Coordinating Perceptually Grounded Categories through Language. A Case Study for Colour.

Luc Steels[1,2] and Tony Belpaeme[1]

(1) Vrije Universiteit Brussel Artificial Intelligence Laboratory

Pleinlaan 2 - 1050 Brussels

(2) SONY Computer Science Laboratory - Paris.

Abstract

The paper proposes a number of models to examine through what mechanisms a population of autonomous agents could arrive at a repertoire of perceptually grounded categories that is sufficiently shared to allow successful communication. The models are inspired by the main approaches to human categorisation being discussed in the literature: nativism, empiricism, and culturalism. Colour is taken as a case study. Although the paper takes no stance on which position is to be accepted as final truth with respect to human categorisation and naming, it points to theoretical constraints that make each position more or less likely and contains clear suggestions on what the best engineering solution would be. Specifically, it argues that the collective choice of a shared repertoire must integrate multiple constraints, including constraints coming from communication.

Keywords: autonomous agents, symbol grounding, colour categorisation, colour naming, genetic evolution, connectionism, memes, cultural evolution, self-organisation, origins of language, semiotic dynamics.

1 Introduction

This paper is about how a perceptually grounded categorical repertoire can become sufficiently shared among the members of a population to allow successful communication. For example, how colour categories like 'red' or 'purple' may become sufficiently shared so that one agent from the population can use the word "red" to get another agent to pick out a red object from a set of coloured objects in a scene.

Our own goal is entirely practical. We want to find out how to design artificial embodied agents (robots) such that they are able to do this task. Although the artificial agents might end up with a quite different categorial repertoire compared to human beings, it is intriguing and challenging to investigate under what circumstances they would arrive at human-like solutions, as this would enable communication with humans. Because the agents will be considered to be autonomous and distributed, we cannot assume telepathy nor central control. Because the real world environments in which they will find themselves will be assumed to be open-ended and unknown at design-time (perhaps the agents are to be sent to a distant planet), we cannot program into the agents a specific repertoire of categories because that would make them unable to adapt to new or unknown circumstances. Moreover it is known to be very difficult, if not impossible, to ground categories in sensory-motor patterns by hand (Harnad, 1990), so some form of learning or evolution will be unavoidable.

It seems a good idea to take as much inspiration as possible from categorisation and naming by humans because that is the only and most impressive natural system achieving shared perceptually grounded categorisation and communication based on a rich open-ended repertoire of categories. Moreover if we can generate categorical repertoires that are similar to those of humans it will make communication between humans and artificial agents more feasible. The question how a population might coordinate their perceptually grounded categories and negotiate a shared set of linguistic conventions to express them is relevant to the computational modeling of the origins of language and meaning, which is receiving increased attention lately (Cangelosi and Parisi, 2001; Briscoe, 2002) and has important applications in man-machine interaction.

With respect to human beings, it is generally acknowledged that human physical embodiment plays a significant role. But it is also clear that this does not yet constrain sufficiently the set of possible categories an agent might utilise to cope with the world. Three approaches have been suggested to aid the coordination of categories over and above the constraints given by embodiment:

- **Approach 1. Nativism.** All humans could be born with the same perceptually grounded categories, as part of their 'mentalese'. So when children learn a language, their categorical repertoire is already shared with that of caregivers and they only have to learn the names of these categories. No influence of language on category formation is deemed to be necessary. Assuming innate perceptual categories implies that the neural mechanisms performing categorisation must be genetically determined and the relevant genes must have evolved through evolution by natural selection. This position is historically associated with rationalism (Fodor, 1983) and often found explicitly or implicitly in evolutionary psychology (Pinker and Bloom, 1990; Durham, 1991; Shepard, 1994). Adopting this position for the design of artificial agents means that we must simulate genetic evolution (Holland, 1975; Goldberg, 1989; Koza, 1992; Fogel, 1999). Agents could be given a genome that determines (through some developmental process) how they categorise the world. We could then use success in communication as the selection pressure acting in artificial evolution, and after some period of time, agents should have zoomed in on a shared set of perceptually grounded categories adequate for communication. If the environment changes or imposes new challenges, genetic evolution could still help the population to adapt.

- **Approach 2. Empiricism.** All human beings share the same learning mechanisms, so given sufficiently similar environmental stimuli and a similar sensory-motor apparatus they will arrive at the same perceptually grounded categories reflecting the statistical structure of the real world. Hence the acquisition of language is again a matter of learning labels for already known shared categories and there is no influence of language on category formation. This view is common among 'empiricist psychologists' (Elman et al., 1996) and researchers in inductive symbolic machine learning (Quinlan, 1993) or connectionism (Rumelhart and McClelland, 1986). If we adopt this approach, the agents will need to have some inductive learning mechanism with which they can derive the perceptually grounded categories relevant in their environment, but it is not necessary to introduce a genetic basis for the categories and hence the genetic structure of the agents can be much simpler. Each agent now needs neural networks, or functionally equivalent clustering algorithms, to perform statistical learning, as well as networks that learn the association between names and categories. To guarantee continued adaptation to an open environment, agents would need to regularly update their repertoire by performing induction on new incoming stimuli.

- **Approach 3. Culturalism.** Although human sensory systems, learning mechanisms, and environments are shared, there might still be sufficiently important degrees of freedom left so that categories are not yet sufficiently shared within a population to support communication. Culturalism therefore argues that language communication (or other forms of social interaction where perceptual categories play a role) is required to further coordinate perceptual categorisation by providing feedback on how others conceptualise the world. So language now plays an additional causal role in conceptual development (e.g. Gumperz and Levinson, 1996; Bowerman and Levinson, 2001; Gentner and Goldin-Meadow, 2003).

  This cultural hypothesis is favoured by those advocating a 'cultural psychology' (Tomasello, 1999) and those viewing language and its underlying conceptual framework as a complex adaptive system that is constantly coordinated by its users (Steels, 1997). If this approach is adopted for the artificial agents, it requires that they are not only given mechanisms to invent or adopt categories and ways to create and adopt associations between names and categories, but also ways to align these choices with other agents based on feedback in communication. The colour categories are now influenced by multiple factors: embodiment constraints, the history of interactions and the adaptation after each interaction, and the collective consensus arrived at through negotiation.

There seems no clear consensus in the cognitive science literature on which approach is most appropriate. We find researchers strongly arguing on the basis of children's early word learning that language acquisition and concept acquisition go hand in hand (Bowerman and Levinson, 2001), take a long time (Bornstein, 1985; Teller, 1998) and require a strong form of cultural learning (Tomasello, 1999), whereas others have argued that perceptually grounded concepts are either innate (Shepard, 1994) or acquired prior and independently of language (Harnad, 1990) without direct linguistic or categorical feedback (Bloom, 2000). So, the engineer is not given a clear choice for what would be the best blueprint for implementing category formation and naming by embodied communicating agents.

1.1 A Case Study for Colour

Colour has become a prototypical case study to investigate issues of category sharing in humans because of the relative ease with which it is possible to gather data (compared to for instance olfactory or gustatory experience) and because colour is well understood as a physical phenomenon (Wyszecki and Stiles, 1982). Colour is of course also one of the primary modes, although surely not the only one, in which artificial robotic agents interact with the world, given the highly advanced state of digital camera technology.

Knowledge about the neurophysiology, the psychophysics, and the molecular genetics of colour vision has been increasing steadily (for an introduction see Gegenfurtner and Sharpe, 1999). In recent years is has become clear that colour perception is perhaps more variable within normal subjects than previously thought (e.g. Bimler et al., 2004). Results from molecular genetics show that there are several allelic variants of opsin genes, and that between 15 and 20% of Caucasian females has the genetic potential to be tetrachromatic instead of trichromatic (Winderickx et al., 1992; Neitz et al., 1993; Sharpe et al., 1999; Mollon et al., 2003).

The impact of the variation of the neural substrate on colour perception, colour categorisation and colour naming is still being investigated. But it is another reason from an engineering viewpoint why it is a good idea to take a closer look at how humans arrive at shared categories. Fabrication processes of complex artifacts like robots or cameras are such that there will always be individual differences, particularly if some form of calibration is involved. So if Nature has found a solution to enable shared categorisation in communication, even if the perceptual apparatus is not exactly the same, then that is very relevant for communicating robots as well.

Psychologists and neurobiologists have been collecting large amounts of data that could help understand how human beings arrive at shared perceptually grounded categories for communication. Data supporting a genetic coding of colour categories are sought by studying the colour categorisation behaviour of new-born children (Bornstein et al., 1976; Gerhardstein et al., 1999). Data supporting the presence of learning are sought in colour tests with pre-language children (Bornstein, 1975; Bornstein et al., 1976; Davies and Franklin, 2002) and in experiments where individuals from one culture learn the colour categories of another one (Rosch-Heider and Olivier, 1972; Roberson et al., 2000).

Anthropologists have also tried to collect empirical data for whether all human beings in the world, whatever their language or culture, use exactly the same colour categories (universalism) or whether there are significant differences (relativism). If colour categorisation is universal then this is of course a very strong indication that either it must be genetically determined due to constraints on physiology (just as each of us has five fingers) or innate categorisation, or that there is enough statistical structure in the real world so that neural systems performing clustering can easily pick it out, as empiricists have been suggesting. In that case, it should be straightforward to use these universal categories as the basis of robotic implementations as well.

The anthropological research has been conducted using colour naming tests and memory tests (Berlin and Kay, 1969; Rosch-Heider and Olivier, 1972; MacLaury, 1997; Davidoff et al., 1999; Kay et al., 2003), as first introduced by Lenneberg and Roberts (1956):

1. The naming experiments require informants to point to the best example for one of the 'basic' colour words in their language. It has consistently been found that subjects are not only capable of doing this, but that there is also a large consensus in a language community about what the focal point is for a particular word, even though there is less of a consensus about the boundaries of its colour region (Berlin and Kay, 1969).

2. The memory experiments require informants to pick out a colour sample seen earlier. It has been found that samples which are closer to focal points are

better remembered than those closer to the boundaries (Rosch-Heider, 1971, 1972).

Berlin and Kay (1969) based on data of naming experiments and memory experiments have argued strongly that the focal points of colour categories are shared by all languages and cultures of the world. Recent analysis by Kay and Regier (2003) of data gathered during the World Color Survey (Kay et al., 2003), confirm that there are cross-linguistic tendencies in colour naming in different languages. Named colour categories of languages across the world appear to cluster at points that tend to be described by English colour names. But researchers like (Davidoff et al., 1999; Roberson et al., 2000; Davidoff, 2001), have presented evidence through the same sort of memory and naming tests that the focal points of English and Berinmo (a Papua New Guinea tribe) are substantially different and that Rosch-Heider's data has been misinterpreted. So despite the abundance of data, no consensus has emerged in the universalism versus relativism debate, on the contrary, colour categorisation seems one of the most controversial areas of cognitive science (e.g. Saunders and van Brakel, 1997; Lucy, 1997; Sampson, 1997).

It is therefore not surprising that no consensus has been reached on how the perceptually grounded categories underlying language communication become shared. The nativist view on colour has been strongly defended, among others, by (Berlin and Kay, 1969; Kay et al., 1991; Shepard, 1992; Pinker, 1994; Kay and Maffi, 1999) based on the identification of universal trends in colour categorisation. Language plays no role in this. As Pinker puts it: "The way we see colors determines how we learn words for them, not vice versa." (Pinker, 1994, p. 63). Other researchers have strongly defended an empiricist position, by trying to find correlations between specific environments and the colour categories of certain communities (Van Wijk, 1959), or by investigating how clustering algorithms can pick out the statistical distributions in natural colour samples (Yendrikhovskij, 2001). The culturalist view on colour categorisation and colour naming has its own defendants, see e.g. (Lucy and Shweder, 1979; Gellatly, 1995; Davies and Corbett, 1997; Davies, 1998; Dedrick, 1998; Jameson and Alvarado, 2003), among others.

