Both approaches yield a sufficient and adequate categorical repertoire for discriminating colour stimuli, but the categories do *not* resemble human colour categories. Although the behaviour of the categorisation is quite similar to human categorisation, every run of the simulation arrives at a different category constellation. Phenomena like Berlin and Kay's evolutionary order in the development of colour categories are not observed in the simulations. The reasons for this are manifold. First, the simulations are built on models of perception, categorisation, learning and evolution which have obvious limits to their realism. Second, the interaction between the individual and the environment is fixed: colour stimuli are presented for discrimination, without their being any bias on the selection of colour stimuli. For example, in constructing the context a saturated red stimulus is not favoured over a desaturated lime stimulus, meaning that the emergence of a category for saturated red or desaturared lime is equally probable. And third, the colour stimuli are devoid of any context. For example, in many human cultures red is related to strong contextual meanings (such as danger, fire, blood or fruit), which might lead to the development or genetic anchoring of a prominent category for red. As colour context is lacking, all stimuli are treated as equally important in the simulations.

7.6 Summary

This chapter has shown a variety of results obtained by running the simulations as described in chapters 5 and 6 *without language*. The simulations demonstrate the diachronic evolution of colour categories and colour lexicons, driven by a task for discriminating perceived colour stimuli. Two strategies are investigated, in the first strategy categories are adapted using an instance-based learning approach. In the second strategy categories are genetically evolved, using Darwinian evolution. The following conclusions can be drawn:

- Both approaches manage to attain a set of categories performing (almost) perfectly at discrimination task at hand, which is not really surprising as both strategies are optimization strategies. The adaptive strategy however deploys learning during the lifetime of an individual, whereas the genetic evolution functions uses generational reproduction to search the solution space.

- Individual learning leads to a certain amount of sharing of colour categories within the population, which can be attributed to the shared physiological, environmental and ecological constraints, but there is no 100% coherence.
- Genetic evolution leads to a complete sharing of colour categories in the population (apart from some small variation due to random mutations). The reason for this being that genetic evolution favours only a few fit individuals of which the genetic information is carried across generations.
- Both individual learning and genetic evolution do not arrive at colour categories shared across populations.

Chapter 8

Results on learning and evolution with language

This chapter reports on experiments in which categories are acquired under influence of linguistic communication. Again two approaches are investigated, (1) categories are learned individually, but are under pressure of linguistic communication and (2) categories are genetically evolved, but individuals are selected on their communicative proficiency. It is shown that both cultural learning and genetic evolution with language result in shared categories, needed for successful meaning transfer between individuals. For (1) it is shown that cultural transmission of category labels and feedback of the communication on the category adaptation results in a shared lexicon and a shared category set. For (2) it shown that genetic evolution of colour categories is possible with pressure from linguistic communication.

8.1 Cultural learning

In this section the simulation models from chapter 5 are used. A population consists of agents playing guessing games. The guessing game consists of an individual component and a social component. In the individual component, implemented as the discrimination game, agents construct category sets for distinguishing stimuli in the environment; results of its behaviour have been demonstrated in section 7.2. The social component consists of a linguistic communication task where two agents have to convey colour meaning to each other. For this agents attach word forms to their private categories and these word forms are then used in a conversation-like interaction. Agents can learn word forms from others and can adapt their internal categorical representations to better represent signified meaning in the environment.

8.1.1 An illustrative experiment

This illustrative experiment serves to present some new parameters for the cultural learning of categories and to gently introduce the reader to the dynamics of the cultural learning experiments.

The population consists of $N = 10$ agents. As input for the agents the simple environment from section 7.2.1 is used. It consists of 11 eleven colour stimuli corresponding to red, orange, yellow, lime, turquoise, blue, bluish purple, purple, white, grey and black. The context consists of $|O| = 2$ stimuli chosen from these possible eleven stimuli, the stimuli in the context are at a minimum distance of $D = 50$. The agents play 5000 guessing games. Each game is played by two agents randomly chosen agents from the population. So on average each agent will have played $\frac{T}{N}$ games, half of it as speaker and half of it as hearer.

A part of the dynamics of cultural learning is the forgetting of word forms and categories, as discussed in section 5.1.4. The parameters for deleting underused categories and forms are set as in table 8.1. Remember that these parameters are merely "household parameters" and serve to rid the lexicons and categorical repertoire of unused and unreacting word forms and categories. If parameters differ for certain simulations, this will be mentioned alongside.

Figure 8.1 shows the average discriminative success $\overline{DS}$, it rises quickly and at game 5000 it is 0.99 ± 0.02. It shows that the agents obtain a categorical repertoire that is sufficient and adequate to discriminate the topic from the other stimulus in the context. The creation and adaptation of categories is not identical to the case where categories are individually learned (see 7.2), as in the guessing game the categories are also adapted by the feedback given during the linguistic interactions.

The average communicative success $\overline{CS}$ in figure 8.2 shows how successful agents are at conveying meaning. The communicative success is the running average ratio over the last $T_a = 20$ games of successful interactions. The communicative success evolves during the simulation, it rises steadily and reaches 0.99 ± 0.02 at game 5000, meaning that from that agents "understand" each other (almost) perfectly. Note that the communicative success is always lower than or equal to the discriminative success: when a stimulus cannot be distinguished from the other stimuli it also impossible to communicate about it.

Figure 8.3 shows the average number of categories for the population. The number of categories rises quickly in the population, the dynamics of the guessing game leave the agents with more categories than are needed to successfully discriminate and communicate about the given context. The mechanism that makes the agents "forget" underused categories kicks in after about 500 games; it removes categories of which the locally tuned units are unresponsive (because their weight w is near zero) or categories

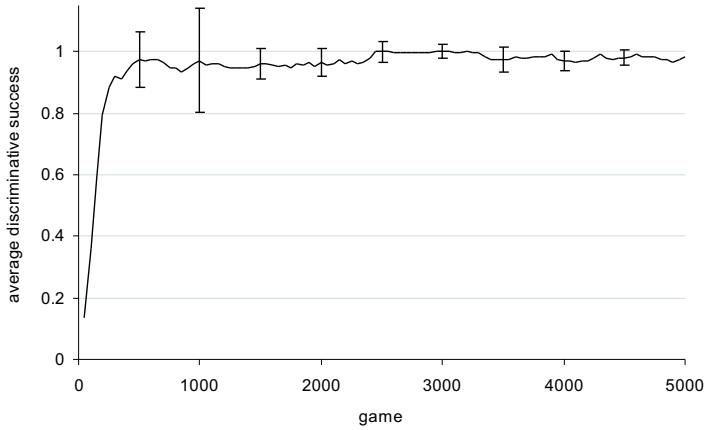


Figure 8.1: Average discriminative success $\overline{DS}$ (cultural learning, simple stimuli set, $N = 10$, $|O| = 2$, $D = 50$).

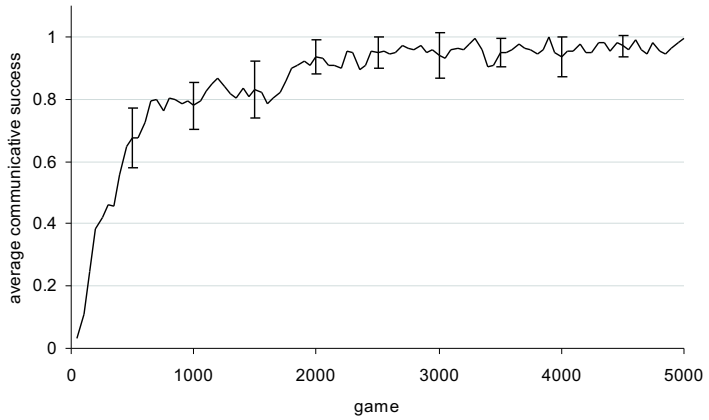


Figure 8.2: Average communicative success $\overline{CS}$ (cultural learning, simple stimuli set, $N = 10$, $|O| = 2$, $D = 50$).

Parameter	Value	Explanation
N	10	Number of agents.
$\|O\|$	2	Number of stimuli in the context.
T	5000	Number of games played.
T_a	20	Number of games over which the running average is taken for computing the DS and CS.
τ_{adapt}	0.95	Threshold for the discriminative success, if the DS is above this threshold, no categories are added anymore, only existing categories are adapted.
σ	5.0	Width of the Gaussian of the locally tuned units.
w_{default}	1.0	Default weight of locally tuned unit, set when creating a new locally tuned unit.
α	1.0	Learning rate for the adaptive network.
β	0.95	Weight decay for the adaptive network.
$\theta'_{\text{last used}}$	50	If a category has not been used in the last 100 games and it has no associated word form, then it will be deleted from the categorical repertoire.
$\theta_{\text{last used}}$	100	If a word form has not been used in the last 200 games and it older than θ_{wfage}, it will be deleted.
θ_{score}	0.2	If a word form's score is lower than 0.2 and it is older than θ_{wfage}, it will be deleted.
θ_{wfage}	50	If a word form is older than 100 games, it will depend on its score or its last use if it will be deleted.

Table 8.1: Parameter settings for cultural learning.

that are not lexicalised and that have not been used by the agent in the last $\theta'_{\text{lastused}}$ games it was involved in. The agents evolve towards having a parsimonious set of categories. At game 5000 the population has on average 4.2 ± 0.4 categories of which 4.2 ± 0.4 categories are lexicalised. The number of lexicalised categories will always be smaller or equal to the number of lexicalised and non-lexicalised categories, because categories do not necessarily need to be associated with word forms: a word form is associated with a category only when that category needs to be communicated. Looking closer at the number of lexicalised categories each agent has, it is worth noticing that the agents need only 4.2 categories on average to distinguish one colour stimulus from one other stimulus, while these can be any of eleven possible stimuli. The agents generalise over their internal colour space, a generalisation that is influenced by the nature of the task (discrimination). The number of categories depends on the complexity of the environment (the number of stimuli and the perceptual distance between these in the context).

In relation to the number of categories in the population, it is also worth following the number of forms in the population. Figure 8.4 shows the

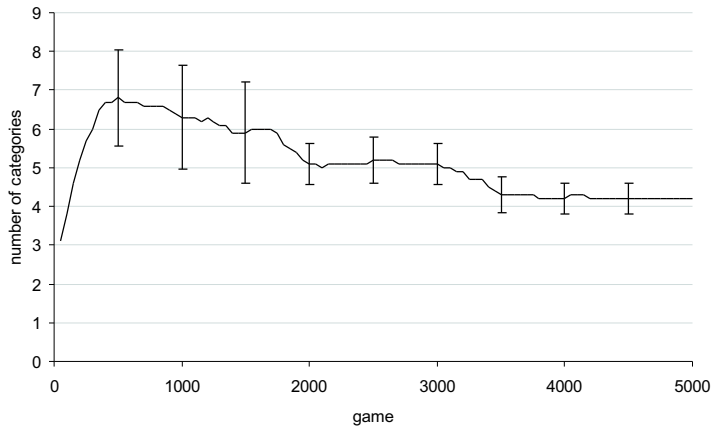


Figure 8.3: Average number of categories (cultural learning, simple stimuli set, $N = 10$, $|O| = 2$, $D = 50$).

number of unique word forms in the population. Initially the agents all create word forms to lexicalise their newly created and thus unlexicalised categories, after playing a number of discrimination games self-organisation takes care of reducing the number of word forms in the population. Unsuccessful word forms will disappear from the population through the mechanism of forgetting unsuccessful word forms: word forms of which the score has dropped below a certain threshold or which have not been used for a certain period of time, are deleted from the agent's lexicon (see 5.1.4). The plot shows how the number of unique word forms in the population decreases from 33 at game 450 to 6 at game 5000. Figure 8.5 plots the score of the word forms of one category of one agent. The positive feedback between the use and success of a word form causes only a few word forms to become used throughout the population. All agents converge towards the same lexicon because as soon as a word starts to become successful in the population its success (represented by the form-meaning score) grows until it takes over in a winner-takes-all effect due to the non-linear nature of the positive feedback loop.

The average interpretation variance $\overline{IV}$ is defined in 5.2 on page 82. It is a measure for the coherence with which agents react to a word form, and is computed by calculating the distance between all categories in the population associated with a word form. The interpretation variance for the entire population is the average interpretation variance for all word forms occurring in the population. In plain words, the interpretation variance shows how well agent agree on the meaning of each word form used in the

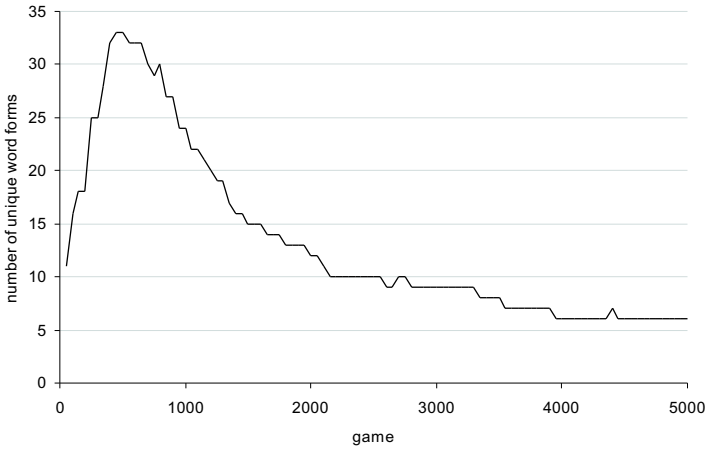


Figure 8.4: Number of unique word forms (cultural learning, simple stimuli set, $N = 10$, $|O| = 2$, $D = 50$).

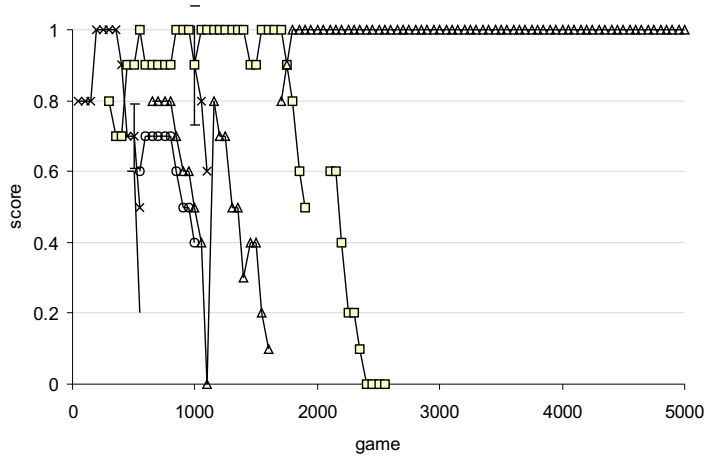


Figure 8.5: The score of five word forms associated with one category. Initially the word are competing until one dominates due to the winner-take-all effect (cultural learning, simple stimuli set, $N = 10$, $|O| = 2$, $D = 50$).

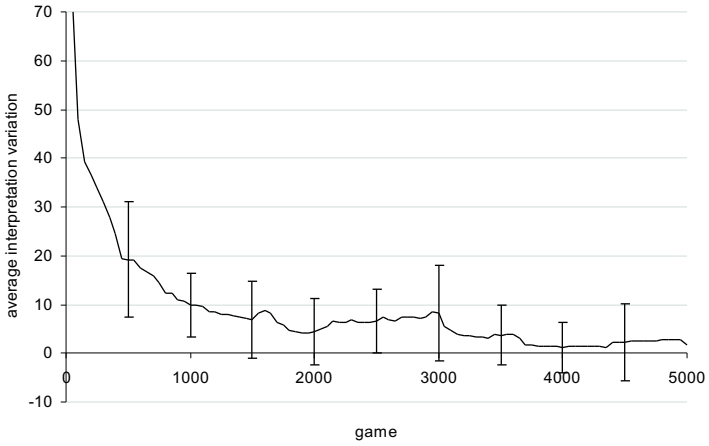


Figure 8.6: The average interpretation variation $\overline{IV}$ (cultural learning, simple stimuli set, $N = 10$, $|O| = 2$, $D = 50$).

population. Note that the interpretation variance for a word form f at its best when $IC_f = 0$, signifying that the agent associates f with a category that is alike to all other categories of other agents associated with f. Figure 8.6 shows the average interpretation variance $\overline{IV}$ for the population. It decreases exponentially and reaches 1.7 ± 7.8 at game 5000, this can be read as all agents having similar meanings for all word forms in the population.

The category variance CV is measure for the similarity between the categories of all agents in the population. If one would like to know how consistent all agents react to the environment, the category variance will tell. The lower CV is, the better the coherence between the categorical repertoires is. If all categorical repertoires would be identical, CV would be zero. Figure 8.7 shows the category variance for our simple experiment. It decreases rapidly, and is 0.44 ± 0.60 at game 5000. A quick glance at figure 7.6 on page 101, which shows the category variance for a similar experiment but *without* language, shows that the coherence between the categories is now markedly better. The influence of linguistic communication is entirely responsible for this: the fact that agents adapt their categories, according to the feedback given during the guessing games, makes their categories converge until linguistic sharing of meaning is obtained.

Finally, figure 8.8 shows the categories of all agents plotted in $L^*a^*b^*$-space. Only the location of the maximal reaction of the category is plotted, it is filled in the colour to which the category reacts best.

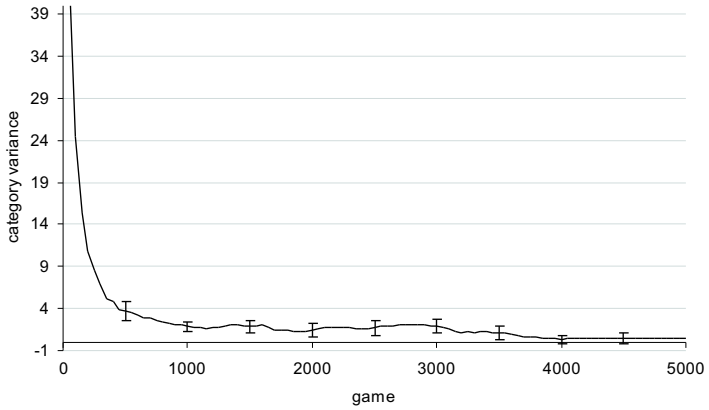


Figure 8.7: The category variation CV (cultural learning, simple stimuli set, $N = 10$, $|O| = 2$, $D = 50$).

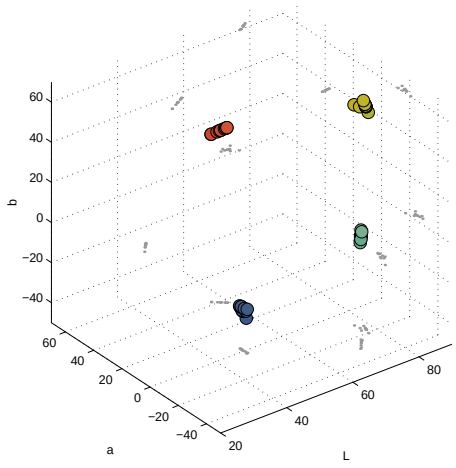


Figure 8.8: The colours to which all categories react highest plotted in the $L^*a^*b^*$-space.

8.1.2 Experiment with full Munsell set

As opposed to the previous experiment, which used a limited set of input stimuli, the next experiment uses the full set of Munsell chips as stimuli. A context now is selected from a possible 1269 stimuli. The experiment has a population containing $N = 10$ agents, the context consists of three colour

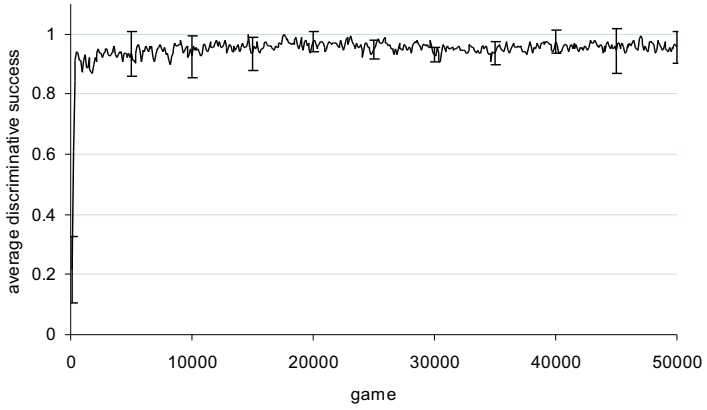


Figure 8.9: Average discriminative success $\overline{DS}$ (cultural learning, full Munsell stimuli set, $N = 10$, $|O| = 3$, $D = 50$).

stimuli $|O| = 3$ which are at a minimum distance of $D = 50$. The agents play $T = 20000$ guessing games. All other parameters are identical to the previous experiment.

Results

Figure 8.9 and 8.10 shows the average discriminative success $\overline{DS}$ and communicative success $\overline{CS}$. The agents rapidly achieve near perfect discrimination, showing that each has sufficient and good enough categories to discriminate the topic from the context. The population's communicative success trails behind, but is 0.87 ± 0.06 after 50000 games, as the population contains 10 agents every agents needs on average 5000 interactions before reaching almost 90% communicative success.

Figure 8.11 shows the average number of categories for the population. At game 50000 the agents have on average 11.1 ± 1.4 categories of which 11.0 ± 1.4 are lexicalised[1]. The number of unique word forms in the population is shown in figure 8.12. In the last games of the simulation, the population contains 21 unique word forms, so on average every category has 1.9 word forms associated to it.

The average interpretation variation $\overline{IV}$ is shown in figure 8.13. The interpretation coherence measures how identical the agents react to the word forms circulating the population. If the agents agree on the meaning of a

[1]Note that this has nothing to do with eleven basic colour categories proposed by Berlin and Kay (1969), other environmental conditions for the simulation will produce different numbers of categories.

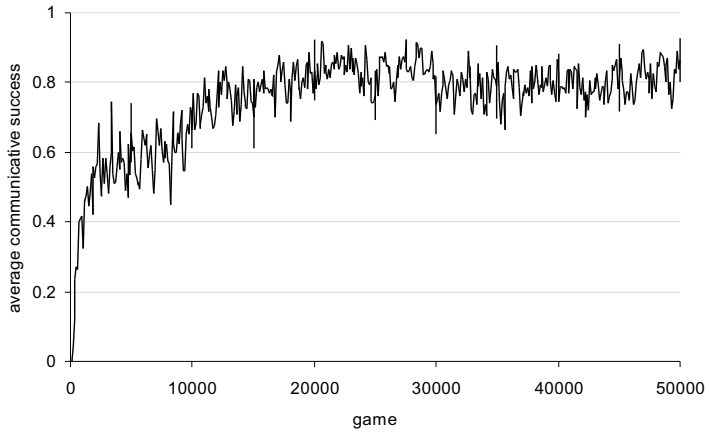


Figure 8.10: Average discriminative success $\overline{CS}$ (cultural learning, full Munsell stimuli set, $N = 10$, $|O| = 3$, $D = 50$).

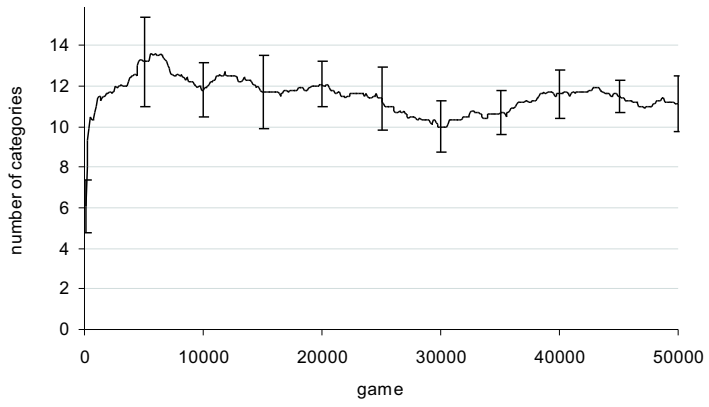


Figure 8.11: Number of categories (cultural learning, full Munsell stimuli set, $N = 10$, $|O| = 3$, $D = 50$).

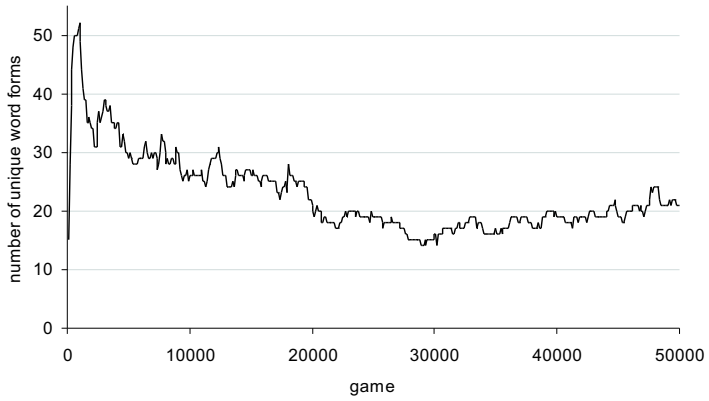


Figure 8.12: Number of unique word forms (cultural learning, full Munsell stimuli set, $N = 10$, $|O| = 3$, $D = 50$).

word form, the $\overline{IV}$ will be lower than when they do not. When the agents's categories are identical, the $\overline{IV}$ will be 0. The category coherence CV on the other hand shows how well the categories of the agents agree. If categories are more similar, CV will decrease. The evolution of the category coherence is plotted in figure 8.14. It decreases exponentially, which means the agents quickly start using categories that are similar. Compared to the category variance in section 8.1.2, where categories are learned without language, it shows how language has a drastic influence on the coherence of the categorical repertoires.

8.1.3 Shared colour lexicons

The mechanism through which the agents each reach a lexicon with which they can communicate colour terms is described in 5.1.2. In a nutshell, the agents keep a score between word forms and categories; word forms that prove successful at communicating a colour stimulus are re-enforced. In selecting a word form the agents will select the word with the highest score. After a certain amount of interactions the agents are left with a set of form-meaning pairs of which the forms are shared in the population.

Figure 8.15 shows a diagram of the number of agents having a certain word form in their repertoire. The population shown here contains 10 agents, during a 18000 games run. A total of 124 word forms have been created during the run, of which only 14 word forms are still in the agents' repertoires after 18000 games. The figure highlights two exemplary word forms; one that stabilises quickly and is kept during the runs, and one that

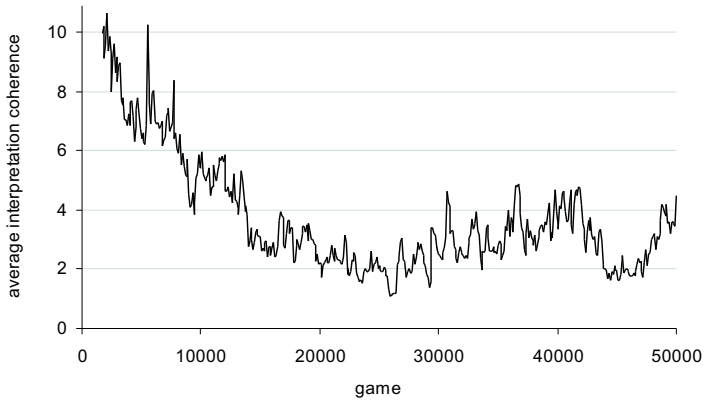


Figure 8.13: Average interpretation variance $\overline{IV}$ (cultural learning, full Munsell stimuli set, $N = 10$, $|O| = 3$, $D = 50$).

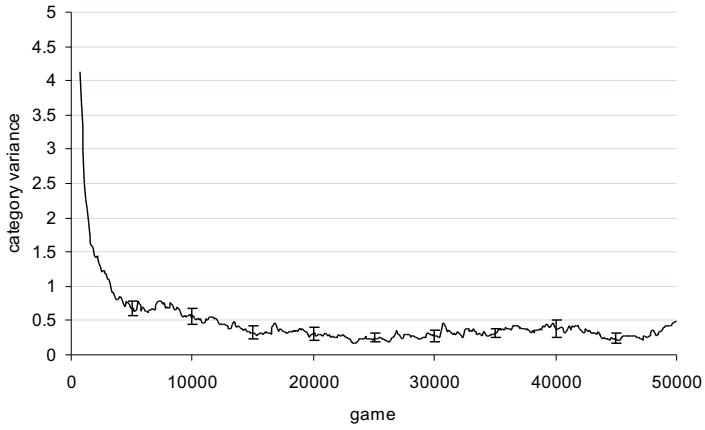


Figure 8.14: Category variance (cultural learning, full Munsell stimuli set, $N = 10$, $|O| = 3$, $D = 50$).

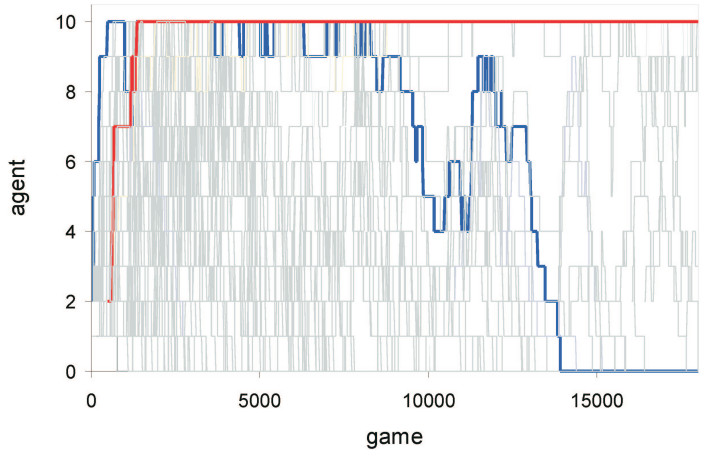


Figure 8.15: A competition diagram for the simulation described in 8.1.2. It plots the evolution of the number of agents having a particular word form in their repertoire. The population contains 10 agents, the context contains 4 stimuli. The evolution of all 124 word forms is plotted, and two are highlighted. One showing a perpetuating word form, spoken by all agents. The other shows a word form that permeates the whole population but then disappears through competition with other word forms.

is used by all agents during the first ten thousand games but after that is lost from all repertoires due to form-meaning competition[2].

8.1.4 Pressure to create colour categories

The influence of the environment on the number of categories created is interesting to follow. It has been stated that the number of colour words in human cultures is correlated to the cultural and technical evolution of its users, would similar circumstances in the model lead to more colour categories? Would the complexity of the environment provide pressure for the agents to create categories that allow to agents to tackle the complex context thrown at them? In the experiment without language it is indeed the case that a more complex environment produces the need for more categories.

In an experiment we vary the perceptual similarity between the colour samples (it was already shown that varying the number of stimuli in the context had little influence on the number of colour categories). The similarity of colour samples is expressed as the Euclidian distance D in the CIE $L^*a^*b^*$ space between colour samples; categories are more similar if D is

[2]More on form-meaning competition and how this depends on stochastic processes and the choice of learning algorithms can be found in (Steels and Kaplan, 1998).