1.2 Objectives

The present paper does not take a stance on whether a nativist, empiricist or culturalist approach is the most appropriate one for interpreting the human data. It focuses on the pragmatic goal of finding the best way to design autonomous embodied agents and leaves it up to future debate what this implies for human categorisation and naming.

Our position is that multiple sources of constraints act on perceptually grounded colour categories, and (at least in the case of artificial agents) all of them play a role:

1. *Constraints from embodiment*: Although there are more variations in the human visual sensory apparatus than usually believed (see references given earlier), there are of course still a large number of similarities in terms of what part of the spectrum human retinal receptors are sensitive to, what perceptual colour appearance model is used, what low level signal processing takes place (for example to calibrate perception to context), etc. Moreover there are also constraints from the kinds of neural processes that are used for categorisation itself, and they show up in human categorisation behaviour, for example through the importance of focal points. Nobody doubts that these constraints help to shape the possible repertoire of perceptually grounded colour categories and it has recently become possible to incorporate many of these constraints in artificial vision systems. We will do so in all the experiments reported in this paper.

2. *Constraints coming from the world*: Although there is significant variation in the environments in which human beings find themselves (compare growing up on the North Pole or in the rainforest), there are obviously considerable similarities. Biological organisms must be adapted to the environment to reach viable performance and this is also true for categorisation. It implies that the statistical structure of the environment has to be a second force shaping the possible categorical repertoire. We can achieve this for artificial agents by giving them stimuli that are taken from real world scenes. Of course, if they have to be adapted to another environment (like Mars) they would have to

be given stimuli from that environment.

3. *Constraints coming from culture*: We want to examine the hypothesis that embodiment and statistical regularity of the environment is not enough to achieve sufficient sharing for communication and that cultural constraints also play a role. Cultural constraints are collective decisions made by a population. For example, one community may decide to drive on the left side of the road whereas another one may decide to drive on the right side. Speakers of English have agreed to call a particular hue "blue" whereas they could just as well have used a different word like "plor". Cultural choice is also available with respect to the perceptually grounded categories that are used in conventionalised communication. Instead of making a categorical distinction between blue and green, a population may decide to combine these into a single category, as indeed many cultures have done. This implies that cultural constraints should be a third force, shaping the perceptually grounded categorical repertoire used for communication.

The first source of constraints is preferred by nativists, and in some extreme versions of nativism, it is argued that these constraints are enough to explain the (universally) shared human colour categories underlying language. This can only be when not only physiological constraints (such as those due to the retinal receptors) are genetically determined but also the colour categories themselves, in other words that the neural microcircuits performing colour categorisation are directly laid down under genetic control. The second source of constraints is preferred by empiricists. They accept of course that there are constraints from embodiment, but these constraints still leave many degrees of freedom so that the categories still need to be shaped for the most part by the environment. Moreover, they do not believe that additional cultural constraints are necessary. The third source of constraints is considered to be crucial by culturalists, even though they do not deny that embodiment and structure in the environment may also play a role. Their position has been the most controversial, perhaps because it is less obvious by what kind of process cultural constraints could play a role. There is a chicken and egg problem: to name a colour category it seems that this category must already exist and be shared, so how can naming influence the shaping of the category?

In order to tease apart the contributions from each source of constraints we have constructed a series of theoretical models and compared their behaviour. Besides the utility for designing artificial autonomous agents, we believe that this effort is also valuable for those exploring human (colour) categorisation and naming. Theoretical models make a particular view explicit and this makes it easier to structure the debate for or against a certain position. Theoretical models bring out the hidden assumptions of an approach, particularly with respect to the cognitive mechanisms that are required and the information they need. Moreover they help to assess the plausibility of certain assumptions, for example with respect to the time that is required to acquire categories or propagate word-meaning pairs in a population. Finally, theoretical models may suggest new experiments for empirical data collection.

Theoretical investigations of the sort undertaken in this paper are very common in many sciences but still surprisingly controversial for psychologists. For example, there is now a large body of game theoretic models which have revolutionised economics. These models are theoretical in the sense that they examine the consequences of certain assumptions about the structure of interactions between agents or the strategies they follow, for example they may show the presence or absence of a Nash equilibrium (Gibbons, 1992). Usually it is not possible to collect the necessary empirical data to make the model predictions empirically grounded, but still a lot can be learned about the possibility of certain outcomes or their plausibility. For example, they might help to infer the effects of certain consumer behaviours on specific business models, without evidence whether consumers actually exhibit these behaviours. Similar theoretical approaches are now widespread in biology. For example, it has been shown that certain observed phenomena, like cycles in predator-prey populations, are due to the mathematical properties of the underlying dynamical system and not to the specific biological instantiation (May, 1986).

The approach in this paper is in the same line and uses the same methodological tools. The verbal interactions between the agents are modeled as multi-agent decision problems, called discrimination games (to categorise the world) and language games (to communicate with others using these categories), and our main goal is to understand what properties follow from the dynamical system implied by the structure of the interactions and the strategies of the agents.

1.3 Overview

Because nobody doubts that embodiment constrains perceptually grounded categories, we have first of all attempted to integrate as well as possible the constraints coming from the physics of light interacting with objects in the real world and the constraints coming from the perceptual apparatus itself, as captured in widely accepted colour appearance models, such as the CIE $L^*a^*b^*$ space. We will also use the same neural networks for categorisation (radial basis function networks) in each of our models. These networks capture the prototypical nature of colour categorisation, as demonstrated by the naming and memory experiments, and are widely believed to be realistic models of the behaviour of biological neural networks. All our models incorporate these same embodiment constraints.

1. To explore position 1 (nativism) we introduce a model of genetic evolution capable of evolving “genes” for focal colours and show how these genes can become shared in a population. Notice that this represents the extreme nativist position arguing that not only embodiment but also the perceptually grounded categories themselves are innate.

2. To explore position 2 (empiricism) we introduce agents using an inductive learning algorithm in the form of a neural network capable of acquiring colour categories, and examine whether colour categories become shared among individual learners when the physiological and environmental constraints are identical.

3. To explore position 3 (culturalism) we strongly couple category formation to the situated use of colour categories in verbal communication and investigate if this enables a population to reach a shared categorical repertoire.

We not only examine for each of these models whether a shared repertoire of categories emerges but also whether a lexicon expressing these categories can arise in the population, and whether categorical sharing is sufficient for successful communication. This allows us to confront the ‘chicken-and-egg’ problem alluded to earlier: How can a self-organising lexicon influence an emergent adaptive categorical repertoire and vice-versa?

The semiotic dynamics generated in the interaction between perception, categorisation, and naming is too complex (in a mathematical sense) to be solved analytically, so we examine its properties through computer simulations, starting from real world physical colour data captured by a multi-spectral camera. The use of computer simulations for examining the behaviour of complex systems is common in all the sciences of complexity, including non-linear physics (Nicolis and Prigogine, 1989) or artificial life (Langton, 1995). It is characteristic for the "methodology of the artificial" (Steels, 2001b) and has been pioneered for colour cognition research by Lammens (1994), who proposed the first concrete computational models exploring colour categorisation and naming. In order to make the simulations feasible, cultural constraints will be exercised exclusively through language, even though language is clearly not the only factor that embodies such constraints. Note that the use of computer simulations does not imply any stance on whether the brain is a computer (we believe it is not), just as the use of computer simulations to make predictive models of the weather does not imply that the weather is seen as a computer.

In the first batch of experiments (section 3 and 4), the presented colour stimuli have no realistic statistical distribution, precisely because we want to examine whether a population can coordinate its colour categorisation and colour naming *even if there is no chromatic distribution in the data.* This therefore forces the question whether coordination is possible, purely based on a structural coupling between categorisation and naming processes. The main conclusion is that this is indeed possible and hence that it is at least plausible that language plays a role to coordinate the coordination of perceptually grounded categories. Our main contribution here is to solve the chicken-and-egg problem by introducing a two-way causality between naming and category formation.

Next, in section 5, we consider what happens when there is a statistical distribution in the samples. This will help us examine whether colour stimuli taken from real world scenes are sufficiently constraining so that no coupling between categorisation and naming is required to explain how a population can coordinate its repertoire of perceptually grounded categories (either through genetic evolution or statistical learning). The main conclusion here is that even if the statistical structure of the world constrains the categories that arise in the agents, it is not so obvious that the statistical structure of the environment alone can explain the sharing of perceptually grounded categories. This confirms that three interacting forces are at work: embodiment, an environment with statistical structure, and cultural negotiation.

Some conclusions and suggestions for further research end the paper.

2 Components for categorisation and naming

This section introduces the basic components needed for making computational models of colour categorisation and colour naming: agents, environments, and tasks.

2.1 Agents

We define an abstract object called an agent. A set of agents is called a population. We use small populations in this paper (typically 10 agents) because we know from other work that the mechanisms being used in our models scale up to populations of thousands of agents (Steels et al., 2002). All agents have the same architecture for perception, categorisation, and naming but each has unique associated information structures, representing its repertoire of categories and its lexicon. The agent's architecture is intended to model what we know today about human colour perception, categorisation and naming. Agents cannot use information structures of other agents, so they have no telepathic access to the categories or lexicons used by other agents. Neither do agents have a global view of what words are used by others. They have only local information coming from the interactions in which they were involved themselves. There is no central authority specifying how the agents should conceptualise reality or speak. Agents only interact by exchanging words and by non-verbal gestural feedback (pointing). The agent population is an example of a distributed multi-agent system (Ferber, 1998), commonly used in artificial life simulations.

Next we define verbal interactions between agents. An interaction has a communicative goal, namely the speaker draws attention of the hearer to an object in the environment. After each interaction, agents adapt their internal states to become more successful in the future. So the framework of evolutionary game theory, which has been used to model genetic and cultural evolution in biology (Maynard Smith, 1982), applies and we therefore call the interactions *language games*. The notion of a language game resonates with the philosophical work of Wittgenstein (1953) who emphasised the situated contextual nature of word meaning. Indeed, the agents in our simulations are grounded, in the sense that their symbols are coupled to the environment through a sensory apparatus (Harnad, 1990), embodied, because the apparatus and subsequent processing reflects human physiology (Kaiser and Boynton, 1996), situated, because the games are embedded in the context of communicative acts in a shared real world setting (Suchman, 1987), and cultural, because the agents are part of a population with recurrent interactions between the members (Sperber, 1996).

Genetic evolution is modelled by introducing change in the population. At regular times, some of the agents are replaced by offspring, i.e. mutated versions of themselves, depending on their success in colour categorisation and colour naming. This is in the spirit of research in genetic algorithms and evolutionary computing (Holland, 1975; Goldberg, 1989; Koza, 1992; Fogel, 1999).

Individualistic learning is modelled by a process by which the categorical repertoires and lexicons of the agents change in interaction with the environment but without interactions among the agents. This is in the spirit of connectionist learning (Elman et al., 1996). Cultural learning is modelled by using a similar connectionist learning algorithm but now with cultural constraints, exercised through language, playing an additional role (Steels, 2001a).

2.2 The environment

The environment consists of 1269 matte finished Munsell colour chips (Munsell, 1976), familiar from anthropological experiments. We use the spectral energy distribution $E(\lambda)$ reflected by physical chips as measured by a spectrometer from 380 to 800 nm in 1 nm steps (Parkkinen et al., 1989). So the simulations do not use monochrome colour samples nor random values in RGB or another colour space, but start from realistic colour data. In each game, the agent(s) are presented with a number of samples randomly drawn from the total set. This set constitutes the context of the game. One of the samples is chosen as the topic. Choice of topic and context reflect the ecological conditions of the environment.