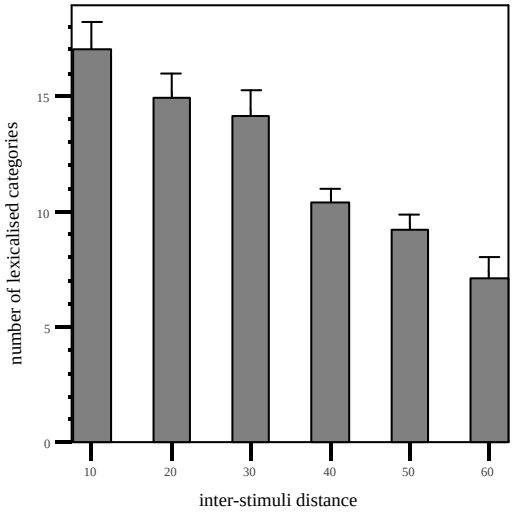


Figure 8.16: The average number of lexicalised categories in function of the inter-stimuli distance D. Population contains 10 agents and context of 4 colour stimuli. The error bars are located at 95%. A higher similarity of stimuli presses the agents to create more categories.

lower. In the experiment D is varied from 10 to 60 in incremental steps of 10, the context contains four colour stimuli and the population contains 10 agents, each simulation runs for 50000 games. Figure 8.16 shows that if the stimuli are closer together, the agents are forced to create more –and therefore more fine-grained– categories.

8.1.5 Nature of the emerged colour categories

The previous sections have illustrated how a population can arrive at common categories through communicative interactions, but little has been revealed of the nature of the categories. The colour domain, in which our perception and categorisation models are situated, should provide some interesting matter for discussion. In its current form, the models and simulations are not under the influence of any bias in the environment. For example, red colours have not got a different status than any other colours. Also, there is no internal bias. The internal colour representation does not prefer categories at opposed locations in colour space or at any other location. Hence it is interesting to see what colour categories are created under these circumstances.

Figure 8.17 shows the categories of two agents from the experiment in 8.1.2 plotted on a Munsell chart. There is a certain resemblance between the category sets, but there is still some variation visible; the categories are all learned individually, the coherence arises because the agents adapt their categories to better facilitate communication. Note that the foci of the categories are not located on the fundamental colours. The model does not contain constraints to arrange the categories on any specific locations.

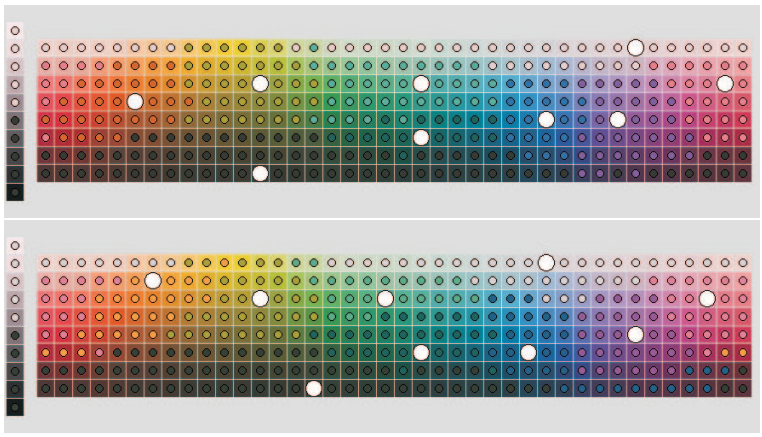


Figure 8.17: Munsell plot of the category sets of two agents, the categories are culturally learned.

8.1.6 Agreement across populations

The category variance shows how well the categories of agents of one population agree (see for example figure 8.14), and we have demonstrated that language has a beneficial influence on the coherence within a population. However, it should be interesting to see if the addition of language could drive different populations towards sharing the same colour categories.

In an experiment 5 simulation were run with identical parameter settings ($N = 10$, $T = 50000$, $|O| = 3$, $D = 40$, full Munsell set as input), but each simulation with a different random seed. If there are strong universal constraints on the cultural learning of the categories, all five populations should have near identical category sets. If not, the category sets of the five populations will show no resemblance. Note that the different populations are not in linguistic contact with each other, they only share the same environment and same the perceptual and cognitive apparatus. Table 8.2 shows the intra-population category variance for five populations. As can be seen, the variance is low within a population, but is still high between two populations. The simulations do not seem to contain constraints steer-

ing different population towards having the same categorical repertoires. Each population finds one of many solutions for discriminating and communicating colour categories.

cv'	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}_3$	$\mathcal{A}_4$	$\mathcal{A}_5$
$\mathcal{A}_1$	0.30				
$\mathcal{A}_2$	4.29	0.45			
$\mathcal{A}_3$	3.83	4.52	0.36		
$\mathcal{A}_4$	5.09	5.60	5.31	0.51	
$\mathcal{A}_5$	5.26	5.80	5.37	6.08	0.55

Table 8.2: Intra-population category variance CV' of 5 populations of which the categories are culturally learned.

8.1.7 Influence of language on shared categories

The influence of linguistic interaction on the nature of the emerging colour categories can be studied by removing the need to communicate colour terms in the simulations. Agents have an opportunistic mechanism of categorising their internal representations, but the pressure to successfully communicate their categories drives them towards sharing those categories. Taking away language leaves the agents with categorical repertoires that are not necesarilly shared.

This is demonstrated here by running two different simulations. One in which the agents only carve up the colour space through playing discrimination games, this should leave each agents with a set of categories adequate to discriminate any colour context. In the second kind of simulations linguistic interaction are introduced; this happens through letting the agents play guessing games, of which ample results have been shown in the previous sections. It is expected that the linguistic interactions will drive the population towards a more or less shared lexicon. What is interesting about it is the amount of coherence that can be reached solely through having to same "cognitive" apparatus and through sharing the same environment. The internal $L^*a^*b^*$ representation of colour stimuli is inherently opponent (blue stimuli are located at the opposite side of yellow stimuli in the $L^*a^*b^*$ space, the same goes for red and green), but is does not have any bias towards preferring categories at these locations: categories can be constructed anywhere in the representational space.

Figure 8.18 shows the category variance CV averaged over five runs with communication and five runs without communication (each run was started with a different random seed). The population contains 10 agents, the context consists of 4 colour samples (with $D = 40$) and the samples were chosen from the full Munsell set. The plot shows how linguistic in-

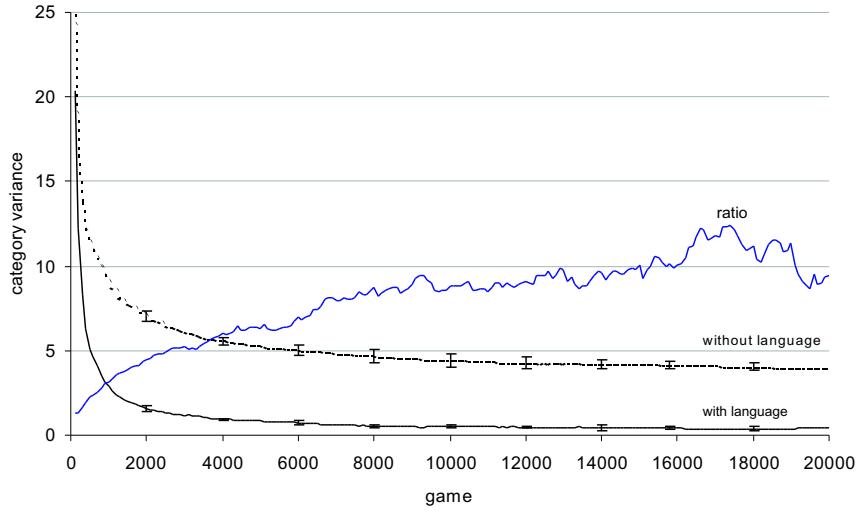
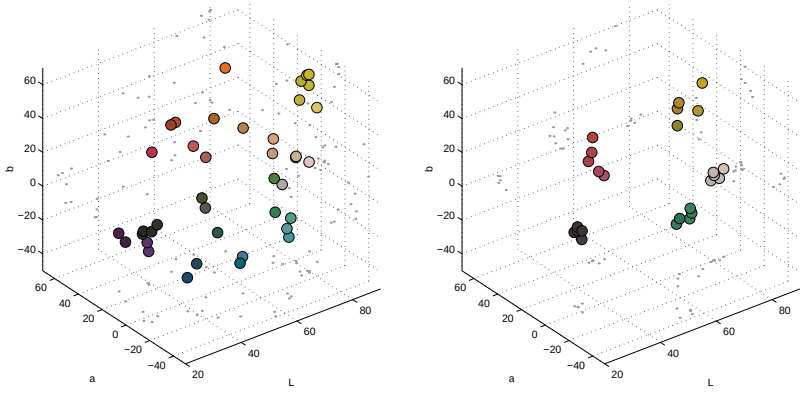


Figure 8.18: The category variance CV averaged over five-fold run with communication (full curve) and an identical five-fold run but without communication (dotted curve); for each curve the standard deviation over the five runs is plotted. The third curve shows the ratio between the category coherence without communication and with communication, note that the variance decreases about 1000% when communication is introduced.

teraction accounts for a significant boost in the coherence of the categories. The figure also shows the ratio between the coherence between the categories created by agents that did not communicate and agents that did communicate; communication increases the coherence by about 1000%.

To again illustrate the influence of language on the coherence of the categories, two simulations containing $N = 5$ agents were run for $T = 10000$ games ($|O| = 3$, $D = 40$). One simulation using individual learning and one using cultural learning. The location of the maximal reaction of all categories is plotted in figure 8.19. With cultural learning, clusters have formed as the categories of the different agents have become more similar due to the coupling of category acquisition and language.



(a) (b)

Figure 8.19: Categories plotted in the $L^*a^*b^*$-space for (a) individual learning and (b) cultural learning.

8.1.8 Learning human colour terms

A model for perceiving, representing, categorising and lexicalising colour should also be adequate to be trained on human colour categories and human colour terms. Using human colour categories and terms for training the agents provides a nice crosscheck for the model and can validate it as a good model for representing human-like colour categories and colour terms.

A single agent is trained on American English colour terms as recorded by (Berlin and Kay, 1969), see figure 8.20. For constructing the chart, American subjects were asked to point out the focus and extent for eleven colour terms. The training set for the simulation consists of all colour stimuli to-

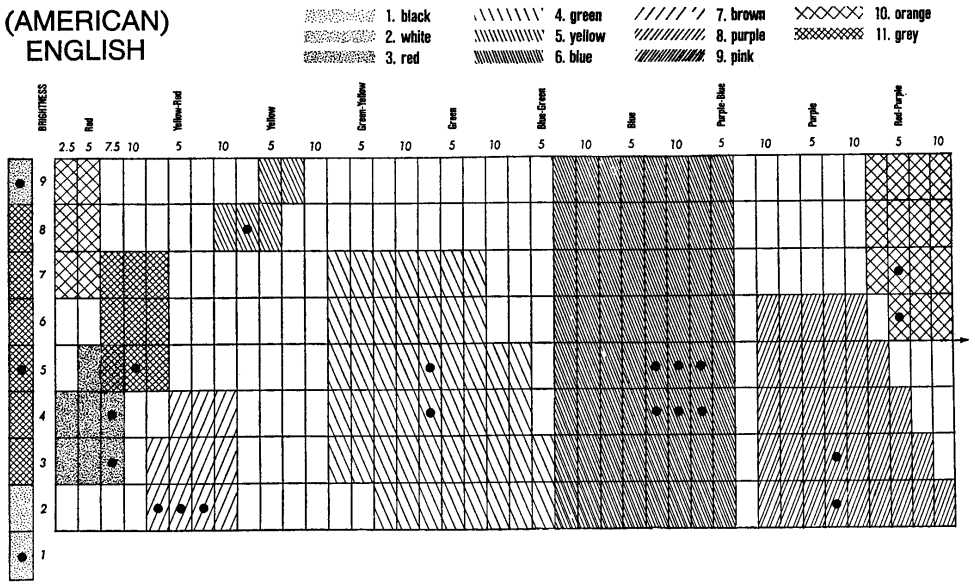


Figure 8.20: The extent and foci of eleven American English colour terms, from (Berlin and Kay, 1969, p. 119). The black dots show the focus of each colour term as pointed out by the subjects, while the pattern fill shows all colour chips corresponding to that colour term. The chart and legend contain some errors: "10. orange" should read "10. pink" and "9. pink" should read "9. orange". Also the pattern in the chart for colour term 9 (third from the left) does not correspond to the pattern in the legend.

gether with a colour label, for example there are 64 colour stimuli with the label "blue" and five stimuli with label "yellow". The chromatic stimuli are presented as spectral power distributions, as measured by (Parkkinen et al., 1989). As Parkinnen and his group only measured chromatic Munsell chips and not the achromatic (white, black and grey) Munsell chips, we had to resort to measurements of (Newhall et al., 1943). The achromatic chips are given in CIE Yxy format. It is impossible to revert the Yxy-values to spectral power distributions; therefore the Yxy-values are converted to CIE XYZ and are as such presented to the agent. In total the training set consists of 213 labelled stimuli.

The agent has the same parameter settings as used in the experiments in previous sections. After playing 4000 games, the agent has on average seen each stimulus 18.8 times. As the agent does not play any discrimination or communication games here, it is only trained to "memorize" the English term and their meaning, it makes no sense to interpret the experiment in terms of discriminative or communicative success, or other measures relat-

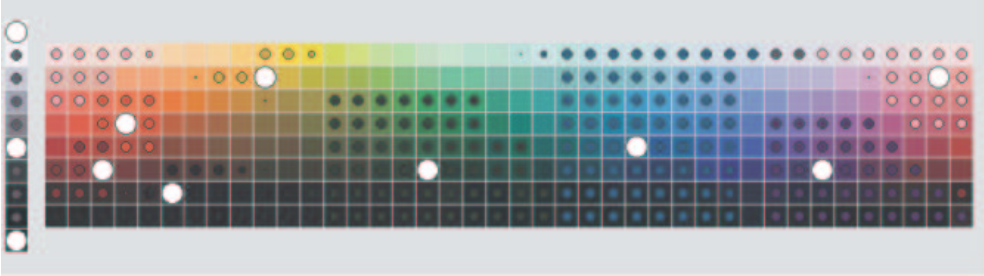


Figure 8.21: Chart showing the results of learning American English colour terms of figure 8.20. The chip to which each category reacts best is marked with a white circle. The extent of each category is shown by circles filled with the colour of the chip to which the category reacts best; the size of the circle represents the output of the category at that location.

ing to population effects, such as category variation or interpretation variation.

Figure 8.21 shows a reproduction of the colour chart used by Berlin and Kay, plotted on it are the foci and extent for the eleven colour categories. The foci are shown with a white circle. The extent of each category is drawn with a circle filled with the colour of the focus; the size of the circles is relative to the reaction of the category to the colour of that chip. The figure allows a subjective interpretation of the models capacity to adapt to label-stimulus training data.

The categories are quite truthful to the training set. Only in regions where the $L^*a^*b^*$-values of Munsell chips are closer together, a certain "overflow" of the categories can be observed. These are the regions where high and low lightness (the upper and lower row on the chart), there the $L^*a^*b^*$-values of the stimuli are close together, as can be seen on figure 7.8 on page 104. The Gaussian receptor functions of the adaptive networks have a fixed width (defined by the σ parameter in equation 4.4 on page 57), and the receptors are too wide for chips close together in the $L^*a^*b^*$-space; therefore categories including very light or very dark colours tend to bleed somewhat. This can be seen for the "blue" category, which includes some light-blue chips that where not in the training data. This also goes for the "pink" category, and to some extent for the "yellow" category. This might be fixed by making the width σ of the receptors adaptable, positive and negative training examples can then be used to fine-tune σ.

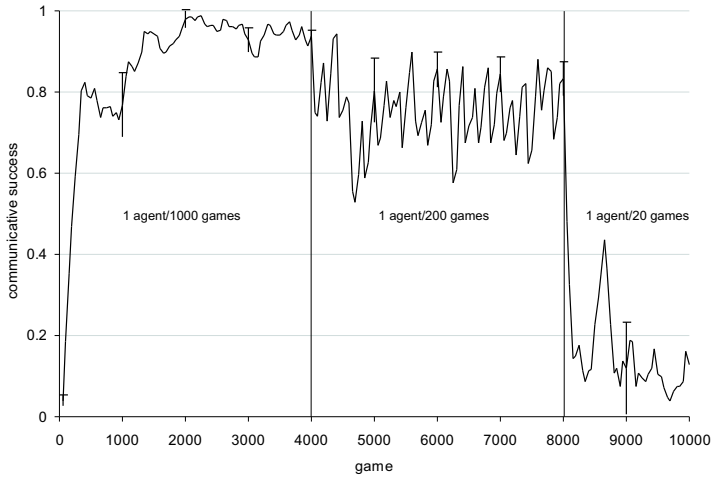


Figure 8.22: Illustration of memetic evolution in a population of 5 agents. In each game the context consists of 4 stimuli ($D = 40$) chosen from the complete Munsell set. A flux is introduction by replacing an agent after 4000 games. Too high a flux destabilises the communicative success.

8.1.9 Memetic evolution

Here we examine what happens when populations engaged in cultural evolution change. This is done by introducing a flux in the population. At regular time intervals an agent is removed from the population and another agent is inserted. The new agent has no prior knowledge of the colour categories nor of the words used in the population. Figure 8.22 shows that at renewal rates that are not too high, communicative success is essentially maintained. New agents obviously fail initially but pick up quickly the words and meanings that are commonly used. This means that the lexicon and the colour repertoire gets transmitted between generations purely through cultural learning. These results are in line with other experiments with much larger agent populations and much larger vocabularies (Steels et al., 2002). This experiment shows how the memetic evolution of language and meaning are possible (Dawkins, 1976; Blackmore, 1999).

8.2 Genetic evolution with language

This section describes results from running the simulation on genetic evolution of categories under influence of language. The categories are evolved as described in 6.1.3, while the language aspect is implemented as guessing games described in 5.1.2 with the understanding that the representation of the categories is not adapted under the influence of the outcome of the

games, but is changed during mutation.

8.2.1 An illustrative experiment

We again start with an illustrative experiment to demonstrate the dynamics of the simulation. The population consists of $N = 20$ agents and runs for $T = 100$ generations. The context contains $|O| = 2$ stimuli at a minimum distance of $D = 50$, the context is chosen from a set of eleven possible stimuli (see 7.2.1). Each generation the $M_{\text{die}} = 2$ oldest agents are replaced by mutated offspring of the two fittest agents (if there are more than M_{die} agents with the same high age, of these M_{die} are randomly selected). The agents are evaluated by playing $N(N-1)$ guessing games: $(N-1)$ in the role of speaker, and $(N-1)^2$ in the role of hearer. The parameter settings for this experiment are summarised in table 8.3.

Parameter	Value	Explanation
N	20	Number of agents.
$\|O\|$	2	Number of stimuli in the context.
D	50	Distance between the stimuli of the context.
T	100	Number of generations.
T_f	50	Number of discrimination games on which the fitness if calculated.
σ	5.0	Width of the Gaussian of the locally tuned units.
w_{default}	1.0	Default weight of locally tuned unit, set when creating a new locally tuned unit.
p_{mut}	0.1	Probability of mutation, see 6.1.1.
σ_{mut}	5.0	Standard deviation of Gaussian random distribution, see 6.1.1.
M_{die}	2	At the end of each generation the oldest 2 agents are replaced by offspring of the 2 fittest agents.
$\theta_{\text{last used}}$	100	If a word form has not been used in the last 200 games and it older than θ_{wfage}, it will be deleted.
θ_{score}	0.2	If a word form's score is lower than 0.2 and it is older than θ_{wfage}, it will be deleted.
θ_{wfage}	50	If a word form is older than 100 games, it will depend on its score or its last use if it will be deleted.

Table 8.3: Parameter settings for genetic evolution with language.

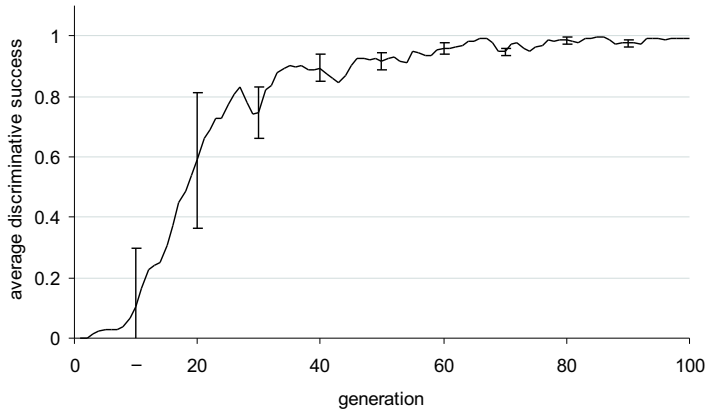


Figure 8.23: Average discriminative success $\overline{DS}$ (genetic evolution with language, simple stimuli set, $N = 20$, $|O| = 2$, $D = 50$).

Figure 8.23 shows the average discriminative success $\overline{DS}$. It rises quickly and at generation 100 it is 0.99 ± 0.01. Figure 8.24 shows the average communicative success $\overline{CS}$, at generation 100 it is 0.93 ± 0.01. As the task on which the fitness is evaluated is a linguistic task (the agents play guessing games) the agents are selected on their communicative success. As discrimination is essential to the guessing game the agents's discriminative success is high as well. We can conclude that the agents evolve a categorical repertoire that is adequate for discrimination, and that the adding word forms to the categories during the guessing game results in successful communication.

Figure 8.25 shows the average number of categories in the population, it is 8.2 ± 1.1 at generation 100, of which 7.0 ± 0.9 categories are lexicalised. The number of categories is lower when language is involved. The need to communicate constrains the creation of categories, as lexicalising and communicating a smaller set of categories is of course easier, it gives agents with a smaller category set an edge over others in the evaluation of their communicative success.

The number of unique word forms in the population is plotted in figure 8.26, it is rather erratic. This flux can be explained by the fact that every generation two agents with a full grown lexicon leave the population, and two without any language enter. During every generation $N\,(N-1)$ guessing games are played, during which new word forms get created and older word forms are removed if they prove to be unsuccessful. So there is quite some "lexical turmoil" in the population.

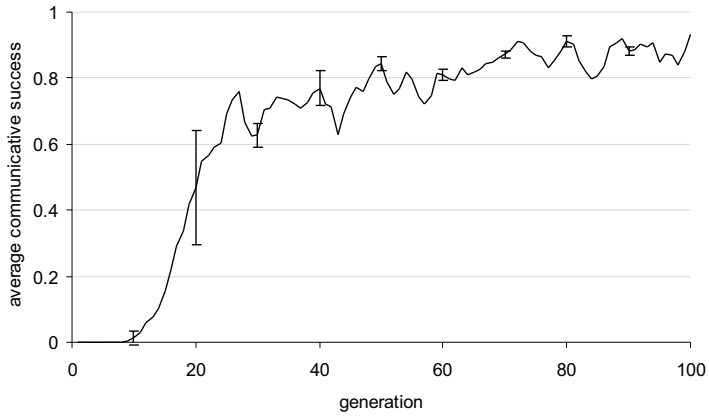


Figure 8.24: Average communicative success $\overline{CS}$ (genetic evolution with language, simple stimuli set, $N = 20$, $|O| = 2$, $D = 50$).

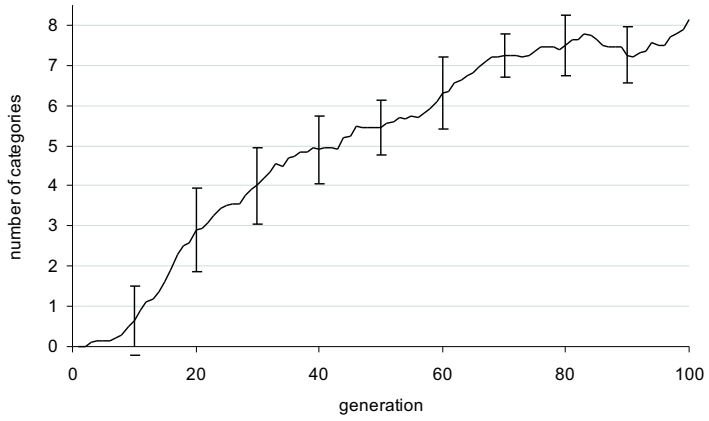


Figure 8.25: Average number of categories (genetic evolution with language, simple stimuli set, $N = 20$, $|O| = 2$, $D = 50$).

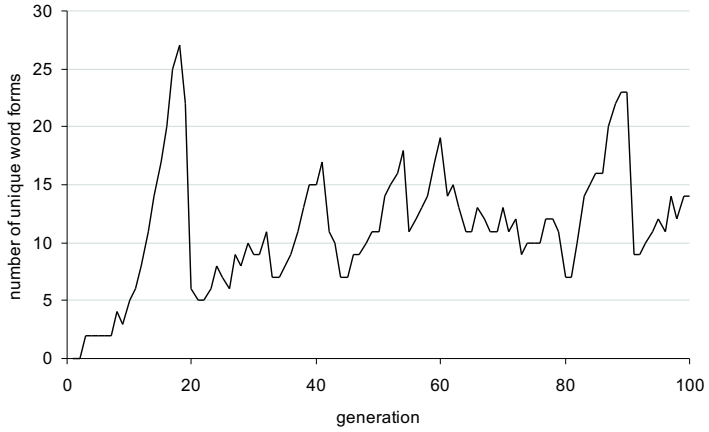


Figure 8.26: Number of unique word forms (genetic evolution with language, simple stimuli set, $N = 20$, $|O| = 2$, $D = 50$).

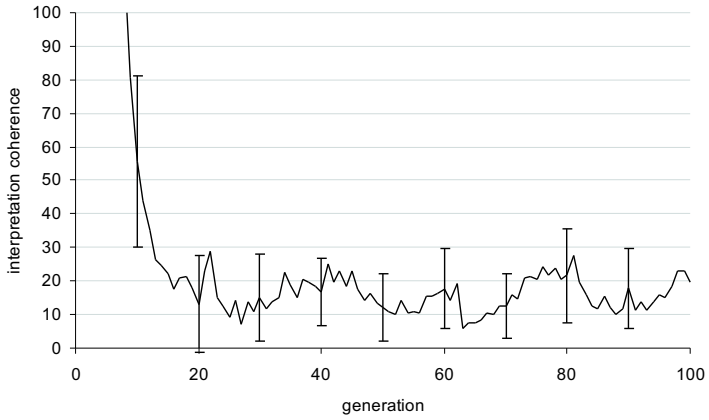


Figure 8.27: Average interpretation variance $\overline{IV}$ (genetic evolution with language, simple stimuli set, $N = 20$, $|O| = 2$, $D = 50$).

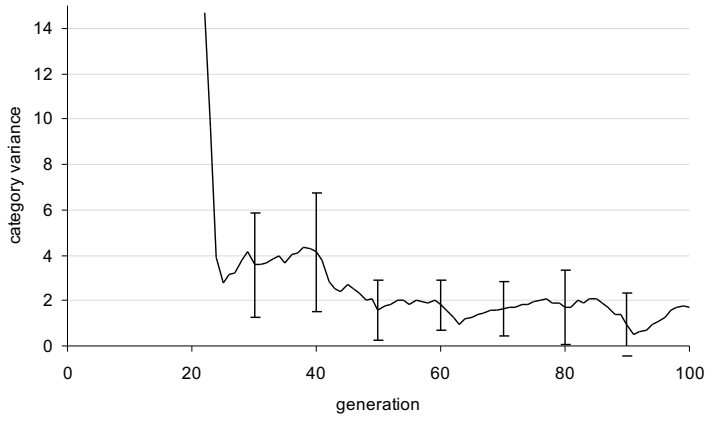


Figure 8.28: Category variance CV (genetic evolution with language, simple stimuli set, $N = 20$, $|O| = 2$, $D = 50$).

The average interpretation variance $\overline{IV}$, showing how well agents agree on the meaning of a word form, is plotted in figure 8.27. The interpretation coherence is lower than in the analogous experiment with cultural learning, but this is not really alarming, as the flux of agents and the mutations are responsible for this. Obviously, the important measure here is the communicative success, which is high. Figure 8.28 shows the category variance for the population, as expected it is low. We already from the experiments on the genetic evolution without language that the categorical repertoires of agents resemble each other because after a while all agents are descendants of only a few successful parents. A successful genome, or in our case categorical repertoire, spreads quickly over the generations leaving every agent with the same set of categories (apart from some variation introduced by the mutation operator).

8.2.2 Experiment with full Munsell set

We now use to full Munsell chip set to select stimuli from (see 7.2.2). The context contains $|O| = 3$ stimuli, with minimal distance $D = 50$. The simulation runs for $T = 400$ generations. Figure 8.29 and 8.30 show the average discriminative success and communicative success. Both increase and at generation 400 they are 0.93 ± 0.02 and 0.83 ± 0.02 respectively. Figure 8.31 shows the average number of categories in the population, and figure 8.32 shows the number of unique word forms in the population. At generation 400 there are 12.85 ± 0.73 categories and 16 word forms present in the

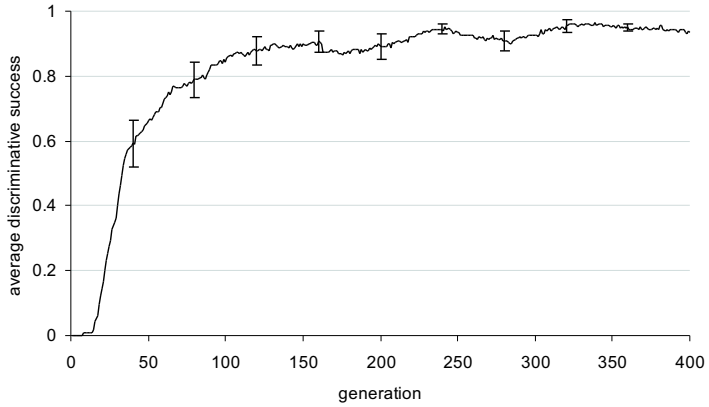


Figure 8.29: Average discriminative success $\overline{DS}$ (genetic evolution with language, simple stimuli set, $N = 20$, $|O| = 3$, $D = 50$).

population. The average interpretation coherence is plotted in figure 8.33, and the category variance in figure 8.34. From these can be read that word forms are coherently interpreted, and that the categorical repertoires are almost identical, apart from some apparent variation introduced by mutation.