The environmental complexity is experimentally controlled by changing the total number of colour samples and the similarity between the samples. The ecological complexity is controlled by varying the properties of the context: the average number of samples in a context and the (shortest) distance from the topic to the other samples in the context. For example, fine shades of orange may constitute the difference between edible and non-edible mushrooms. Mushroom eaters will therefore need to acquire the ability to distinguish these fine shades of orange. If the distinction is much clearer (for example because all edible mushrooms are orange and all non-edible ones are white), the agents' colour distinctions can be less fine-grained, even though the same diversity of orange shades might still occur in the environment. In general, when there are more samples and they are closer together, finer categorical distinctions are needed and the lexicon can be expected to contain more colour words. This dependency between environmental and ecological complexity on the one hand and cognitive complexity on the other is a property of the proposed models but is not further discussed in this paper (see Belpaeme, 2001).

2.3 Agent architecture

2.3.1 Perception

All agents are assumed to have exactly the same perceptual process. Perception starts from a spectral energy distribution $S(\lambda)$ and is converted into tristimulus values in CIE $L^*a^*b^*$, which is considered to be a reasonable model of human lightness perception (L^*), and the opponent channels red-green (a^*) and yellow-blue (b^*). This colour coding handles certain aspects of the colour constancy problem as well (Fairchild, 1998, p. 219).

The spectral energy distributions are converted to XYZ coordinates using the following equations.

$$\begin{aligned} X &= k \int S(\lambda)\, \bar{x}(\lambda)\; d\lambda \\ Y &= k \int S(\lambda)\, \bar{y}(\lambda)\; d\lambda \\ Z &= k \int S(\lambda)\, \bar{z}(\lambda)\; d\lambda \end{aligned} \qquad (1)$$

$\bar{x}(\lambda)$, $\bar{y}(\lambda)$ and $\bar{z}(\lambda)$ are the 1931 2° CIE colour matching functions, describing how an average observer reacts to chromatic stimuli[1]. The CIE $L^*a^*b^*$ colour coding is computed directly from these CIE XYZ values using standard formulae (Wyszecki and Stiles, 1982, p. 166).

Obviously the realism of this model can be improved. For example, Lammens (1994) has started from the neural response functions proposed by (De Valois and De Valois, 1975; De Valois et al., 1966) and showed how tristimulus values in another colour space can be derived. This space, though carefully constructed and founded on neurophysiological data, is not as suited for colour categorisation as is the CIE $L^*a^*b^*$ space (Lammens, 1994, p. 142). So for categorising colour perception, CIE $L^*a^*b^*$ remains a good choice[2].

2.3.2 Categorisation

Categorisation is based on the generally accepted notion that colours have prototypes and a region surrounding each prototype (Rosch, 1978) with fuzzy boundaries (Kay and McDaniel, 1978). Categorisation can therefore be modelled with adaptive networks, a modification of radial basis function networks (Medgassy, 1961), which are widely assumed to have a high biological plausibility (Hassoun, 1995). Input to the network is a tristimulus $\mathbf{x}$ in CIE $L^*a^*b^*$ space.

An adaptive network consists of locally reactive units. These units have a peak response at a central value $\mathbf{m}$ and an exponential decay around this central value. The regional extent around $\mathbf{m}$ is determined by a normalised Gaussian function, of which the width[3] is defined by parameter σ, thus giving rise to the magnet effect typically found in categorical perception (Harnad, 1990). The behaviour of each unit j is defined as follows:

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

Rather than using a single decision unit, as in the work of Lammens (1994), an adaptive network is used for each colour category[4]. Each network contains weighted locally reactive units, so that colour regions do not have to be symmetrical - as is the case with only a single decision unit. Each unit in the network reacts to an incoming stimulus $\mathbf{x}$, as in eq. (2). The reaction of an adaptive network for category k with J locally reactive units has the following form, familiar from perceptron-like feed forward networks (Minsky and Papert, 1969), where w_j is a weight factor with a range between 0 and 1.

$$y_k\left(\mathbf{x}\right) = \sum_{j=1}^{J} w_j z_j\left(\mathbf{x}\right) \tag{3}$$

Each colour category has its own adaptive network and all networks consider the input in parallel. The 'best matching' colour category b for a given tristimulus value $\mathbf{x}$ is determined by a winner-take-all process based on the output of each categorical network.

$$\forall c \in C : y_b(\mathbf{x}) \leq y_c(\mathbf{x}) \tag{4}$$

The various components of the adaptive networks are summarised in figure 1. Physiological evidence for locally reactive units in the domain of vision have been found in the macaque monkey visual cortex (Komatsu et al., 1992) and these neurons have been modeled by Lehky and Sejnowksi (1999).

2.3.3 Naming

Naming is modelled with an associative memory network $\mathcal{L}$. One word form can be associated with several categories (because the agent must be able to maintain multiple hypotheses about what the meaning is of a word) and one category with several word forms (because the agent must be able to maintain multiple hypotheses about which word to use for a specific meaning). Given a set $\mathcal{C}$ of n categories and a set $\mathcal{F}$ of m word forms, this network consists of $n \times m$ relations, each having

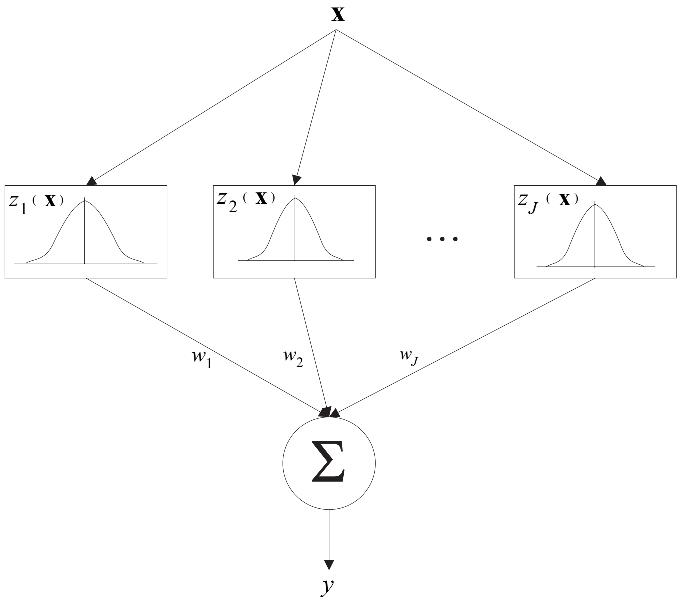

Figure 1: Categorisation is performed by an adaptive network consisting of locally reactive units fully connected to a summing output unit. Each such network corresponds to one colour category.

a strength $s \in [0.0, 1.0]$, so that $\mathcal{L} = \mathcal{C} \times \mathcal{F} \times [0.0, 1.0]$. Words are randomly selected from a finite alphabet of syllables. The strength of the association between a category and a word can be varied, as explained later. When a word form f is needed for a category c, there is a winner-take-all competition and the word form with the highest strength wins. Conversely, to find the category given a word form f, there is again a competition. The category c with the highest strength is taken as the winner.

2.4 Tasks

We will explore two types of interactions: The first one requires an individual agent to discriminate a sample (called the topic) from a set of other samples. This means that the agent must not only categorise all the samples in the context but must also find a categorisation of the topic which is unique for the topic and does not apply to any other sample in the context. We call this a *discrimination game* (Steels, 1996a). The second interaction is between two agents in a shared context playing the role of speaker and hearer. The speaker chooses the topic, categorises it, using a discrimination game, and names the categorisation. The hearer must identify the topic based on the category name. We call this a *guessing game* because the hearer has to guess the object intended by the speaker through verbal means. We have been using the guessing game in a wide variety of experiments investigating the origins of language (Steels and Kaplan, 1999; Steels et al., 2002), including experiments on autonomous mobile robots (Steels, 2001a; Steels and Kaplan, 2002; Vogt, 2003).

2.4.1 The Discrimination Game

The discrimination game has been chosen so as to introduce the ecological dimension in the models. As already mentioned, suppose that there are various types of mushroom which all have similar form and shape but are only distinguishable based on their colour, and only one type of mushroom is edible. Given a specific situation with a number of mushrooms on the table, the agent must play a discrimination game where the topic is the edible mushroom and the other objects in the context are the non-edible ones. So ecology is concretised through what objects form a context, and which ones are topics that need to be distinguished. Similar examples could be given for distinguishing predators or prey based on colour marks, distinguishing members of the group from outsiders using the colour of clothes, etc. In later simulations, contexts are chosen randomly and any sample in the context can be the topic, so there is no strong distinction between environmental constraints (what stimuli are present in the environment) and ecology (which stimuli are functionally significant to the agent).

The discrimination game is defined more precisely as follows. An agent has a, possibly empty, set of categories C. A random context $O = \{o_1, ..., o_N\}$ is created and presented to the agent. It contains N colour stimuli o_i of which one is the topic o_t. These colour stimuli take the form of spectral distributions of energy against wavelength. The topic has to be discriminated from the rest of the context. The game proceeds as follows.

1. Context $O = \{o_1, ..., o_N\}$ and the topic $o_t \in O$ are presented to the agent.

2. The agent perceives each object o_i and produces a sensory representation for each object: $S_{o_i} = \{s_1^{o_i}, ..., s_N^{o_i}\}$. The sensory representation is the CIE $L^*a^*b^*$

value computed from the spectral distribution as discussed earlier.

3. For all N sensory representations, the 'best' category $c_{S_o} \in C$ is found, according to

$$c_{S_o} = \arg\max_{C} \left(y_c \left(S_o \right) \right) \tag{5}$$

   y_c is the output of the adaptive network belonging to category c, and S_o is the sensory input for an object o.

4. The topic o_t can be discriminated from the context when there exists a category whose network has the highest output for the topic but not for any other sample in the context.

$$\text{count} \left(\left\{ \mathrm{c}_{\mathrm{S}_{\mathrm{o}_1}}, ..., \mathrm{c}_{\mathrm{S}_{\mathrm{o}_\mathrm{N}}} \right\}, \mathrm{c}_{\mathrm{S}_{\mathrm{o}_\mathrm{t}}} \right) = 1 \tag{6}$$

2.4.2 The Guessing Game

The guessing game has been chosen because it is the most basic language game one can imagine. It is a game of reference where the speaker wants to get something from the listener and identifies it through language, as opposed to gestures. Language presupposes a categorisation of reality because words name categories and not individual objects. The ecological relevance of guessing games is obvious. For example, two people sit around the table on which there are various fruits of the same form and shape but with different colours. The speaker wants a particular type of fruit (the topic). She says for example "could you give me the red one", whereby the hearer has to apply the category which is the meaning of "red" to the objects in the context and identify the desired fruit. The meaning of "red" is the category which discriminates the topic from the other objects in this context. So the guessing game implies a discrimination game.

The guessing game is more precisely defined as an interaction between two agents, one acting as the *speaker* and the other as the *hearer*. The agents have an associative memory relating colour categories with colour names. Each association has an associated strength. The game consists of the following steps.

1. A context $O = \{o_1, \ldots, o_N\}$ is presented to both the speaker and the hearer. Only the speaker is aware of the topic $o_t \in O$.

2. The speaker tries to discriminate the topic from the context by playing a discrimination game. If a discriminating category c^s is found the game continues, otherwise the game fails.

3. The speaker looks up the word forms associated with c^s. If no word forms are found, the speaker creates a new random word form f by combining syllables from a pre-given repertoire and stores an association between f and c^s. On the other hand, if there are word forms associated with c^s, the one with the highest strength s is selected. The speaker conveys word form f to the hearer.