So, using a larger set of input stimuli does not produce any surprises. The agents all evolve a shared categorical repertoire under linguistic pressure, showing that genetic evolution of (colour) categories guided by communication is theoretically possible.

8.2.3 Nature of the colour categories

The only pressure on the evolution of the categories is that they should enable communication about a context of colours, implicitly this means that the category set should have sufficient discriminative power and that the categories and their labels should be shared in the population. Figure 8.35 shows the categorical repertoires of two agents of experiment 8.2.2. As can be seen, the categories are located at places where human categories would never be expected. For example, some categories code for adjacent colours on the Munsell plot. This has never been observed in human colour categories, as they are well distributed in the perceptual colour space. The peculiar category constellation is of course due to the genetic evolution: categories can be created anywhere in the colour space, and as long as this constellation delivers a better discriminative or communicative success than the other constellations in the population it will propagate over gen-

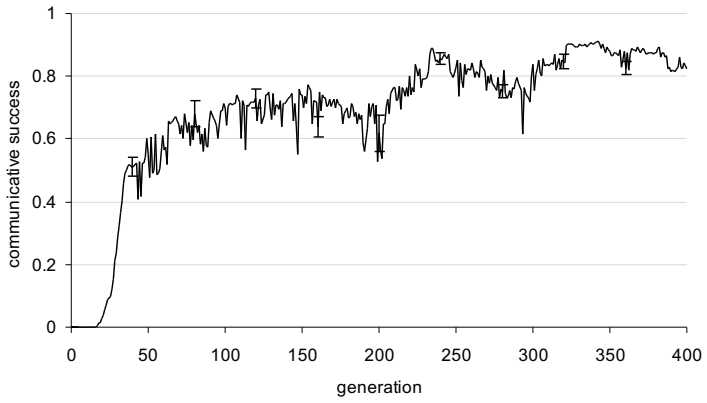


Figure 8.30: Average communicative success $\overline{CS}$ (genetic evolution with language, simple stimuli set, $N = 20$, $|O| = 3$, $D = 50$).

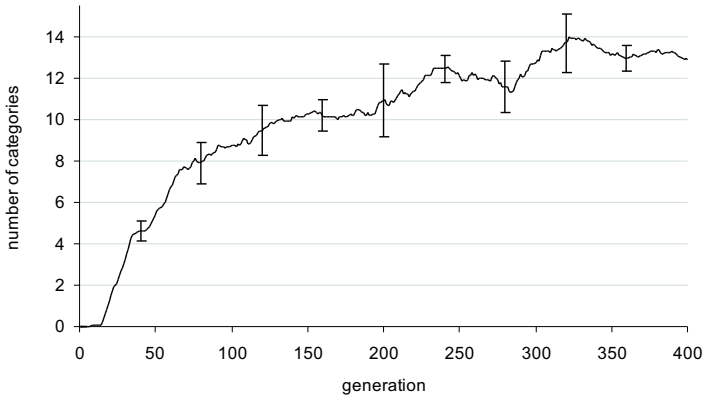


Figure 8.31: Average number of categories (genetic evolution with language, simple stimuli set, $N = 20$, $|O| = 3$, $D = 50$).

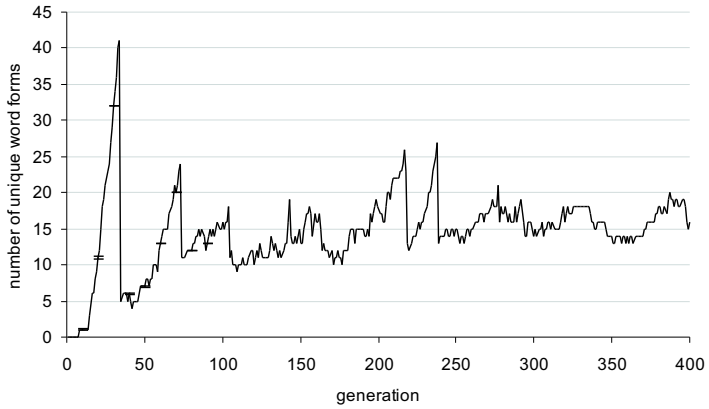


Figure 8.32: Number of unique word forms in the population (genetic evolution with language, full Munsell set, $N = 20$, $|O| = 3$, $D = 50$).

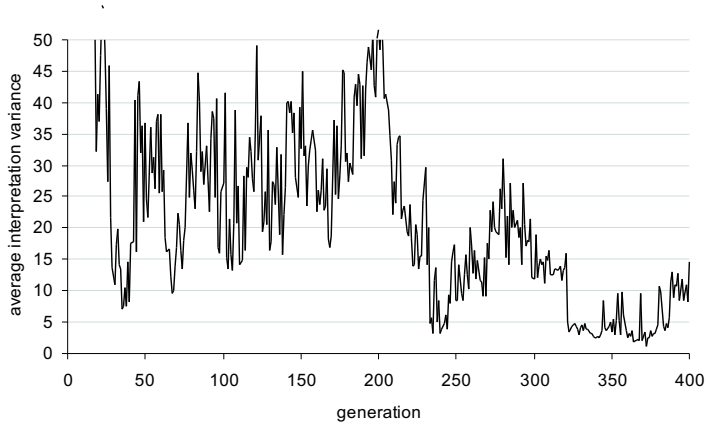


Figure 8.33: Average interpretation variance $\overline{IV}$ (genetic evolution with language, simple stimuli set, $N = 20$, $|O| = 3$, $D = 50$).

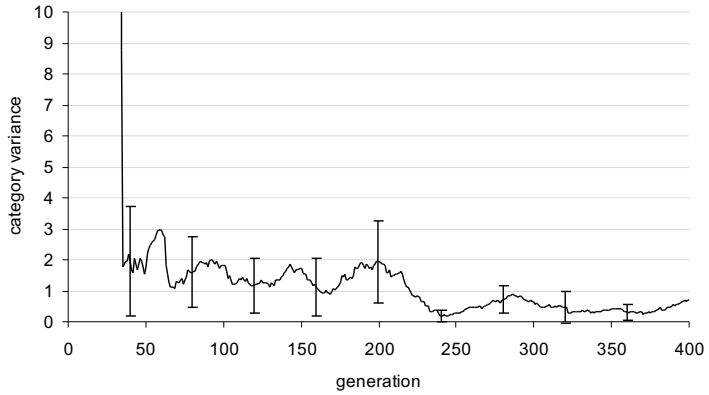


Figure 8.34: Category variance CV (genetic evolution with language, full Munsell set, $N = 20$, $|O| = 3$, $D = 50$).

erations. Evolution often produces artefacts that are not the best solution to the problem at hand, but do provide a good enough solution. Also, evolution sometimes drags along useless remnants of what were once functional and useful adaptations[3].

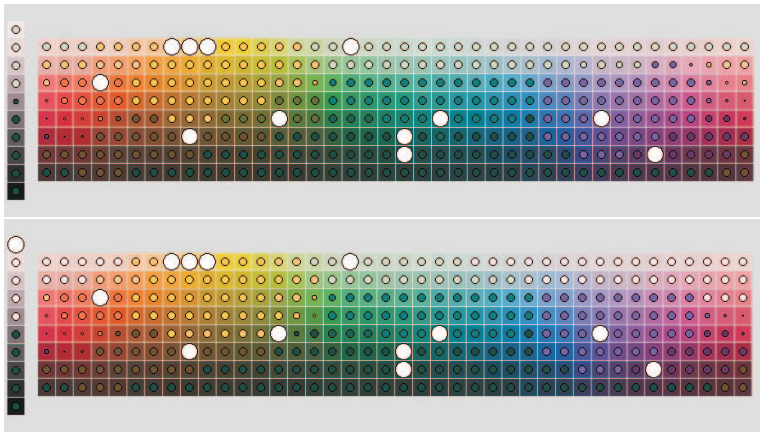


Figure 8.35: Munsell plot of the category sets of two agents, the categories are genetically evolved categories under influence of language.

[3]The appendix is an example of this, evolutionists call this a vestigial organ.

8.2.4 Agreement across populations

As before, we measure the agreement across populations using 5 identical simulations ($N = 20$, $T = 400$, $|O| = 3$, $D = 40$, full Munsell set as input). The question to be answered is: can evolution of a categorical repertoire under pressure of linguistic communication yield categories that are shared between populations. From 7.3.6 we know that evolution without language could not, so might language have that something extra which pushes all agents in different populations to have shared categories. The answer is, as might be expected, no.

Table 8.4 shows the intra-population category variance CV' for the five populations. The variance within a population is low, meaning that the category sets of the agents in a population are similar. But across population, the variance is too high to conclude that agents have the same categories across populations. Again, each population has evolved its own solution to the communication task, a solution different from the ones found in other populations.

cv'	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}_3$	$\mathcal{A}_4$	$\mathcal{A}_5$
$\mathcal{A}_1$	0.91				
$\mathcal{A}_2$	6.90	0.86			
$\mathcal{A}_3$	5.89	6.56	0.98		
$\mathcal{A}_4$	7.45	7.54	7.40	6.98	
$\mathcal{A}_5$	6.35	7.34	6.80	7.01	0.88

Table 8.4: Intra-population category variance CV' of 5 populations of which the categories are genetically evolved under influence of language.

8.3 Discussion

This chapter demonstrated computational models of two different attitudes towards explaining the origins of shared colour categories; (1) cultural learning, the categories are learned under influence of language and (2) genetic evolution with language, the categories are evolved under pressure of language. Let us first discuss the former.

As shown in previous chapter, it is possible to learn colour categories for distinguishing colour stimuli. Cultural learning adds a social dimension to the acquisition of categories: categories not only have to serve to distinguish stimuli, their meaning also needs to be communicated to other individuals. The extra requirement here is that categories need to be "shared" in the population. As shown in the experiment, it is indeed possible to obtain shared categories through learning. Re-enforcement during linguistics shapes categories and the lexicon to enable communication. There is

a two-way structural coupling between category formation and language. Language communication stimulates the formation of categories as it calls for a discrimination game that will create new categories of needed. In its turn category formation will stimulate language because when a category is created, it needs a new word form to be communicated. The language provides feedback on the usefulness of categories and word form, and thus provides cultural constraints.

Note that first learning colour categories and only then learning words, as advocated by those arguing against such a causal influence, would not work because language learning is crucial for the convergence of colour categories. When agents learn categories independently of language their categories diverge too much to support communication later. So both must be learned at the same time in a co-evolutionary dynamics. This shows that the Sapir-Whorf thesis is not only feasible but the best way to reach categorical coherence, and this based on cultural evolution only. Even with the same environmental, physiological and ecological constraints, two populations without contact with each other would develop different colour categories and consequently they would develop colour names with different meanings. Multiple solutions are possible but only one solution gets culturally frozen and enforced through language in each population.

The positive feedback loop between use and success causes self-organisation, in the sense of non-linear dynamical systems theory (Stengers and Prigogine, 1986). The agents converge towards the same lexicon because once a word starts to become successful in the population its success grows until it takes over in a winner-take-all effect. Lexical incoherence may remain in the population if different categories are compatible with a large set of contexts (for example a particular word may for a long time be associated with yellow and green as long as this does not cause conflict during a guessing game). The meaning will be disentangled when a situation arises where yellow and green have to be seen as two separate meanings. This relates to Quine's observation (1960): using the example of a linguist studying an exotic tribe and hearing one of the natives exclaim "gavagai" when a rabbit runs by. The linguist can never be sure if gavagai means rabbit, or hopping or even a temporal slice of a four dimensional space-time rabbit. Only through through induction during linguistic interactions can he arrive at the correct meaning.

The communicative success of both approaches does not rise to 100%, but is in any case over 85%. This compares favourable to human performance. In experiments by Lantz and Stefflre (1964) a similar communication task was used to measure the communicatability of colours. In these experiments two groups of subjects were used: the encoders named a colour and the decoders had to pick the colour from an array. Results from Spanish and Yucatan speakers (Stefflre et al., 1966) show that on average the decoders are able to point out the right colour or a colour in the

neighbourhood, but their performance is never perfect.

The results also show that with the current simulations and environment, the nature of the resulting colour categories is stochastic: each population ends up with a categorical repertoire unlike that of other populations. Even so, the environment and ecology have a slight influence on the position of the colour categories in space: when two populations evolve in a similar environment, it is not unthinkable that some similarity will emerge. These results are not in accord with observations of human colour categories, where a strong cross-cultural categorical reaction is found to (almost) the same colour stimuli; leaving some anomalies aside (MacLaury, 1987; Saunders, 1995). As mentioned several times before, there are two opponent explanations for this cross-cultural agreement, both boil down to a nature-nurture discussion of colour categories. The cultural relativist account focuses on the fact that colour categories find their origins in the shared environmental and linguistic experiences of individuals. As individuals experience the world in a largely identical manner across cultures, this explains cross-cultural sharing. Nativists argue that colour categories or the mechanisms leading to the ontogenetical expression of colour categories, are genetically passed on, thus explaining the cross-cultural sharing of colour categories.

The experiments show that the biases present in the simulations on cultural learning are not enough to reach cross-cultural coherence: language and the limited constraints present in the environment and ecology do not arrive at categorical repertoires identical across populations. This however does not invalidate culturalism, adding environmental and ecological constraints will solve our problem and will make different populations converge onto identical categorical repertoires. Though there are undoubtedly many such constraints, to my knowledge no study has yet studied and identified them. Its seems credible that some colours are ecologically more salient than others, for example red is in all cultures associated with danger, blood and fire[4].

As a closing note: if colour categories are indeed culturally learned, this suggests that languages with identical roots or which are geographically close will have approximately the same division of the colour space. Of course, more research on this would be required.

The second simulation is based on genetic evolution of colour categories under influence of language. Category sets are now selected on their ability to support communication about colour. This differs from genetic evolution where the category sets were directly selected on their discriminative power. The categorical repertoires of the agents are shared, as was already the case in the experiments on genetic evolution without language. The experiments demonstrate that linguistic pressure can steer evolution.

[4]But of course, red could also derive its importance from its strong genetic basis.

Implicitly the categories are also fit for distinguishing colour stimuli, as discrimination is a prerequisite for communication. The colour categories are however not shared cross-culturally, again confirming that the environmental and ecological constraints in the simulations are insufficient for explained the shared nature of human colour categories.

8.4 Summary

This chapter shows results of experiment *with* language as described in chapter 5 and 6. Category sets are acquired either through cultural learning or through genetic evolution. The results demonstrate the diachronic evolution of the categorical repertoire and the associated lexicon. The following conclusions can be drawn:

- Both approaches manage to attain a set of categories performing well on the communication task. Cultural learning uses self-organising cultural dynamics, while genetic evolution uses generational reproduction and selection based on the communicative performance of the individuals.
- Both cultural learning and genetic evolution arrive at shared categories within the population. The mechanism by which they do is different: cultural learning adaptively tunes the categories of the individuals to facilitate communication, while sharing in genetic evolution is the result of a one fit genotype spreading in the population, oppressing other less fit solutions.
- Both cultural learning and genetic evolution do not arrive at cross-cultural sharing.
- There are no constraints present, both in cultural learning and genetic evolution with language, to form categories representing fundamental hues (black, white, red, green, yellow and blue) or intermediate hues (orange, purple, grey, pink, brown). In the current model, the unique hues do not have special status.
- The colour categories obtained by both approaches do not bear resemblance to human colour categories, exposing the need for other or more constraints on the environment and the ecology.