4. The hearer looks up f in its lexicon. If f is unknown to the hearer, the game fails and the speaker reveals the topic o_t to the hearer by pointing to it. The hearer then tries discriminating the topic o_t from the context. If a discriminating category is found, the word form f is associated with it; if no discriminating category is found, a new category is created to represent the topic and f is associated with it.

5. If the hearer does have the word form f in its lexicon, it looks up the associated category c^h and identifies the topic by selecting the stimulus in the context with the highest activation for this category c^h. The hearer then points to this sample.

6. The speaker observes to which sample the hearer is pointing and if this is the one that it choose as topic, the game is successful. If not, the speaker identifies the topic and the hearer adapts its categorical network and its lexicon as in (4) to become better in future games.

When agents only engage in discrimination games, the formation of colour categories is influenced by physiological, environmental and ecological constraints only. When agents perform a discrimination game *and* a guessing game a cultural dimension is brought in (through language). Guessing games are therefore an effective way to study the potential causal relation between language and category acquisition. Another reason for using the guessing game is that the colour chip naming experiments widely utilised in anthropological research (Lenneberg and Roberts, 1956;

Lantz and Stefflre, 1964; Berlin and Kay, 1969; Rosch-Heider, 1972; Kay et al., 1991, 1997; MacLaury, 1997) are equivalent to guessing games. So, if needed, the results of our simulations can be compared with anthropological data obtained with human subjects. One difference is that the context in most anthropological studies usually consists of all the Munsell chips and the topic is the best representative or prototype of a colour name. We believe that it would be desirable that anthropological experiments are made more realistic by asking subjects to name topics within ecologically valid contexts (see also Jameson and Alvarado, 2003). Presenting all the Munsell chips at once is obviously an unusual problem setting for human subjects, no wonder that some report difficulties doing it.

We now discuss a series of computer simulations exploring different ways in which colour categories and colour names can be acquired. The first series (section 3) assumes that there is no causal role of language in concept formation, so agents only play discrimination games. The next section (section 4) uses guessing games to explore the interaction between conceptualisation and language. As mentioned earlier, no statistical structure is present in the data, in order to find out whether coordination of categories takes place even in the absence of such a structure. In section 5, we then examine colour samples drawn from real world data where a clear chromatic structure is present.

3 Learning without Language

We have seen earlier that there could be two approaches for the problem how concepts are acquired: either they are learned or they are innately present, the latter implying that they have evolved through genetic evolution. Both possibilities are now explored in sections 3.2. and 3.3. respectively. The discrimination context is the same for both experiments and consists of 4 stimuli chosen from a total of 1269 Munsell chips. In the learning case, the agents adapt their categorical networks during their lifetime in the spirit of connectionist learning systems (Churchland and Sejnowski, 1992). In the genetic evolution case, the agents have a fixed network and change only takes place when there is a new generation whose “colour genes” have undergone some mutation, in the spirit of genetic algorithms (Holland, 1975). But first we need some measures to follow the progress and adequacy of concept formation (for more details on these measures see (Belpaeme, 2002)).

3.1 Measures

To play a discrimination game i the agent A is given a context that consists of a set of (randomly chosen) colour samples. One sample from this context (also randomly chosen) is the topic. The agent then exercises its categorisation network. There are two possible outcomes:

1. The colour sample is uniquely categorised. Agent A is therefore capable of discriminating the topic from the other colour samples. The discriminative success for game i is $ds_i^A = 1$.
2. No unique category was found for the topic. The discrimination game has failed. $ds_i^A = 0$.

The discriminative success of the agent for a specific environment ideally reaches 100 percent. In this case we say that the agent has acquired an adequate repertoire of colour categories for that environment. The cumulative discriminative success at game j for a series of n games is defined as:

$$DS_j^A = \frac{\sum ds_i^A}{n} \tag{7}$$

The average success of a population of m agents at game j is defined as

$$\mathcal{DS}_j = \frac{\sum DS_j^A}{m} \tag{8}$$

The category variance cv between the categorical repertoires of the different agents is measured by computing the cumulated distance between the categories of the agents of a population $\mathcal{A} = \{A_1, \ldots, A_n\}$, as in

$$cv(\mathcal{A}) = \frac{1}{\frac{1}{2}n(n-1)} \sum_{i=2}^{n} \sum_{j=1}^{i-1} D\left(A_i, A_j\right) \tag{9}$$

$D(A_i, A_j)$ is a distance measure between the category sets of two agents[5].

3.2 Individualistic Learning

We now present a model of individualistic learning. The update rule used by an agent after playing a game is as follows:

When successful The weights w_i of each locally reactive unit i of the discriminating category network are increased according to the following rule

$$w_i = w_i + \beta\, z_i\left(S_{o_t}\right) \tag{10}$$

Where $z_i\left(S_{o_t}\right)$ is the output of unit i for the topic S_{o_t}, and β is the learning rate[6].

When not successful The discrimination game scenario can fail in two ways. First, the agent has no categories yet ($C = \emptyset$); in this case the agent creates a new category centred on the topic. Second, no discriminating category can be found because the category found for the topic is also applicable to the other objects. When the discriminative success of the agent is lower than a predefined threshold (set at 95%), a new category is created. Otherwise, the best matching category network is adapted by adding a new locally reactive unit to its network.

Adding a new category is done by creating a category with only one locally reactive unit centred on the sensory representation of the topic ($\mathbf{m} = S_{o_t}$). Adapting a category is similarly done by just adding a new locally reactive unit sensitive to the topic.

After playing a discrimination game, the weights of all the locally reactive units of all categories of the agent are decreased with a small factor. The weight decay, a learning rule standard in the literature (Rumelhart and McClelland, 1986; Krogh and Hertz, 1995), is defined as

$$w_j = \alpha\, w_j \tag{11}$$

where $\alpha \leq 1$ is a non-negative value. This takes care of a slow "forgetting" of unused categories and thus of the reshaping of categories to remain adapted to changes to the environment or the ecology.

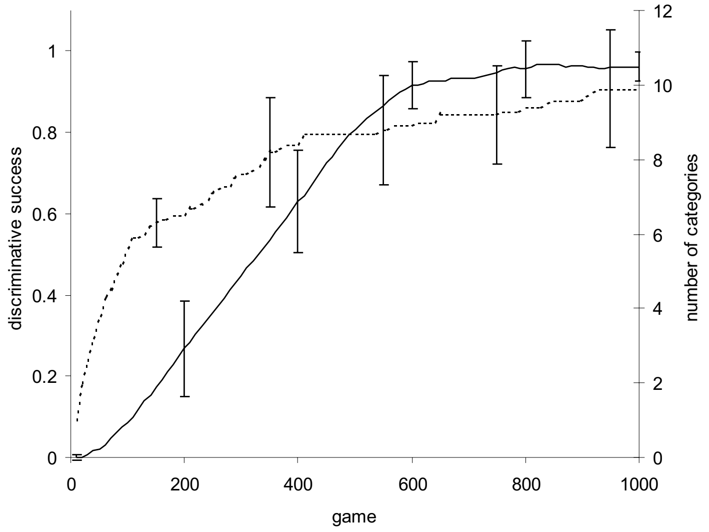

Figure 2: Average discriminative success $\mathcal{DS}$ and average number of categories (dotted line) of 10 agents playing discrimination games.

The following graphs show the outcome of simulations exploring this model. Agents play successive discrimination games with random sets of samples from the environment and randomly chosen topics within each set. In a first illustrative experiment (figure 2) a population of 10 agents plays a series op 1000 discrimination games. The context of a game contains four colour stimuli chosen randomly from the complete set of over 1269 Munsell chips, of which one has to be discriminated from the other three. The chips are at a minimum Euclidean distance of 50 from each other in $L^*a^*b^*$-space. Agents take random turns playing a game. Two agents are randomly selected from the population to play one discrimination game. The x-axis maps to consecutive games. The left y-axis of figure 2 shows the average success rate in the discrimination game with the learning rules used here. We see clearly that discriminative success increases to almost 100 %, proving that the agents are capable of developing a repertoire of colour categories adequate for the given environment[7]. The right y-axis plots the size of the categorical repertoire. It stabilises when the agents have become successful in discrimination. It is undeniable that a repertoire forms which is adequate for the given environment and ecology. When the environment or the ecology is more complex, agents take longer and the number of categories increases, but the same trend is seen.

A mapping of the extent and focal points of the different colour categories for two agents onto the Munsell array is shown in figure 3.

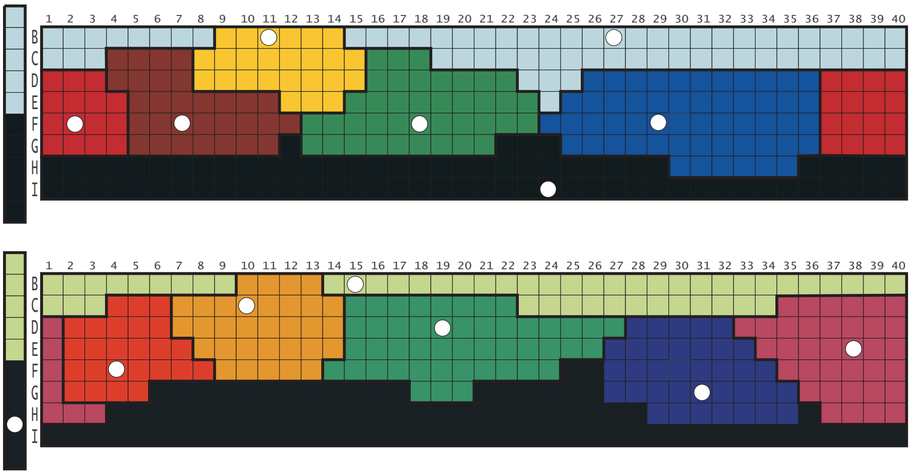

Figure 3: The maximum (white circle) and the extent (colour coding) of the categories of two agents after playing 1000 discrimination games. The chart consists of saturated Munsell chips, following (Berlin and Kay, 1969). Observe how categories are distributed across the Munsell chart, and how both agents end up with different categories.

Figure 4 shows that agents endowed with adaptive networks are capable of coping with changes to the environment. The agents start now with a context of four stimuli randomly chosen from a total of seven stimuli. The stimuli are equal to the Munsell chips[8] corresponding to red, yellow, green, blue, purple, black and white. The categorical repertoires stabilise and after 50 games four more stimuli are added as potential choices[9]. We see at first a dip in discrimination success. Then the agents quickly adapt to the more complex situation by expanding their colour repertoires. Note that the population does not change during the course of the simulation and agents do not interact with each other. The observed behaviour is entirely based on individualistic learning.