Chapter 9

Discussion

This thesis investigated four extreme positions in the discussion on the origins of colour categories. Four computer models were constructed and experiments were run to explore the claims related to these four positions, the models implemented individual learning of colour categories, cultural learning, genetic evolution without cultural interaction, and genetic evolution with cultural interaction.

9.1 Summary

Colour perception has been the topic of many studies in psychology, linguistics and anthropology. All seek to the answer the question of whether colour categories are innate for all humans or whether they are learned. And if colour categories are indeed learned, are they then acquired through a individual learning process (empiricism), or does cultural have a causal influence on the formation of colour categories (culturalism).

Evidence supporting genetic determinism has been found in the opponent-colours theory and in the neurobiology of colour vision. However, the link between neural opponent processing channels and the nature of colour categories obtained through studying their linguistic labels is still debated. Other evidence for universalism comes from naming experiments and memory experiments; there, a striking cross-cultural agreement between the colour categories is found and results show that focal categories are memorised better than others, strengthening the case nativism makes. Relativism challenges these views and results. Among them, empiricists argue that the environment, the ecology and human biology contains enough constraints to learn universal categories, without the need for any innate constraints on the formation of categories. Culturalism admits that there are indeed environmental and ecological constraints, but that these are not enough to explain shared colour categories in cultures. An additional constraint is required in the form of language. Language thus provides

pressure to arrive at coherent meaning structures, making communication possible. So the formation of colour categories, and other perceptual categories, is not an individual process, but it is self-organising process taking place in a population of language users.

These topics have been investigated in this thesis using computer simulations. The basic entity of the simulation is the *agent*, which represents the individual. Each agent is endowed with perception, categorisation, and — in the models where the influence of language is explored— lexicalisation and communication. The colour stimuli are not offered to the agents as abstract symbols but as real stimuli in the form of spectral power distribution measurements of actual colour samples. The perception is modelled as a function that maps the colour stimuli to a representation in an internal space. For the internal colour space I have chosen the CIE LAB space as it is perceptually uniform colour space, lending itself well to define categories on. A colour category is represented by an adaptive network; these are a modification of radial basis function networks known from the field of machine learning. Adaptive networks return a scalar membership value, which is the sum of the reaction of many locally tuned receptors in the internal colour space. The response function of an adaptive network can be shaped using three different methods: (1) they can be trained using an instance-based learning algorithm, (2) their response can be adapted according the feedback gained from a task and (3) they can be evolved in a Darwinian manner. Finally, the lexicalisation associates labels to categories; the strength between a label and a category is represented by a scalar value: the score. The score is changed according to feedback obtained from linguistic interactions.

The simulations exist in four variations. In *individual learning* the agents learn a categorical repertoire through playing discrimination games, in such a game the agent has to make a categorical distinction between one colour stimulus and one or more other colour stimuli. During the game, the agent extends its repertoire with categories representing the observed stimuli. In *cultural learning* a social layer is added to this in the form of guessing games; two agents, one acting as speaker to other as hearer, convey colour meaning using only lexical labels and pointing. The categorical repertoire now forms again through discrimination and, additionally, through adapting the colour categories to facilitate communication. In *genetic evolution* of colour categories, the population of agents is submitted to Darwinian selection process. The fitness is now evaluated on the discriminative performance of the categorical repertoires of the agents. Agents reproduce by copying their categorical repertoire to their offspring with a slight mutation needed to obtain diversity in the population. The fourth simulation implements *genetic evolution with language*; categories are again evolved, but now the selection is performed on the agent's communicative success.

Following observations have been made:

- Both learning and genetic evolution can explain how a repertoire of colour categories emerges; the categories adapted to a set of physiological, environmental and ecological constraints.
- The number of categories produced by learning or evolution is dependant on the complexity of the environment.
- It is theoretically plausible that language has a causal influence on category acquisition. For this a structural coupling is needed between the performance in language games and the formation or adaptation of categories by the agents.
- It is theoretically plausible that language has a causal influence on the evolution of a categorical repertoire. For this the fitness evaluation of an agent has to integrate the agent's communicative success.
- A shared lexicon can self-organise in a population. The learning process must include a positive feedback loop between the success of the word forms in the communication and the selection of word form for expressing meaning.
- Sharing of colour categories is not observed when agents individually learn a repertoire of colour categories. The physiological, environmental and ecological constraints as implemented in the simulations are not enough to drive agents to the same solution.
- Sharing of colour categories is observed when colour categories are genetically evolved. This is due to the nature of genetic evolution, which propagates successful categorical repertoires, so that eventually the population contains the same "colour genes".
- Sharing of colour categories is observed in cultural learning. This is due to the strong cultural coupling between category acquisition and lexicon formation.
- Sharing of categorical repertoires *across* populations is not observed in cultural learning or in genetic evolution. The stochastic nature of the simulation and the lack of strong biological, environmental or ecological constraints causes each population to arrive at a unique solution.
- Genetic evolution operates on a much slower timescale than individual and cultural learning. Genetic evolution takes several generations to adapt to changing environmental conditions, while learning adapts during the lifetime of an individual.
- Cultural learning can cope with a certain degree of fluctuation in the population.

The following points are not demonstrated:

- No comparison has been made between the colour categories produced by the simulations and human colour categories. As the simulations do not produce consistent categorical repertoires (each simulation arrives at a unique solution), comparing the emerged repertoires to human repertoires is not applicable.
- The evolutionary order in which colour terms emerge in human languages (Berlin and Kay, 1969) was not consistently observed in the simulations.

9.2 Critical notes

The claims put forward in this work are based on existence proofs. It has been shown in computer simulations that it is possible to construct circumstances under which we observe the characteristics described in section 9.1. It was not shown that these conclusions hold under all circumstances; there is a large set of parameter settings for the environment, for the ecology or for the cultural interactions for which the observations will not hold. For example, if the communication contains too much noise (in the form of imperfect communication) the agents will not be able to arrive at shared categories using cultural learning. Or when the flux in the population is too high, a shared lexicon will not establish and all observations based on the influence of language will be invalid. However, the conclusion is that the certain phenomena, such as cultural learning of shared perceptual categories, are theoretically possible.

To allow a clear discussion, we have taken extreme positions on the main points of the debate. Obviously, intermediate stances exist as well, but these would only obfuscate the discussion. An example of an intermediate position would be that there is some innate functionality leading to a preference for creating categories at certain locations in the perceptual space. Or, some categories might be innate (for example for the unique colours), while others might be the result of learning influenced by the physiology, ecology and language. These intermediate positions could be explored with the same mechanisms described in this thesis.

The ecology in the simulations is not strong. No colour stimulus has special status (for example by having a higher information content for the observer) or is presented more often than other stimuli. I agree that a different presentation of stimuli —in the sense that some stimuli are presented more often— would produce different outcomes. A non-uniform distribution of colour stimuli indeed has an influence on the learning or genetic evolution of categories. The distribution of the environmental stimuli could be such that a specified set of colour categories would emerge resembling

colour categories found in humans. However, no one has yet been able to identify such a distribution of stimuli or any other environmental constraints. This does not mean that they are nonexistent; they just have not been identified yet, due to the magnitude of the research task.

Next to adding more constraints to the environment, a more constrained model of the neurophysiology of colour perception could also lead to shared colour categories under individual learning. However, these physiological and neurological constraints have not been identified yet. There has been speculation on the influence of neurophysiology on the nature of colour categories (Ratliff, 1976; Kay and McDaniel, 1978; Bornstein, 1987; Hardin, 1988), but this is only based on an assumed link between opponent-colour processing and colour categorisation and colour naming.

Finally, I have taken colour in isolation, but there is more to colour than just chromatic content. Actual object colours have next to chromatic content also specular, spatial and temporal content. Size, shape, texture, lustre, glitter, iridescence, transparency, and so on, of an object might also have an influence on how one perceives colour. I have taken this approach not because it would be impossible to model at least some of these properties, but because the psychological and anthropological literature also uses decontextualised colour in the majority of colour naming and colour memory experiments. The approach taken here allows an uncomplicated comparison with the research from these fields.

9.3 Suggestions for future research

As the model and simulations are restricted in their realism, it might be commendable to try to enhance the realism. First of all, the representation of the categories is open for discussion. The adaptive networks used in this thesis are "parameter-burdened" and the influence of an uncomplicated representation on the dynamics of the simulation might be investigated. This representation might be a Gaussian bell of which the centre and the width can be learned, or a simple point representation. I would expect that the behaviour of the simulation will be largely identical, expect for some specifics such as the non-symmetrical extent of colour categories.

An arguable aspect of the category representation is the updating of the weights of the locally tuned units of the adaptive network point (see 5.1.1). Weights decrease at a constant rate, and are only increased again when the category has been used successfully in a discrimination game. This means that the categories of an agent depend on whether or not they are observed in the environment. Though this works well for the simulations at hand, it is unrealistic (we don't forget a category for RED just because we haven't seen a red stimulus for a while) and could be problematic in simulations with different environmental conditions (i.e. where the stimulus presenta-

tion is more erratic). A solution could be to make the weight dependant on the success of a category; or to inhibit the weight decay depending on the age of the agent, thereby modelling a period where the categorical repertoire is first plastic, but then becomes fixed.

Another addition might be the introduction of multi-word language games (Van Looveren, 2000). Often humans resort to using a concatenation of two colour words when describing an intermediate colour. Implementing this could enhance the realism of the simulation and would probably reduce the number of categories the agents need to successfully communicate colour.

As this work is to my knowledge one of the first using fuzzy meanings defined in a continuous representation space, it would be interesting to see how the dynamics behave under extreme circumstances (i.e. large populations, complex environments). Kaplan (2000) has already demonstrated how large populations are indeed able to agree on a lexicon associated with a fixed set of meanings. But some exploratory experiments on large populations show that as meaning is not clearly delineated, agents often have difficulties in converging towards shared meaning. Adding an age structure to the population —where younger agents are more plastic and learn from older agents— or a geographical distribution —where one agent is allowed to interact with only a few other agents— have produced promising results.

As mentioned before, the simulations have not been able to confirm the evolutionary order proposed Berlin and Kay. There exists the remote possibility that the evolutionary order is an artefact of Berlin and Kay's observations, but if the evolutionary order should indeed be reproducible in the simulations one might try to the change ecology such that the agents form colour categories in a specific order. As shown in the chapters reporting on the results, it is possible to modify the presentation such that certain categories have a higher probability of emerging (an extreme example of this is an ecology which only presents red and green stimuli, which will results in the agents learning one red and one green category). However, many different ecological conditions exist that will all produce an evolutionary order (and many more conditions exist under which no evolutionary order is produced at all). To find which ecological conditions, together with cultural constraints, might produce a certain colour category repertoire or a certain order of development, we could try to build a backward-running experiment. In this experiment the space of possible ecologies could be searched (using an evolutionary computation approach) for the ecologies that are responsible for certain categorical repertoires.

A further suggestion would be to look at colour categories as an optimal distribution in a representational space. Liljencrants and Lindblom (1972) ran computational simulations to see if a simple energy minimization strategy might predict realistic vowel systems. They suspected that the

quality of a vowel system is related to the distinctiveness of vowels. In their simulations a fixed number of vowels indeed spread in the acoustic space until they occupied a minimal energy configuration, just as if magnets floating in basin might push each other away and end up in minimum-energy equilibrium. It would be interesting to see if the same mechanism might be applied to colour categories[1]. The biological constraints are formed by the psychophysical colour space in which the categories are placed. Suppose one uses the CIE LAB space, then two categories would indeed have a minimal energy constellation if they were placed at the location of white and black. However, the model already goes wrong with three categories. Two categories would still represent white and black, but the third category might be anywhere on the maximally saturated rim of the colour space. Psychology states that the third category should represent red. This suggests that either CIE LAB lacks information on the "importance" of colours, or that other ecological or cultural constraints are involved. If a colour space could be found, together with real-world colour data, which does produce emerging colour categories in the evolutionary order of Berlin and Kay, this would make a strong point for empiricism. More research on this would be required.

9.4 Conclusion

This thesis has investigated how colour categories could emerge under different paradigms. It has been argued that individual learning under environmental, ecological and biological constraints as implemented in the simulation is not able to explain colour sharing within a population. However, category acquisition coupled with language through cultural learning can indeed explain colour sharing, but so can genetic evolution of colour categories. However, the genetic encoding of colour categories is implausible as genetic evolution is too slow to adapt to changing environmental conditions, whereas learning clearly outperforms evolution in its power to adapt to new challenges posed by the environment or by culture. Additionally, there are many different solutions to the problem of attaining a colour category repertoire for discriminating and communicating. This again speaks against the universal nature of colour categories, which claims that only one repertoire of colour categories is found in all human cultures. The experiments have also shown that an influence of language on categorisation is possible and even beneficial, thereby confirming the theoretical plausibility of the Sapir-Whorf thesis. However, as simulations merely provide a reflection of reality, these conclusions should be accepted in the light of the obvious limits of simulation. As Oscar Wilde paraphrased: "The truth is rarely pure and never simple."

[1]Thanks to Bart de Boer for the suggestion.

Appendix A

Symbols

$m \pm s$	m is the mean of a set of values, s is the standard deviation computed as $\sqrt{\frac{n\sum x^2 - (\sum x)^2}{n^2}}$.
$\overline{x}$	A value averaged over the population.
$\{a_1, ..., a_n\}$	The set containing elements a_1 up to a_n. Each element can occur only once in a set and the elements are unordered.
$\langle a_1, ..., a_n \rangle$	The sequence containing elements a_1 up to a_n. An element can occur more than once and all elements are ordered.
$\|S\|$	Number of elements in set S.
$\|a - b\|$	Distance metric between to points a and b.
α	Factor for decreasing the weight of a hidden unit.
A	The agent A.
$\mathcal{A}$	A population of agents.
a_f	The number of agents having word form f.
β	Factor for increasing the weight of a hidden unit.
CS	Communicative success of an agent.
$\overline{CS}$	Average communicative success of all agents in the population.
CV	Category variance of the population.
$\overline{CS}$	Average communicative success of all agents in the population.
δ_{increase}	Amount with which the score of a word form is increased.
δ_{decrease}	Amount with which the score of a word form is decreased.
D	Euclidian distance between two sets.
D_{set}	Distance measure between two point sets.
$D_{\text{categoryset}}$	Distance measure between two sets of categories.
DS	Discriminative success of an agent.
$\overline{DS}$	Average discriminative success of all agents in the population.
f	A word form.
F	A set of word forms $F = \{f_1, \ldots, f_{\|F\|}\}$

IV_f	Interpretation variance of word form f.
$\overline{IV}$	Average interpretation variance of all word forms in the population.
$\mathbf{m}_{c,i}$	The centre of the i-th hidden unit belonging to the adaptive network c.
N	A discrete number.
o	An object or stimulus.
O	The context offered to the agents, $O = \langle o_1, \ldots, o_N \rangle$ is a sequence containing N stimuli or objects o_i.
p_{cf}	Probability of creating a new form.
p_{lf}	Probability of learning a form.
p_{mut}	Probability of one of the mutation operators being chosen.
σ	Width of the locally tuned unit of an adaptive network.
σ_{default}	Default value for the width σ of locally tuned unit.
σ_{mut}	Standard deviation of the gaussian random distribution used in mutating colour categories.
s	Internal sensory representation of object o.
$score_{\text{default}}$	Initial score of a word form when it is created or learned.
τ_{adapt}	Threshold used in a discrimination game for deciding if a category should be adapted to be more discriminative or if a new category should be created.
τ_{merge}	Threshold deciding when two categories are close enough together to merge.
θ_{lastused}	Threshold for removing a word form if it has not been used for a certain number of games.
$\theta'_{\text{lastused}}$	Threshold for removing a category if it has no been used for a certain number of games.
θ_{score}	Threshold for removing a word form if its score is lower than a certain threshold.
θ_{wfage}	Age threshold for removing a word form.
T	Number of games or generations in a simulation.
T_a	Number of games over which a running average is taken.
T_f	Number of games used to evaluate fitness of an individual.

Bibliography

Allott, R. M. (1974). Some apparant uniformities between languages in colour-naming. *Language and Speech*, 17(4):377–402.

Arbib, M. A. (1995). *The Handbook of Brain Theory and Neural Networks*. The MIT Press, Cambridge, MA.

Armstrong, S. L., Gleitman, L. R., and Gleitman, H. (1983). What some concepts might not be. *Cognition*, 13:263–308.

Baldwin, D. A. (1993). Infant's ability to consult the speaker for clues to word reference. *Journal of Child Language*, 2:395–418.

Batali, J. (1999). Computational simulations of the emergence of grammar. In Hurford, J. and Studdert-Kennedy, M., editors, *Approaches to the evolution of language: social and cognitive bases*. Cambridge University Press, Cambridge.

Belpaeme, T. (1998). Tracking objects using an active camera. In La Poutré, H. and van den Herik, J., editors, *Proceedings of the 10th Belgium-Netherlands Conference on Artificial Intelligence (BNAIC98), Amsterdam, The Netherlands*.

Belpaeme, T. (1999). Evolution of visual feature detectors. In *Proceedings of the First European Workshop on Evolutionary Computation in Image Analysis and Signal Processing (EvoIASP99, Göteborg, Sweden)*, University of Birmingham School of Computer Science.

Belpaeme, T. (2001a). Reaching coherent color categories through communication. In Kröse, B., editor, *Proceedings of the 13th Belgium-Netherlands Conference on Artificial Intelligence (BNAIC01), Amsterdam, The Netherlands*, pages 41–48.

Belpaeme, T. (2001b). Simulating the formation of color categories. In Nebel, B., editor, *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'01)*, pages 393–398, Seattle, WA. Morgan Kaufmann, San Francisco, CA.

Belpaeme, T. and Birk, A. (1997a). On the watch. In *Proceedings of the 30th International Symposium on Automotive Technology and Automation conference (ISATA 97), Florence, Italy*. ISATA Press.

Belpaeme, T. and Birk, A. (1997b). A real-world ecosystem featuring several robot species. Technical Report 97-07, Artificial Intelligence Lab, Vrije Universiteit Brussel. Presented at the European Conference on Artificial Life (ECAL97).

Belpaeme, T. and Birk, A. (2001). Hungry robots. *ACM Crossroads*, 8(2).

Belpaeme, T., Steels, L., and van Looveren, J. (1998). The construction and acquisition of visual categories. In Birk, A. and Demiris, J., editors, *Proceedings of the 6th European Workshop on Learning Robots, Lecture Notes on Artificial Intelligence*, Lecture Notes on Artificial Intelligence, Berlin. Springer.

Berlin, B. and Kay, P. (1969). *Basic Color Terms: Their Universality and Evolution*. University of California Press, Berkeley, CA.

Bhaskar, R., editor (1978). *A Realist Theory of Science*. Hemel Hempstead, UK, Harvester Wheatsheaf, 2nd edition.

Billard, A. and Dautenhahn, K. (2000). Experiments in social robotics: grounding and use of communication in autonomous agents. *Adaptive behavior*, 7(3/4).

Birk, A. and Belpaeme, T. (1998). A multi agent system based on heterogeneous robots. In Drogoul, Tambe, and Fukuda, editors, *Proceedings of the Collective Robotics Workshop*, Lecture notes on Artificial Intelligence 1456. Springer, Berlin.

Birk, A., Walle, T., Belpaeme, T., and Kenn, H. (1999). The VUB AI-lab robocup'99 small league team. In Veloso, Pagello, and Kitano, editors, *Robocup'99: Robot Soccer World Cup II*, Lecture Notes on Artificial Intelligence. Springer, Berlin.

Birk, A., Walle, T., Belpaeme, T., Parent, J., De Vlaminck, T., and Kenn, H. (1998). The small league robocup team of the VUB AI-lab. In Asada and Kitano, editors, *Proceedings of the 2nd Robocup Workshop*, Lecture notes on Artificial Intelligence 1604. Springer, Berlin.

Blackmore, S. (1999). *The meme machine*. Oxford University Press, Oxford.

Bornstein, M. H. (1975). The influence of visual perception on culture. *American Anthropologist*, 77:774–798.

Bornstein, M. H. (1987). Perceptual categories in vision and audition. In Harnad, S., editor, *Categorical Perception*, pages 287–300. Cambridge University Press, Cambridge.

Bornstein, M. H., Kessen, W., and Weiskopf, S. (1976). Color vision and hue categorization in young human infants. *Journal of Experimental Psychology*, 2:115–129.

Boynton, R. M. (1990). Human color perception. In Leibovic, K., editor, *Science of Vision*, pages 211–253. Springer Verlag, New York.

Brainard, D. H. (2001). Color appearance and color difference specification. In Shevell, D., editor, *The Science of Color*. To appear.

Braitenberg, V. (1984). *Vehicles: Experiments in Synthetic Psychology*. The MIT Press, Cambridge, MA.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, 46:139–159.

Broomhead, D. S. and Lowe, D. (1988). Multivariate functional interpolation and adaptive networks. *Complex Systems*, 2:321–355.

Brown, R. W. (1976). Reference, in memorial tribute to Eric Lenneberg. *Cognition*, 4:125–153.

Brown, R. W. and Lenneberg, E. H. (1954). A study in language and cognition. *Journal of Abnormal and Social Psychology*, 54:454–462.

Bullock, S. (1997). *Evolutionary Simulation Models: On their character, and Application to Problems Concerning the Evolution of Natural Signalling Systems*. PhD thesis, School of Cognitive and Computing Sciences, University of Sussex.

Burnham, R. W. and Clark, J. R. (1955). A test of hue memory. *The Journal of Applied Psychology*, 39(3):164–172.

Cairns, J., Overbaugh, J., and Miller, S. (1988). The origins of mutants. *Nature*, 335:142.

Callaghan, T. (1984). Dimensional interaction of hue and brightness in preattentive field segregation. *Perceptual Psycophysics*, 36:25–34.

Camazine, S., Deneubourg, J., Franks, N., Sneyd, J., Theraulaz, G., and Bonabeau, E. (2001). *Self-Organization in Biological Systems*. Princeton University Press, Princeton.

Casson, R. W. (1997). Color shift: evolution of english color terms from brightness to hue. In Hardin, C. L. and Maffi, L., editors, *Color Categories in thought and language*, pages 224–240. Cambridge University Press.

Chandler, D. (2001). *Semiotics: the Basics*. Routledge, London. Also available at http://www.aber.ac.uk/media/Documents/S4B/semiotic.html.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. The MIT Press, Cambridge, MA.

Clark, H. H. and Clark, E. V. (1977). *Psychology and Language: An introduction to psycholinguistics*. Harcourt Brace.

Corbett, G. and Morgan, G. (1988). Colour terms in Russian: Reflections of typological constraints in a single language. *Journal of Linguistics*, 24:31–64.

Cornsweet, T. N. (1970). *Visual Perception*. Academic Press, New York.

Cotter, J. R. (1990). The visual pathway: an introduction to structure and organization. In Leibovic, K., editor, *Science of vision*. Springer, New York.

Crystal, D. (1997). *The Cambridge Encyclopedia of Language*. Cambridge University Press, Cambridge, 2nd edition.

Darwin, C. (1859). *The Origin of Species*. Wordsworth, Hertfordshire, UK. 1998 edition.

Davidoff, J. (1991). *Cognition through color*. The MIT press, a Bradford book, Cambridge, MA.

Davidoff, J. (2001). Language and perceptual categorisation. *Trends in Cognitive Sciences*, 5(9):382–387.

Davidoff, J., Davies, I., and Roberson, D. (1999). Colour categories in a stone-age tribe. *Nature*, 398:203–204.

Dawkins, R. (1976). *The Selfish Gene*. Oxford University Press, Oxford.

de Boer, B. (2001). *The origins of vowel systems*. Oxford University Press, Oxford, UK.

de Jong, E. (1999). Autonomous concept formation. In Dean, T., editor, *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI99)*, pages 344–349. Morgan Kaufmann, San Francisco, CA.

de Jong, E. and Vogt, P. (1998). How should a robot discriminate between objects? In *Proceedings of the Fifth International Conference of the Society for Adaptive Behavior (SAB'98)*, Cambridge, MA. The MIT Press.

de Saussure, F. (1974). *Course in General Linguistics*. Fontana/Collins, New York. Manuscript from 1916, translated by Baskin, W.

De Valois, R., Abramov, I., and Jacobs, G. (1966). Analysis of response patterns of LGN cells. *Journal of the Optical Society of America*, 56(7):966–977.

Deacon, T. W. (1997). *The symbolic species: the co-evolution of language and the brain*. W.W. Norton, New York.

Dedrick, D. (1998). *Naming the rainbow: Colour language, colour science, and culture*, volume 274 of *Synthese Library*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

DeMonasterio, F. M. and Gouras, P. (1975). Functional properties of ganglion cells of the rhesus monkey retina. *Journal of Physiology*, 251:39–48.

Derrington, A., Krauskopf, J., and Lennie, P. (1982). Chromatic mechanisms in lateral geniculate nucleus of the macaque. *Journal of Physiology*, 357:241–265.

Di Paolo, E., Noble, J., and Bullock, S. (2000). Simulation models as opaque thought experiments. In *Proceedings of Artificial Life VII: the Seventh International Conference on the Simulation and Synthesis of Living Systems*. Reed College, Portland, Oregon.

Dournes, J. (1978). Les races de couleurs. In Tornay, S., editor, *Voir et Nommer les Couleurs*. Laboratoire d'Ethnologie et de Sociologie Comparative, Nantes.

Dowling, J. (1987). *The Retina: an approachable part of the brain*. The Bellknap Press of Harvard University Press, Cambridge.

Drescher, G. L. (1991). *Made-up Minds: a constructivist approach to artificial intelligence*. The MIT Press, Cambridge, MA.

Dubois, D. (2000). Categories as acts of meaning: the case of categories in olfaction and audition. *Cognitive Science Quarterly*, 1:33–66.

Durbin, M. (1972). Basic terms - off-color? *Semiotica*, 6:257–278.

Durham, W. H. (1991). *Coevolution: Genes, Culture and Human Diversity*. Stanford University Press, Stanford.

Edelman, S. (1999). *Representation and Recognition in Vision*. The MIT Press, a Bradford book, Cambridge, MA.

Eiter, T. and Mannila, H. (1997). Distance measures for point sets and their computation. *Acta Informatica*, 34(2):109–133.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., and Plunkett, K. (1996). *Rethinking Innateness: A Connectionist Perspective on Development*. The MIT Press, Cambridge, MA.

Etcoff, N. and Magee, J. (1992). Categorical perception of facial expressions. *Cognition*, 44:227–240.

Fairchild, M. (1998). *Color Appearance Models*. Addison-Wesley, Reading, MA.

Ferber, J. (1998). *Multi-agent systems: an introduction to distributed artificial intelligence*. Addison-Wesley, Reading, MA.

Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. Clarendon Press, Oxford, UK.

Fogel, D. B., editor (1998). *Evolutionary Computation: The fossil record*. IEEE Press, New York.

Fogel, L. J. (1999). *Intelligence Through Simulated Evolution: Forty years of evolutionary programming*. Wiley Series on Intelligent Systems. John Wiley and sons, New York.

Gegenfurtner, K. R. (1999). Reflections on colour constancy. *Nature*, 402:855–856.

Gellatly, A. (1995). Colourful whorfian ideas: Linguistic and cultural influences on the perception and cognition of colour, and on the investigation of them. *Mind and Language*, 10(3):199–225.

Ghosh, J. and Nag, A. (2000). Radial basis function networks. In Howlett, R. and Jain, L., editors, *Radial Basis Function Neural Network Theory and Applications*. Physica-Verlag.

Gleason, H. (1961). *An Introduction to Descriptive Linguistics*. Holt, Rinehart and Winston, New York.

Gouras, P. (1984). Color vision. In Osborn, N. and Chader, J., editors, *Progress in retinal research*. Pergamon Press, Oxford.

Hardin, C. and Maffi, L., editors (1997). *Color categories in thought and language*. Cambridge University Press, Cambridge.

Hardin, C. L. (1988). *Color for Philosophers: Unweaving the Rainbow*. Hacket, Indianapolis.

Hassoun, M. (1995). *Fundamentals of Artificial Neural Networks*. The MIT Press, Cambridge, MA.

Hawkins, J. A. (1992). Innateness and function in language universals. In Hawkins, J. A. and Gell-Mann, M., editors, *The Evolution of Human Languages*, Santa Fe Institute, Studies in the Sciences of Complexity, pages 87–120. Addison-Wesley.

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

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

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

Holland, J. H. (1975). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. University of Michigan Press, Ann Arbor.

Hurford, J. (1989). Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua*, 77(2):187–222.

Jackendoff, R. (1992). *Languages of the mind*. The MIT Press, Cambridge, MA.

Jackendoff, R. (1993). *Patterns in the mind: Language and human nature*. Harvester Wheatsheaf (Paramount Publishing).

Jameson, D. and Hurvich, L. M. (1955). Some quantitative aspects of an opponent-colors theory. I. Chromatic responses and spectral saturation. *Journal of the Optical Society of America*, 45(7):546–552.

Jennings, N. and Wooldridge, M. (1998). *Agent Technology: Foundations, Application and Markets*. Springer Verlag.

Jolliffe, I. (1986). *Principal component analysis*. Springer, New York.

Kaiser, P. and Boynton, R. (1996). *Human Color Vision*. Optical Society of America, Washington DC.

Kaplan, F. (2000). *L'émergence d'un lexique dans une population d'agents autonomes*. PhD thesis, Université de Paris VI, France.

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

Kay, P. and Kempton, W. (1984). What is the Sapir-Whorf hypothesis? *American Anthropologist*, 86(1):65–79.

Kay, P. and McDaniel, C. (1978). The linguistic significance of the meanings of basic color terms. *Language*, 54(3):610–646.

Kirby, S. (1999). *Function, selection and innateness: the emergence of language universals*. Oxford University Press, Oxford.

Koza, J. R. (1992). *Genetic programming: on the programming of computers by means of natural selection*. The MIT Press, Cambridge, MA.

Krogh, A. and Hertz, J. A. (1995). A simple weight decay can improve generalization. In Moody, J., Hanson, S., and Lippmann, R., editors, *Advances in Neural Information Processing Systems 4*, pages 950–957. Morgan Kauffmann, San Mateo, CA.

Laakso, A. and Cottrell, G. (2000). Content and cluster analysis: assessing representational similarity in neural systems. *Philosophical Psychology*, 13(1).

Lammens, J. M. (1994). *A computational model of color perception and color naming*. PhD thesis, State University of New York.

Langton, C., editor (1989). *Artificial Life: Proceedings of the Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems (ALIFE'87)*. Addison-Wesley, Redwood City, CA.

Lantz, D. and Stefflre, V. (1964). Language and cognition revisited. *Journal of Abnormal and Social Psychology*, 69(5):472–481.