Clearly the proposed mechanisms solve the acquisition problem, but what about the sharing problem? Figure 5 compares the repertoire of the different agents for the same run as in figure 2, using the category variance metric *cv* defined earlier. Although the agents are all capable of discrimination, they use different repertoires. And although the repertoires tend to become more similar as the simulation pro-

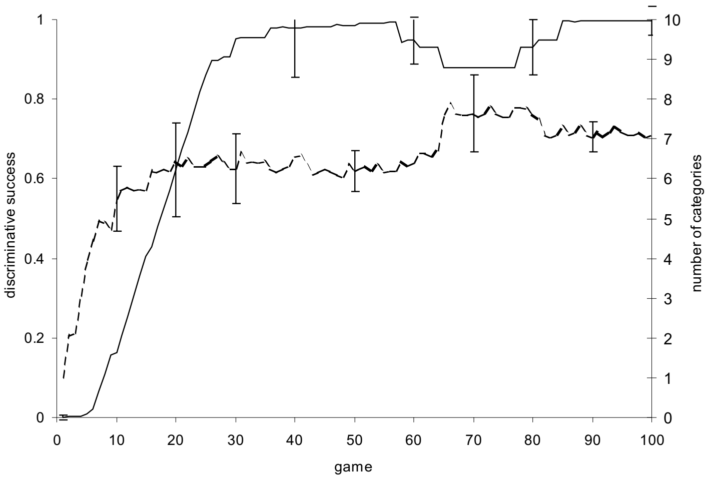

Figure 4: Average discriminative success $\mathcal{DS}$ and average number of categories (dotted line) for a population of 10 agents which *learn* colour categories. In the first 50 games the context is chosen from a simple stimuli set, after 50 games the set of stimuli is extended to increase complexity. The graph shows how the agents cope to reach again a discriminative success of 100 %.

gresses, the similarity is not absolute (if all categories would be similar, the category variance would be zero). This demonstrates that the constraints which are at work, namely the physiological constraints (perception and cognitive architecture) and the environmental and ecological constraints, are not enough to drive the agents to the same solution space. Different solutions are possible for the same task in the same environment. More sophisticated physiological models will probably not alter that fact. Indeed, it confirms why it has not been proven possible to explain basic colour categories based on physiological constraints alone (e.g. Gellatly, 1995; Jameson and D'Andrade, 1997; Saunders and van Brakel, 1997). If different populations exposed to different environmental stimuli and ecological challenges are compared, the repertoires of the agents in the population would be even more different.

Table 1 shows the inter-population category variance cv', a metric used to show how well categories compare *across* populations. It is the average of the category variance computed between all agents of two different populations P and P'. n and m are the number of agents in the respectively population P and P', assumed to be equal for all populations being compared.

$$cv'(P, P') = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} D\left(A_i, A'_j\right) \tag{12}$$

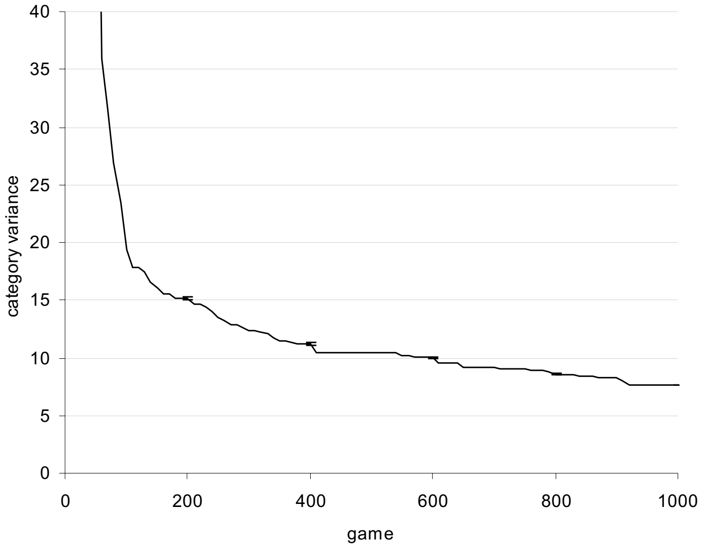

Figure 5: The category variance of a population of 10 agents playing discrimination games (for the same simulation as reported in figure 2). The graph shows how the categories of all agents start to resemble each other due to ecological pressure, but do not become equal.

Table 1 shows that the category sets of agents *within* and *across* populations are quite dissimilar (an intuitive grasp can be obtained by comparing the values in this table with other category variance tables in the following sections). If the categories of agents are similar between two populations, cv' would decrease. Populations where all individuals have identical categories have $cv' = 0$.

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 1: Inter-population category variance cv' of 5 populations of which the categories have been learned under identical experimental settings, except for the initial random seed.

We conclude that

1. Individualistic learning leads to the development of an adequate repertoire of colour categories.

2. There is a certain percentage of sharing of colour categories within a population, which can be attributed to shared physiological, environmental and ecological constraints, but there is no 100 % coherence.

3. The colour categories are not shared across populations.

3.3 Genetic evolution

This section turns to the properties of genetic evolution. We examine a variation of the previous model which includes different generations of agents. Each agent has a set of "colour genes" which directly encode its categorical networks, so we shortcut the problem of modelling gene expression. The networks do not change during the lifetime of the agent. Agents play exactly the same discrimination game as before. They have a cumulative score, reflecting their success in the game, as defined earlier. This score will be used as the fitness of the agent. The m fittest agents (where m is equal to 50% in the present simulation) are retained in the next generation and the others are discarded. A single mutated copy is made of each remaining agent so that the size of the population always remains constant. Mutations, which happen with a probability inversely proportional to discriminatory success, can take four forms with equal probability:

1. A new category network is added with a single locally reactive unit whose centre is at a random point in the $L^*a^*b^*$ space.

2. A randomly chosen category network is expanded by adding a new locally reactive unit whose centre $\mathbf{m}$ is at a random deviation from the centroid $\mathbf{c}$ of the category. The centroid $\mathbf{c}$ of the category is computed as in (eq. 13). The centre of the added locally reactive unit is randomly chosen from a normal distribution with mean $\mathbf{c}$ and standard deviation σ.

$$\mathbf{c} = \frac{\sum w_i \mathbf{m}_{c,i}}{\sum w_i} \tag{13}$$

3. A randomly chosen existing category network is restricted by removing one, randomly chosen locally reactive unit. If no unit is left the category network itself is removed.

4. An existing, randomly chosen category network is removed.

Only one mutation is allowed for each copy. Note that the mutation operator does not use any intelligence about what might be good changes to the categorical repertoire, as indeed it should be.

Figure 6 shows the behaviour of this model using the same environmental stimuli as in the learning case discussed earlier (a context consists of 4 stimuli chosen from a total of 1269 Munsell chips.) The x-axis plots the different generations of agents. The y-axis displays the success rate after n generations. This success rate is based on the outcome of 50 discrimination games. We see that after several generations a population of agents is reached which have adequate categorical repertoires for the given environment. When this environment is made more complex (in a similar way as in figure 4), genetic evolution generates more colour categories and after a number of generations there is again an adequate repertoire (figure 7). Figure 8 shows the focus and extent of the categories of two agents plotted on the two dimensional Munsell colour chart.

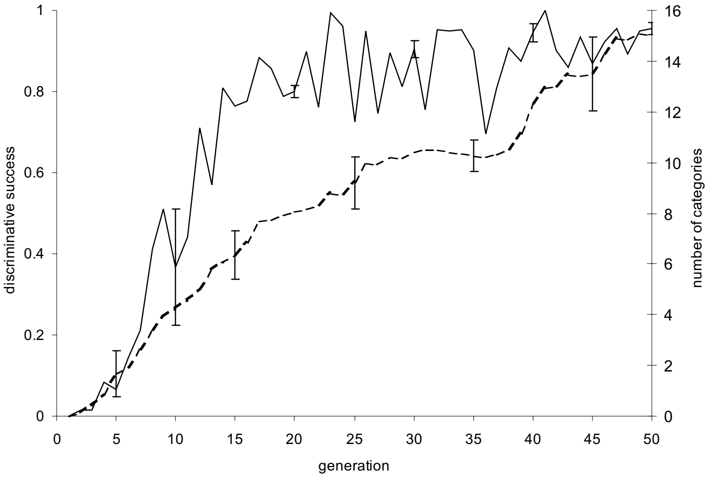

Figure 6: Average discriminative success $\mathcal{DS}$ and average number of categories (dotted line) for a population of 10 agents of which the colour categories are *evolving* in a genetic fashion.

These results show that our model of genetic evolution is also capable of evolving agents that have adequate repertoires of colour categories. There is of course a profound difference between the learning and genetic scenarios. In the learning scenario, agents start their life with no colour categories, develop an adequate reper-

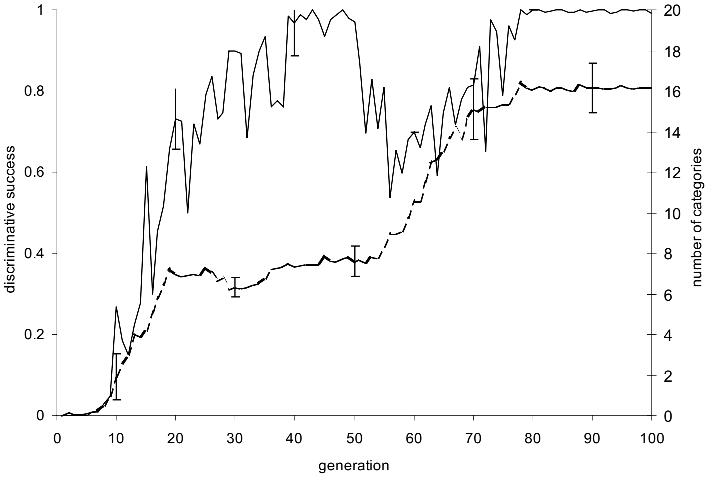

Figure 7: Average discriminative success $\mathcal{DS}$ and average number of genetically evolved categories (dotted line) for a population of 10 agents. In the first 50 games the context is chosen from a simple stimuli set, after 50 games the set of stimuli is extended to increase complexity.

toire within their lifetime, and adapt to environmental changes (for example caused by the availability of new dyes) also within their lifetime. In the genetic scenario, successive generations of agents are needed before a generation arises that has an adequate repertoire. So genetic evolution is much slower than learning, which is of course a well known fact. This is borne out by the simulation results shown in figure 7 which uses the same data as the learning case in figure 4. Rather than adapting after two dozen more games, the agents need about 20 generations (which would amount to at least 400 years of evolution if such a mechanism was to be applied to an equally small population of ten humans, counting a modest 20 years per generation). On the other hand, once genetic evolution has established a repertoire, agents do not have to learn anything but get born with a ready-to-use categorical repertoire.

Figure 9 displays the category variance between the categorical repertoires of the agents in the case of genetic evolution. We see clearly that genetic evolution not only solves the acquisition problem but also the sharing problem. The population evolves towards the same categorical repertoire for all the agents. This is in strong contrast with the learning scenario where the final repertoires were never identical. The cause of this sharing lies in the nature of genetic evolution. The

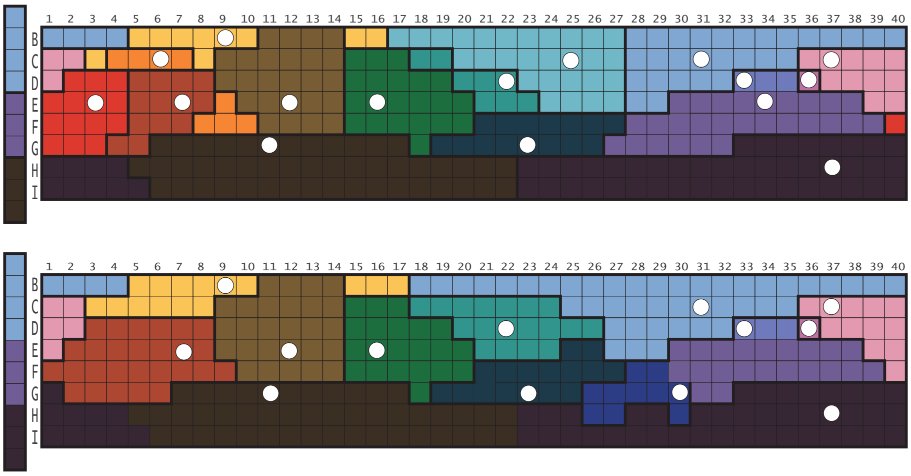

Figure 8: The maximum (white circle) and the extent (colour coding) of the categories of two agents with genetically evolved categories. Because of the dynamics of the evolutionary process, most categories of both agents are identical.