Lee, S. and Kil, R. (1988). Multilayer feedforward potential function networks. In *Proceedings of the IEEE Second International Conference on Neural Networks (San Diego)*, volume I, New York. IEEE Press.

Leibovic, K., editor (1990). *Science of Vision*. Springer Verlag, New York.

Lenneberg, E. H. (1961). Color naming, color recognition, color discrimination: A re-appraisal. *Perceptual Motor Skills*, 12:375–382.

Lenneberg, E. H. and Roberts, J. M. (1956). The language of experience: A study in methodology. *International Journal of American Linguistics*, memoir 13.

Lenz, R., Osterberg, M., Hiltunen, J., Jaaskelainen, T., and Parkkinen, J. (1996). Unsupervised filtering of color spectra. *Journal of the Optical Society of America*, 13(7):1315–1324.

Levine, M. (2000). *Fundamentals of sensation and perception*. Oxford University Press, Oxford, 3rd edition.

Liljencrants, L. and Lindblom, B. (1972). Numerical simulations of vowel quality systems. *Language*, 48:839–862.

Lucy, J. A. (1996). The scope of linguistic relativity: an analysis and review of empirical research. In Gumperz, J. J. and Levinson, S. C., editors, *Rethinking linguistic relativity*, Studies in the Social and Cultural Foundations of Language 17. Cambridge University Press, Cambridge.

Lucy, J. A. (1997). The linguistics of "color". In Hardin, C. L. and Maffi, L., editors, *Color Categories in Thought and Language*, pages 320–346. Cambridge University Press, Cambridge.

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

MacLaury, R. E. (1987). Color-category evolution and shuswap yellow-with-green. *American Anthropologist*, 89:107–124.

MacLaury, R. E. (1992). From brightness to hue: An explanatory model of color-category evolution. *Current Anthropology*, 33(2):137–186.

Medgassy, P. (1961). *Decomposition of Superposition of Distributed Functions*. Hungarian Academy of Sciences, Budapest.

Medin, D. (1989). Concepts and conceptual structure. *American Psychologist*, 44:1469–1481.

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill, New York.

Moody, J. and Darken, C. (1989). Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1(2):281–294.

Munsell (1976). *Munsell book of color, matte finish collection*. Munsell Color Company, Baltimore, MD.

Nathans, J. (1989). The genes for color vision. *Scientific American*, 260(2):28–35.

Nathans, J. (1999). The evolution and physiology of human color vision: Insights from molecular genetic studies of visual pigments. *Neuron*, 24:299–312.

Neitz, J., Neitz, M., He, J., and Shevell, S. (1999). Trichromatic color vision with only two spectrally distinct photopigments. *Nature Neuroscience*, 2(10):884–888.

Newhall, S., Nickerson, D., and Judd, D. (1943). Final report of the O.S.A. subcommittee on spacing of the Munsell colors. *Journal of the Optical Society of America*, 33:385–412.

Nickerson, D. (1940). History of the Munsell color system and its scientific application. *Journal of the Optical Society of America*, 30:575–586.

Noble, J. (1998). *The Evolution of Animal Communication Systems: Questions of Function Examined through Simulation*. PhD thesis, School of Cognitive and Computing Sciences, University of Sussex.

Nowak, M., Komarova, N., and Niyogi, P. (2001). Evolution of universal grammar. *Science*, 291:114–118.

Oliphant, M. (1996). The dilemma of saussurean communication. *BioSystems*, 37(1-2):31–38.

Oliphant, M. and Batali, J. (1997). Learning and the emergence of coordinated communication. *The newsletter of the Center of Research in Language*, 11(1).

Parkkinen, J., Hallikainen, J., and Jaaskelainen, T. (1989). Characteristic spectra of Munsell colors. *Journal of the Optical Society of America*, 6(2):318–322.

Parkkinen, J., Jaaskelainen, T., and Kuittinen, M. (1988). Spectral representation of color images. In *Proceedings of the Ninth International Conference on Pattern Recognition*. IEEE.

Peirce, C. S. (1931-1958). *Collected Writings*, volume 1 - 8. Harvard University Press, Cambridge, MA.

Pfeifer, R. and Scheier, C. (1999). *Understanding Intelligence*. The MIT Press, Cambdrige, MA.

Piaget, J. (1977). *The essential Piaget*. Routledge and Kegan Paul, London. Edited by Gruber, H.E. and Vonéche, J.

Pinker, S. (1994). *The language instinct: How the mind creates language*. W. Morrow, New York.

Poggio, T. and Hurlbert, A. (1994). Observations on cortical mechanisms for object recognition and learning. In Koch, C. and Davis, J., editors, *Large-scale Neuronal Theories of the Brain*, pages 153–182. The MIT Press, Cambridge, MA.

Quine, W. (1960). *Word and Object*. The MIT Press, Cambridge, MA.

Quinlan, J. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Francisco, CA.

Ramon, J. and Bruynooghe, M. (2001). A polynomial time computable metric between point sets. *Acta Informatica*. To appear.

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

Ratner, C. (1989). A sociohistorical critique of naturalistic theories of color perception. *Journal of Mind and Behavior*, 10:361–372.

Regan, B., Julliot, C., Simmen, B., Viénot, Charles-Dominique, P., and Mollon, J. (2000). Fruits, foliage and the evolution of primate colour vision. *Philosophical Transactions of the Society of London*, 356:229–283.

Reynolds, C. (1987). Flocks, herds, and schools: a distributed behavioral model. *Computer Graphics*, 21(4):25–34.

Rivers, W. H. R. (1901). Colour vision. In *Reports of the Cambridge anthropological expedition to Torres Straits. Volume 2. Physiology and psychology*. Cambridge University Press, Cambridge. Also published as Rivers, W.H.R. (1901) Primitive Color Vision. *Popular Science Monthly*, 59:44-58.

Rivers, W. H. R. (1903). Observations on the vision of the Urális and Sholagas. *Bulletin of the Madras Government Museum*, 5:3–18.

Rivers, W. H. R. (1905). Observations on the senses of the Todas. *British Journal of Psychology*, 1(4):321–396.

Roberson, D., Davies, I., and Davidoff, J. (2000). Color categories are not universal: replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, 129(3):369–398.

Rosch, E. (1978). Principles of categorization. In Rosch, E. and Lloyd, B., editors, *Principles of categorisation, in cognition and categorisation*, pages 27–48. Erlbaum, Hillsdale, NJ.

Rosch, E. and Lloyd, B. (1978). *Principles of categorisation, in cognition and categorisation*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Rosch, E., Mervis, C., Gray, W., Johnson, D., and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8:382–439.

Rumelhart, D. and McClelland, J. (1986). *Parallel Distributed Processing: Exploration in the microstructure of cognition*. The MIT Press, Cambridge, MA. Volume 1 and 2.

Sacks, O. (1996). *The island of the colorblind*. Picador Books, London.

Sahlins, M. (1976). Colors and cultures. *Semiotica*, 16:1–22.

Sandell, J. H., Gross, C. G., and Bornstein, M. H. (1979). Color categories in macaques. *Journal of Comparative and Physiological Psychology*, 93(4):626–635.

Saunders, B. (1995). Disinterring basic color terms: a study in the mystique of cognitivism. *History of the Human Sciences*, 8(4):19–38.

Saunders, B. and van Brakel, J. (1997). Are there nontrivial constraints on colour categorization? *Behavioral and Brain Sciences*, 20(2):167–228.

Savage-Rumbaugh, S. E. (1986). *Ape Language: From Conditioned Response to Symbol*. Columbia University Press, New York.

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science*, 237:1317–1323.

Shepard, R. N. (1994). Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin & Review*, 1:2–28. Reprinted in Behavioral and Brain Sciences, 24(3).

Slaughter, M. (1990). The vertebrate retina. In Leibovic, K., editor, *Science of Vision*, pages 53–83. Springer Verlag, New York.

Smith, A. D. (2001). Establishing communication systems without explicit meaning transmission. In Kelemen, J. and Sosik, P., editors, *Proceedings of the European Conference on Artificial Life (ECAL01), Prague*, pages 381–390, Berlin. Springer.

Steels, L. (1996a). Perceptually grounded meaning creation. In Tokoro, M., editor, *Proceedings of the International Conference on Multiagent Systems (ICMAS-96)*, pages 338–344, Menlo Park, CA. AAAI Press.

Steels, L. (1996b). Self-organizing vocabularies. In Langton, C., editor, *Proceedings of the Conference on Artificial Life V (Alife V) (Nara, Japan)*.

Steels, L. (1997a). Construction and sharing perceptual distinctions. In van Someren, M. and Widmer, G., editors, *Proceedings of the European Conference on Machine Learning*, Berlin. Springer Verlag.

Steels, L. (1997b). The origins of syntax in visually grounded robotic agents. In Pollack, M., editor, *Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI97) (Los Angeles, California)*, San Francisco, CA. Morgan Kauffman Publishers.

Steels, L. (1997c). The synthetic modeling of language origins. *Evolution of Communication*, 1(1):1–34.

Steels, L. (1998a). The origins of ontologies and communication conventions in multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 1:169–194.

Steels, L. (1998b). The origins of syntax in visually grounded robotic agents. *Artificial Intelligence*, 103:1–24.

Steels, L. (1999). The talking heads experiment. Available from the VUB Artificial Intelligence Laboratory, Brussels, Belgium.

Steels, L. (2001a). The methodology of the artificial. *Behavioral and Brain Sciences*, 24(6). A reply to Webb, B. (2001) Can robots make good models of biological behavior? *Behavioral and Brain Sciences*, 24(6).

Steels, L. (2001b). Social learning and language acquisition. In McFarland, D. and Holland, O., editors, *Social robots*. Oxford University Press, Oxford, UK.

Steels, L. and Belpaeme, T. (2002). Computational simulations of colour categorisation and colour naming. In preparation.

Steels, L. and Kaplan, F. (1998). Stochasticity as a source of innovation in language games. In Adami, G., Belew, R., Kitano, H., and Taylor, C., editors, *Proceedings of the Conference on Artificial Life VI (Alife VI) (Los Angeles, California)*, Cambridge, MA. The MIT Press.

Steels, L. and Kaplan, F. (1999a). Bootstrapping grounded word semantics. In Briscoe, T., editor, *Linguistic evolution through language acquisition: formal and computational models*. Cambridge University Press, Cambridge.

Steels, L. and Kaplan, F. (1999b). Situated grounded word semantics. In Dean, T., editor, *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI'99) (San Francisco, CA)*, San Francisco, CA. Morgan Kauffman Publishers.

Steels, L. and Kaplan, F. (2002). AIBO's first words: The social learning of language and meaning. *Evolution of Communication*, 4(1).

Steels, L., Kaplan, F., McIntyre, A., and Van Looveren, J. (2002). Crucial factors in the origins of word-meaning. In Wray, A., editor, *The Transition to Language*. Oxford University Press, Oxford, UK.

Stefflre, V., Vales Castillo, V., and Morley, L. (1966). Language and cognition in yucatan: a cross-cultural replication. *Journal of Personality and Social Psychology*, 4(1):112–115.

Stengers, I. and Prigogine, I. (1986). *Order Out of Chaos*. Bantam Books, New York.

Sumner, P. and Mollon, J. (2000). Catarrhine photopigments are optimized for detecting targets against a foliage background. *Journal of Experimental Biology*, 203:1963–1986.

Tomasello, M. (1988). The role of joint attention in early language development. *Language Sciences*, 11:69–88.

Tomasello, M. (1999). *The cultural origins of human cognition*. Harvard University Press, Cambridge, MA.

Tomasello, M. and Barton, M. (1994). Learning words in nonostensive contexts. *Journal of Developmental Psychology*, 30(5):639–650.

van Brakel, J. (1993). The plasticity of categories: the case of colour. *British Journal for the Philosophy of Science*, 44:103–135.

Van Looveren, J. (2000). An analysis of multiple-word naming games. In Van den Bosch, A. and Wiegand, H., editors, *Proceedings of the 12th Belgium-Netherlands Conference on Artificial Intelligence (BNAIC00), Kaatsheuvel, the Netherlands*.

Vautin, R. and Dow, B. (1985). Color cell groups in foveal striate cortex of the behaving macaque. *Journal of Neurophysiology*, 54:273–292.

Vogt, P. (2000). *Lexicon grounding on mobile robots*. PhD thesis, Artificial Intelligence Lab, Vrije Universiteit Brussel, Belgium.

Vogt, P. (2001). Evolution of grounded symbols in communicating robots. *Evolution of Communication*, 4(2). To appear.

von Kries, J. (1902). Chromatic adaptation. In MacAdam, D., editor, *Sources of Color Science*. The MIT Press, Cambridge, MA. (1970).

Wachtler, T., Lee, T.-W., and Sejnowski, T. (2001). The chromatic structure of natural scenes. *Journal of the Optical Society of America*, 18(1):65–77.

Webster, M. A., Miyahara, E., Malkoc, G., and Raker, V. E. (2000). Variations in normal color vision. II. Unique hues. *Journal of the Optical Society of America A*, 17(9):1545–1555.

Whorf, B. L. (1940). Science and linguistics. *Technology review*, 42(6):229–231, 247–248.

Whorf, B. L. (1956). *Language, Thought and Reality: selected writings of Benjamin Lee Whorf*. The MIT Press, Cambridge, MA. Edited by Carrol, J.B.

Wilson, R. A. and Keil, F. C. (1999). *The MIT Encyclopedia of the Cognitive Sciences*. The MIT Press, Cambridge, MA.

Wooldridge, M. and Jennings, N. (1995). Intelligent agents: Theory and practice. *Knowledge Engineering Review*, 2(10).

Worden, R. (1995). A speed limit for evolution. *Journal of Theoretical Biology*, 176:137–152.

Wurm, L., Legge, E., Isenberg, L., and Luebker, A. (1993). Color improves object recognition in normal and low vision. *Journal of Experimental Psychology: Human Perception and Performance*, 19:899–911.

Wyszecki, G. and Stiles, W. (1982). *Color Science: Concepts and Methods, Quantitative Data and Formulae*. John Wiley and sons, New York, 2nd edition. Reprinted in 2000.

Yanco, H. and Stein, L. (1993). An adaptive communication protocol for cooperating mobile robots. In Meyer, J., Roitblat, H., and Wilson, S., editors, *From Animals to Animats 2: Proceedings of the Second International Conference on Simulation of Adaptive Behavior*, pages 478–485, Cambridge, MA. The MIT Press.

Zeki, S. (1983). Colour coding in the cerebral cortex: the reaction of cells in monkey visual cortex to wavelenghts and colours. *Neuroscience*, 9(4):741–765.

Zeki, S. (1993). *A Vision of the Brain*. Blackwell Scientific Publications, Oxford, UK.

Zipf, G. K. (1929). Relative frequency as a determinant of phonetic change. In *Harvard Studies in Classical Philology*, 15.

Zollinger, H. (1988). Biological aspects of color naming. In Rentschler, I., Herzberger, B., and Epstein, D., editors, *Beauty and the Brain: Biological Aspects of Aesthetics*, pages 149–164. Birkhäuser, Basel.

Zuidema, W. H. (2001). Emergent syntax: the unremitting value of computational modeling for understanding the origins of complex language. In Kelemen, J. and Sosík, P., editors, *Advances in Artificial Life (Proceedings 6th European Conference on Artificial Life, Prague)*, volume 2159 of *Lecture Notes in Computer Science*, pages 641–644. Springer, Berlin.

Index