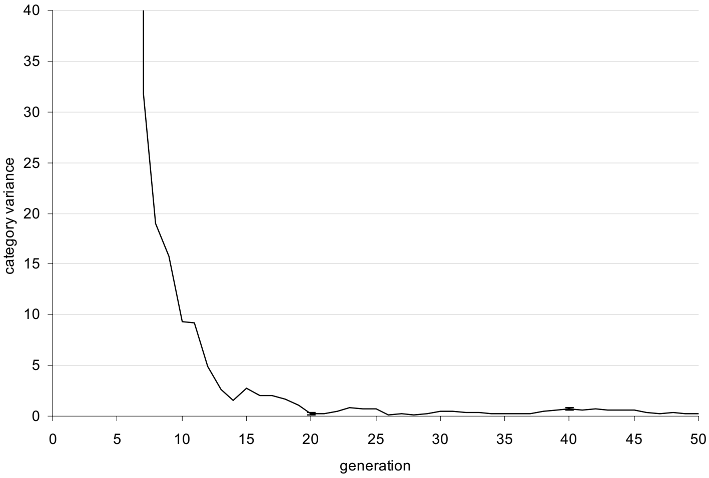

Figure 9: The category variance of a population of 20 agents after evolving for over 50 generations (under the same conditions are figure 6). It can be seen that there is hardly any variation between the categories of the agents.

colour genes coding the categorical networks of more successful agents propagate in the population and so after some time these ’genes’ completely dominate. Which colour categories come out depends on environmental, ecological and physiological constraints, but there are multiple solutions. Genetic evolution randomly selects one solution which then spreads to the rest of the population. This is clearly seen by doing another simulation with exactly the same parameters (for the environment, genetic mutation rates, etc.) but starting from another random seed. Due to the randomness inherent in the genetic search process, both repertoires are very different. This is shown in table 2: The variation within a population is almost nonexistent (≤ 0.40) but across populations the variation is considerable. With different ecological and environmental constraints the variation would be even more dramatic.

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 2: Inter-population category variance of 5 populations of which the categories have been evolved using the discriminative success as fitness measure.

We conclude that

1. Genetic evolution leads to the development of an adequate repertoire of colour categories.
2. The colour categories are completely shared among the individuals within a population.
3. The colour categories are not shared across populations.

4 Learning with language

The previous section compared individualistic learning with genetic evolution. Both were capable to explain how categories may be acquired by individuals, but only genetic evolution could also explain how colour concepts could become shared. In the next series of experiments, we study the impact of language (and thus of culture) on the formation of colour categories, by letting the agents play guessing games, and discrimination games as part of a guessing game. Again we are interested to model both cases: learning (section 4.2) and genetic evolution (section 4.3). First we need some additional measures to follow the progress in the experiment.

4.1 Measures

There are three possible outcomes of a guessing game:

1. The topic pointed at by the hearer is equal to the topic chosen by the speaker. The game i is a success for both agents A: communicative success $cs_i^A = 1$.
2. The topic pointed at by the hearer is not equal to the topic chosen by the speaker. The game i is a failure for both agents A: $cs_i^A = 0$.
3. The game got stuck somewhere halfway, either because the speaker or the hearer did not have a discriminating category, or the speaker did not have a word for the category or the hearer did not know the word. In this case the game is also a failure for both agents A: $cs_i^A = 0$.

The cumulative communicative success CS_j^A of an agent A at game j for the last series of n games is defined as

$$CS_j^A = \frac{\sum cs_i^A}{n} \tag{14}$$

.

The cumulative success $\mathcal{CS}_j$ for a population of m agents $\mathcal{A}$ for the last series of n games at game j is defined as

$$\mathcal{CS}_j = \frac{\sum CS_j^{\mathcal{A}}}{m} \tag{15}$$

.

4.2 Lexicon acquisition

No one has ever proposed that humans acquire the vocabularies of their language by genetic evolution, simply because lexical evolution is too rapid, most humans are bilingual, and children clearly go through a long phase in which they acquire new words (de Boysson-Bardies, 1999; Bloom, 2000). Nevertheless a number of mathematical and computational models have appeared that show that genetic evolution can in principle do the job (Nowak and Krakauer, 1999; Cangelosi, 2001). These models code the lexicon as part of an agent's genome, use communicative accuracy as selection pressure, and propose gene spreading as the mechanism by which the group reaches coherence. Here we stick to the more realistic view that lexicons are learned and that coherence arises through self-organisation in the population. Two kinds of computational models have been proposed in such a case: observational learning models that do not use negative evidence (Hurford, 1989; Oliphant, 1996) and active learning models that use both positive and negative evidence (Steels, 1996b). It is the latter approach that is used further in this paper.

The word learning algorithm for the hearer and the speaker works as follows:

1. Assume that a *speaker* has associated the word forms $\{f_1, ..., f_m\}$ with the discriminating category c_k and assume that f_j is the word form with the highest strength s_{kj} between f_j and c_k.

- If the communication was successful, the speaker increases the strength s_{kj} by $\delta_{inc} = 0.1$ and decreases the strength of connections with other categories by δ_{inh} (this mechanism is called *lateral inhibition*).
- If the communication was unsuccessful, the speaker decreases the strength s_{kj} by δ_{dec}.

2. Assume that the *hearer* has associated categories $\{c_1, ..., c_m\}$ with the word f_k and assume that c_j is the category that had the highest strength for f_k.

- If the communication was successful, the hearer increases the strength s_{jk} by δ_{inc} and decreases the strength of competing words associated with the same category by δ_{inh}.
- If the communication was unsuccessful, the hearer decreases the strength s_{jk} by δ_{dec}.

The algorithm has therefore three parameters. In later simulations we use $\delta_{inc} = \delta_{inh} = \delta_{dec} = 0.1$. Lateral inhibition is based on positive evidence (a successful game) and is necessary to damp synonyms. When $\delta_{dec} > 0$ negative evidence plays a role, and this has been found to be necessary to damp homonymy.

When a speaker does not have a word yet for a category that needs to be expressed, it creates a new word form (by generating a random combination of syllables from a pre-specified repertoire) and adds an association between this word and the category in its associative memory with initial strength $s = 0.5$. This ensures that new words enter into the population and explains how a group of agents may develop a grounded lexicon from scratch. When a hearer does not have the word used by the speaker in its associative memory, it stores the new word with a category that is capable of discriminating the topic pointed at by the speaker with initial strength $s = 0.5$.

The positive feedback loop between use and success causes self-organisation, in the sense of non-linear dynamical systems theory (Nicolis and Prigogine, 1989), (Stengers and Prigogine, 1986). An example of self-organisation is path formation in an ant society. Ants deposit pheromone when returning to the nest with food. This attracts other ants which also deposit pheromone, and so there is a a positive feedback loop which causes all ants to assemble on the same path (Camazine et al., 2001). In a similar way, the more speakers adopt a word and the meaning underlying it, the more successful communication with that word will be and hence the more speakers will adopt it. The positive feedback loop between the use of a word and its success in shared communication, causes words to spread in the population like viruses and eventually dominate. This is illustrated in figure 10, taken from a large-scale experiment in lexicon formation discussed in (Steels and Kaplan, 1998). 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 due to the non-linear nature of the positive feedback loop.

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 bright and yellow if in most situations the brightest object is also the one that is uniquely yellow). This relates to Quine's well known puzzle (1960). A linguist observing a native can never be sure if *gavagai* means rabbit, or hopping, or a temporal slice of a four dimensional space-time rabbit. Incoherence will be disentangled when situations arise where the two meanings are incompatible, for example a bright object which is blue. This type of disentanglement is also observed with the mechanisms described here, see figure 11 taken from (Steels and Kaplan, 1999), which discusses this "semiotic dynamics" in more detail.

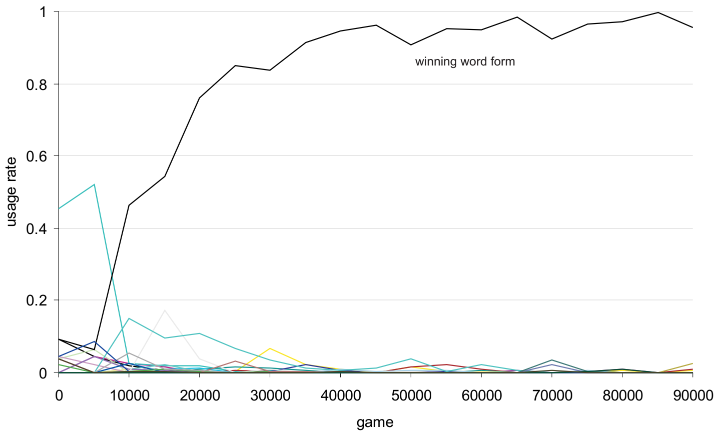

Figure 10: This graph plots the usage rate of all possible words for the same meaning in a consecutive series of language games. Initially many words are competing until one dominates due to a winner-take-all effect.

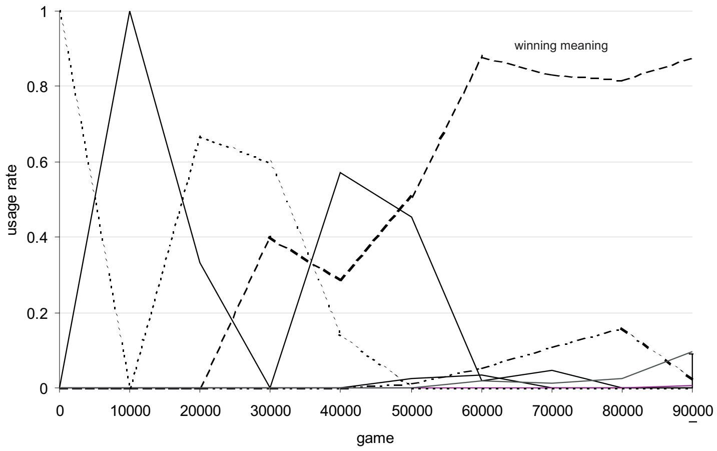

Figure 11: This graph plots (on the y-axis) the usage percentage of different meanings associated with the same word. Different meanings may co-exist until a situation arises that disentangles them.

4.3 Cultural Learning

Given these processes we can now begin to study the interaction of word learning and category acquisition. The first experiment uses learning both for categories and for words. When a category has been successful in the language game, i.e. it led to a successful communication, it is re-enforced by increasing the weights of its network according to eq. 10. This increases the probability that the category stays in the repertoire of the agent and that it is the category of choice when a similar situation arises in the future. So there is a two-way structural coupling (Maturana and Varela, 1998) between category formation and language: Language communication stimulates the formation of categories because it calls for a discrimination game that might lead to the learning of new categories. Category formation in turn stimulates language because if the discrimination game generates a new category, this leads to the creation of a new word. The discrimination game itself provides feedback whether a particular category is successful and so it embodies environmental and ecological constraints. The language game provides feedback whether the category worked in the communication, and so it exercises a cultural constraint.

Figure 12 shows that these components lead to a satisfactory outcome. The agents reach discriminative success and communicative success[10]. The graph plots on the x-axis the number of games and on the y-axis average discriminative success (top) and communicative success (bottom). The latter goes up to 90 percent. This experiment shows therefore that cultural learning is capable of establishing a shared repertoire of words in a population. It also shows that the categories underlying the words are culturally coordinated, even though there is no telepathic access of an agent to the categories used by another agent and even though the colour categories are not innately given "at birth".

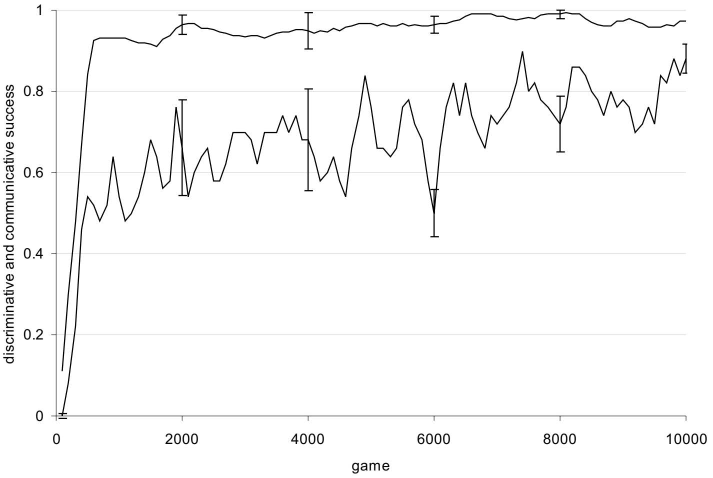

Figure 12: Average discriminative success $\mathcal{DS}$ (top) and average communicative success $\mathcal{CS}$ (bottom) for a population of 10 agents of which the colour categories are learned under influence of linguistic communication.

Figure 13 looks at the similarity between the categorical repertoires of the agents.

We see that now the agents do have similar repertoires - in contrast with the experiment in individualistic learning (section 3.2). This is due to the structural coupling between the category formation process and language. Success (or failure) in language communication feeds back into whether new categories are created or maintained in an agent's repertoire. So this experiment shows that the Sapir-Whorf thesis, advocating a causal influence of language on category acquisition (Sapir, 1921; Whorf, 1956), is entirely feasible from a theoretical point of view. Even more so, it shows that only due to such a causal influence will the agents develop a sufficiently shared categorical repertoire to allow successful communication. This does NOT imply the colour categories are not influenced by embodiment and statistical structure of the environment also. Hence these results do not imply that colour categories are arbitrary. The point is simply that language communication is a very effective way for a population of agents to go the final stretch in arriving at a shared categorical repertoire.

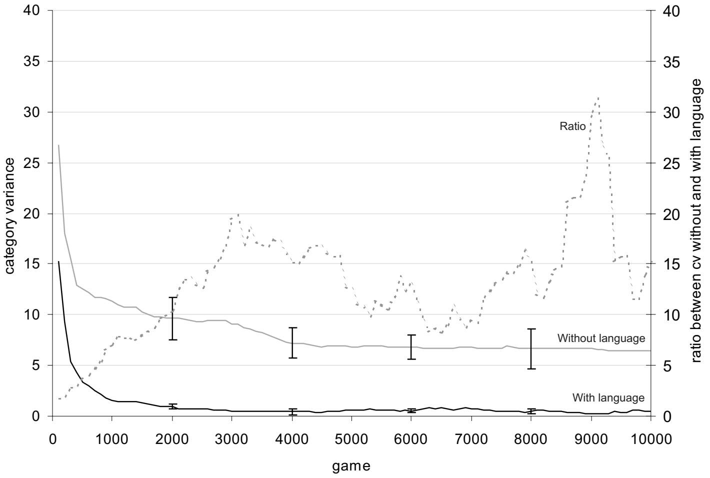

Figure 13: The category variance for the simulation described in figure 12. To show the influence of language, the category variance for exactly the same circumstances but now without language is plotted as well. The ratio between the two clearly demonstrates how the similarity of the colour categories is drastically increased by using language.

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 (as they do in the experiments discussed in section 3.2) 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 coupling category formation to language. Even with the same environmental, physiological and ecological constraints, two populations without contact with each other would develop different colour categories and consequently colour names with different meanings. Multiple solutions are possible but only one solution gets culturally frozen and enforced through language in each population. This is further illustrated in table 3 which shows that the inter-population coherence between agents in one population is high, but between populations it is far lower.

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 3: Inter-population category variance of 5 populations of which the categories are *learned under linguistic pressure*.

To conclude this section, we examine what happens when populations with this kind of semiotic dynamics 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 14 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). They are among the first concrete computer simulations showing how the memetic evolution of language and meaning are possible (Dawkins, 1976; Blackmore, 1999).

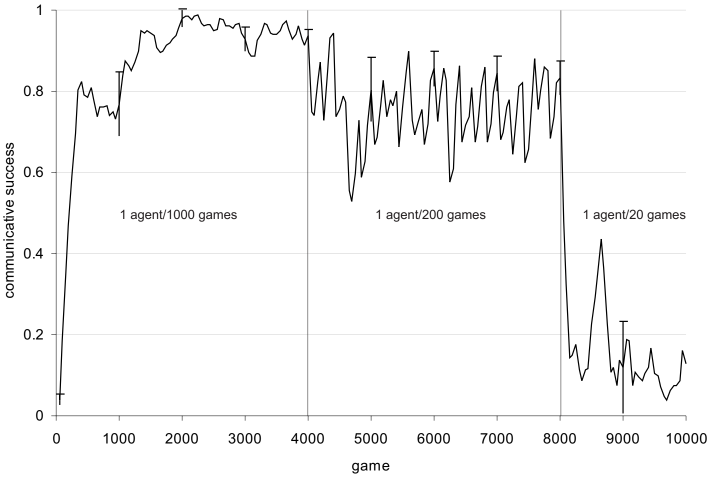

Figure 14: Illustration of memetic evolution in a population of 5 agents. In each game the context consists of 4 stimuli chosen from the complete Munsell set. A flux is introduction by replacing an agent after n games (where n=1000, 200 and 20 respectively). Too high a flux destabilises the communicative success.

We conclude that

1. Cultural learning leads to the development of an adequate repertoire of colour categories and an adequate repertoire of colour terms.

2. The colour categories are shared among the members of a population.

3. The colour categories are not shared across populations.

4.4 Genetic evolution

The next experiment tests the potential influence of language on the genetic evolution of colour concepts. It uses the same genetic model as used in section 3.3, and the learning algorithm for the acquisition of colour words explained in section 4.2. Rather than using discriminatory success to determine fitness, communicative success is used, so that the colour repertoire of the agents, genetically encoded in their genes, is not only influenced by physiological, environmental and ecological constraints but also by cultural constraints as embodied in language, despite the fact that the lexicon itself is not genetically transmitted but learned by each generation. The agents that remain in the population keep their lexicons so that they can be acquired by the new agents resulting from mutation.

Figure 15 shows the outcome of the experiment. It displays communicative success for successive generations of agents (bottom graph) and the discriminative success (top graph). As discrimination is a prerequisite for further communication, communicative success can only be reached when there is also discriminative success. We see that the same sort of results are obtained as in the previous models. The agents manage to evolve a shared repertoire of colour concepts - although now they do it in a genetic way - and evolve a language for expressing these concepts - in a cultural way.

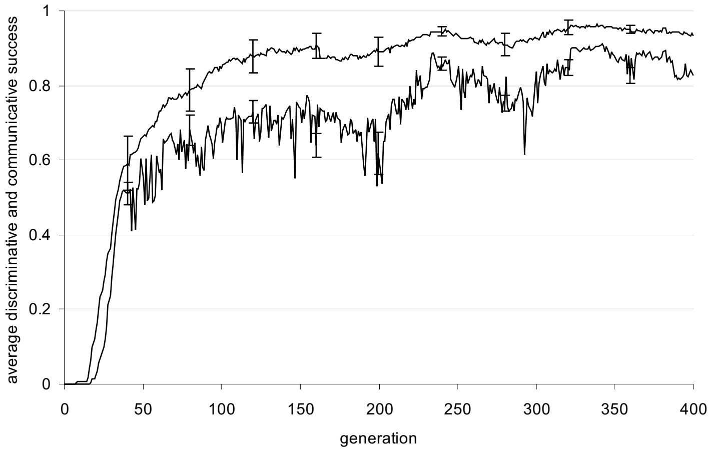

Figure 15: Average discriminative success $\mathcal{DS}$ (top graph) and average communicative success $\mathcal{CS}$ (bottom graph) for a population of 20 agents of which the colour categories are *evolved*. The fitness of the agents is based on success in the guessing game.

As in the previous genetic evolution experiment, the colour repertoires of different populations (and of course also the vocabularies that emerge) diverge, even if the same physiological, environmental and ecological constraints are used. As explained earlier, the randomness inherent in genetic evolution causes the exploration of different parts of the search space. When cultural factors play a role in fitness, as is the case here, this divergence is even more pronounced.

We conclude that

1. Genetic evolution leads to the development of an adequate colour repertoire of colour categories, even if the selectionist force includes learned cultural habits.

2. The colour categories are completely shared among the members of the population.

3. The colour categories are not shared across populations.

5 The Role of Chromatic Distributions

We can already draw a number of conclusions from the experiments so far. The first two results constitute a necessary baseline that prove that our models satisfy at least minimal working conditions:

1. The self-organisation of a shared lexicon in a population was shown to occur through adaptive language games. The learning process must include a positive feedback loop between the choice of which words to use and their success in use (section 4.3).

2. The formation of a repertoire of colour categories was shown to occur through consecutive discrimination games, both for individual learning (section 3.2) and for genetic evolution (section 3.3).

The next results are about the possible causal influence of language on category acquisition:

1. Language may have a causal influence on category acquisition, both in the case of cultural learning, if there is a structural coupling between success in the language game and adoption of categories by the agents (section 4.3), and in the case of genetic evolution (section 4.2), including if the fitness function integrates communicative success (section 4.4).

2. When there is this causal influence, the colour categories of agents within the same population become coordinated in the case of cultural learning because of the strong structural coupling between concept acquisition and lexicon formation (section 4.3). Colour categories also become shared within the same population in the genetic evolution model, because of the proliferation of 'successful' colour categorisation genes (section 3.3).

3. On the other hand sharing *across* populations did not occur neither for genetic evolution nor for cultural learning. Genetic evolution necessarily incorporates randomness in the search process which causes divergence as soon as two populations develop independently, even when exactly the same constraints are active. Different ecological and cultural circumstances, which are inevitable in split populations, will only increase this divergence (section 4.4). Learning adapts even faster to ecological and cultural circumstances and so as soon as these circumstances diverge, colour categories diverge as well (sections 3.2, 4.3).

So both a cultural learning hypothesis (with causal influence of language on category acquisition) and a genetic evolution hypothesis (with integration of communicative success into fitness) could explain how agents in a population can reach a shared repertoire of categories and a shared lexicon for communicating about the world using these categories. The difference between the two models appears to be in terms of the time needed to adapt to the environment or reach coherence. Genetic evolution is orders of magnitude slower than cultural learning and so it could only work when almost no change takes place in the environment nor the ecology of the agents. The larger the population and the more it is spread out, the longer it takes for genes to become universally shared. Moreover genetic evolution requires that a lot more information is stored in the genome, and that the developmental process will be more complex, as it requires fine-grained genetic control of neural micro-circuits (including genetic coding of the weights in networks). We leave it up to geneticists and neurobiologists to judge the plausibility of such an assumption in the case of humans (Worden, 1995). But there can be no doubt that for designing autonomous robots the cultural learning solution is preferable.

However we have not examined yet what happens when the sensory data presented to the agents has a statistical structure. That might also lead to the creation of a repertoire of shared categories - even in the absence of language interaction. So we will now introduce samples taken from real world scenes as stimuli. This will allow a fair examination of the empiricist argument that colour categories are coordinated precisely because the real world environment has enough statistical structure so that any kind of clustering algorithm (and ipso facto a neural network that embodies a statistical clustering algorithm) would allow the population to arrive at shared categories. It would also give support to the nativist position because environmental constancy and regularity is required for genetic evolution to zoom in on these statistical regularities (e.g. Shepard, 1992).

5.1 Categories from Real World Samples

Chromatic data of natural surfaces and the frequency with which these stimuli occur in natural scenes are available (see e.g. Hendley and Hecht 1949; Burton and Moorhead 1987; Howard and Burnidge 1994), but it is obviously difficult to get data reflecting the ecological importance of colour stimuli for a particular culture and thus the data can never show what aspects of real world scenes people actually pay attention to. Nevertheless, Yendrikhovskij (2001) has investigated how colour categories can be extracted from the statistics of natural images. He uses a clustering algorithm to extract colour categories from a sample of natural colours, and concludes not only that categories can be reliably extracted, but also that the extracted colour categories resemble the basic colours identified by universalists, and that this is due to the chromatic distribution of the perceived environment. Also, increasing the k parameter (where k is the number of desired clusters) leads to a growing set of categories which more or less corresponds to the evolutionary order as proposed by Berlin and Kay (o.c.). This is a very important and relevant result for the present discussion and so we decided to replicate it.

The neural networks used in previous sections for modeling categorisation are sensitive to the statistical distribution of colours in the environment. Indeed, Radial Basis Function networks (on which the categorical networks are based) stem from linear models research in statistics and have been generally used to induce a function from sample input-output pairs (Medgassy, 1961). It therefore makes sense to use real world colour samples as source of data in discrimination games and see what categories come out. We have collected two batches of data: one from natural environments and another one from urban environments. The natural data set contains 25,000 pixels drawn randomly from photographs of animals, plants and landscapes, while the urban data contains 25,000 pixels drawn from photographs of buildings, streets, traffic, shops and other urban scenes. Both data sets have a specific distribution, with an abundance of lowly saturated colours and much fewer highly saturated colours (as already observed by Hendley and Hecht, 1949). To allow comparison, a third data set containing 25,000 uniformly random sampled Munsell chips is also used. All constraints on embodiment used in earlier experiments, including the use of the CIE $L^*a^*b^*$ colour appearance model have been maintained.

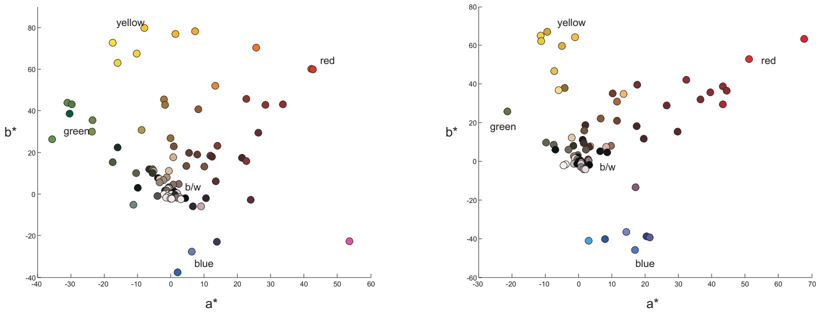

Figure 16: Results of discrimination game experiments for natural (left) and urban data (right). The centroids of all colour categories of 10 agents are plotted. Agents arrive at focal points which are more constrained than for random data but not sufficiently to explain sharing.

Results of discrimination games with these data are shown in figure 16. The left side shows the focal points of 10 agents for natural environments, and the right side the same for urban environments. Agents were left to play discrimination games until they each reached on average eleven categories[11]. Results from another experiment where agents were given samples from a randomly distributed data set are shown in figure 17. For reference, the location of human basic colour categories are shown as well in all diagrams (Sturges and Whitfield, 1995).

We see that the statistical structure in the data clearly helps the agents to reach a higher degree of categorical sharing than would otherwise be the case. There is for example a clustering around the origin $a^* = b^* = 0$ for both natural and urban environments, whereas we do not see these clusters in randomly distributed samples. This comparison is made more precise in figure 18. Notice however that there is still significant categorical variance between the agents exposed to the same type of environment.

These results clearly show that even if there is a statistical structure, there is increased sharing but the sharing is surely not complete, neither among the

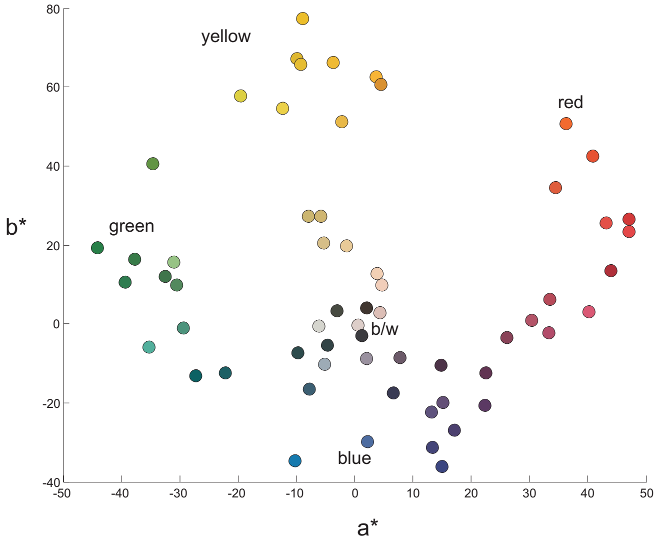

Figure 17: Results of experiments in statistical learning of colour categories for random data. Note how categories are spread out over the colour space.

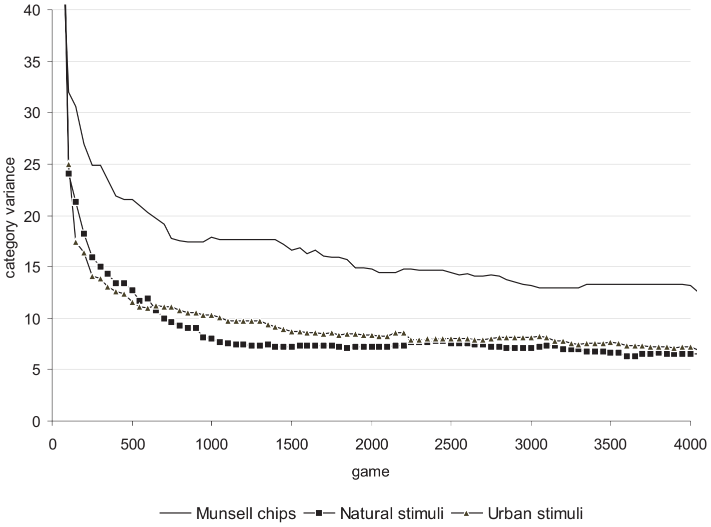

Figure 18: Category variance for three runs of 10 agents playing discrimination games. The agents have in each run been offered different kinds of colour stimuli: Munsell stimuli, having a uniform statistical distribution, and natural and urban colour stimuli, having a non-uniform distribution. The statistical structure of the natural and urban stimuli aids the agents in achieving more coherent categories.

members of the population nor among different populations. There are several reasons why this is the case. Although the natural and urban data now have natural chromatic distributions, there is random sampling going on within these data sets so agents within the population do not get exactly the same data series. The opposite would be a very unrealistic assumption anyway, both for human beings and for autonomous agents. Second, the influence of the two environments (urban versus natural) works also against sharing, simply because the statistical structure of the two environments is different. Anthropological observations show however that individuals growing up in different environments but speaking the same language have the same colour categories, and vice versa, individuals growing up in similar environments but speaking different languages often have diverging colour categories – cfr. Papua New Guinean cultures (Kay et al., 2003).

It could be argued that the sensitivity observed here is due to the specific clustering method used, namely discrimination games and adaptive RBF networks. But this is not the case. We applied the clustering algorithm used by Yendrikhovskij (2001) and used his method of sampling, and similar results were obtained. Figure 19 shows the category variance[12] for categories extracted from random, natural and urban stimuli. Categories extracted from natural and urban stimuli have approximatly half the variance of categories extracted from stimuli with a uniform distribution. From this we can conclude that learning without the influence of language in a structured environment indeed increases the sharing of categories across agents, but the sharing is never absolute.

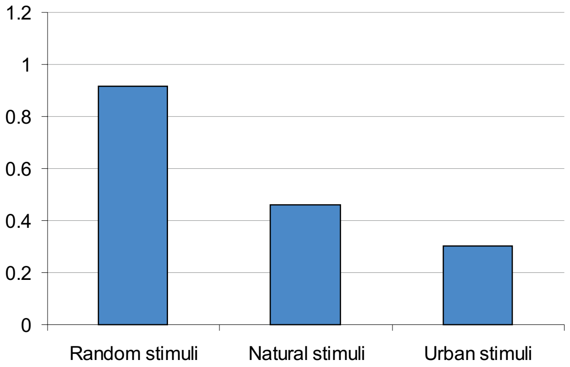

Figure 19: Category variance for categories extracted from three different types of chromatic stimuli: random, natural and urban stimuli. The agents now used a clustering algorithm instead of discrimination games and each agent extracted 11 categories.

Yendrikhovskij (2001) used the CIE $L^*u^*v^*$ colour appearance model instead of the CIE $L^*a^*b^*$ model used in this paper, and so we compared the outcome for both colour appearance models (figure 20) and even between these we see significant variation. The fact that clustering is sensitive to the colour appearance model shows that –even small– variations in colour perception, as surely occurring in humans (Gegenfurtner and Sharpe, 1999; Neitz et al., 2002), drives a purely empiricist acquisition of colour categories to diverging results.

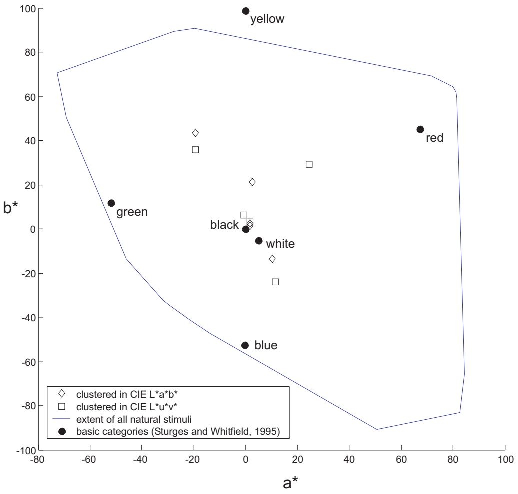

Figure 20: Clusters extracted from natural chromatic data. Five clusters are extracted in the CIE $L^*a^*b^*$ space (diamonds) and five are extracted in the CIE $L^*u^*v^*$, but then mapped onto and displayed in the $L^*a^*b^*$ space (squares). The clusters differ to a large extent, demonstrating how the colour space influences the clustering.

Perhaps even more importantly, the categories that agents end up with (still using Yendrikhovskij's clustering algorithm and the same data sets) vary significantly both with respect to natural versus urban environments and with respect to the basic human colour categories proposed in the literature (Sturges and Whitfield, 1995). This is shown in table 4 below. It shows the correlation[13] between (a) categories extracted by the clustering algorithm[14] and (b) human colour categories (as measured by Sturges and Whitfield, 1995). Perhaps surprisingly, statistical extraction of categories from natural colour data with a clear statistical structure does not deliver categories that resemble human colour categories more than do categories extracted from random data. Even more, the correlation between categories extracted from natural, urban or random colour data is approximately equal. This demonstrates that the non-uniform chromatic distributions –i.e. the urban and natural data– do not lead to categories that are similar. They correlate as much with each other as with categories extracted from random data.

	human	natural	urban	random
human	1	0.615	0.580	0.562
natural		1	.593	0.622
urban			1	0.462
random				1

Table 4: Correlation between human colour categories and 11 clusters extracted from the natural data set, the urban data set, and a random data set.

Clearly the chromatic distribution of colours in the environment can influence which colour categories are adopted by a population and how similar they are, but it is far from obvious that it alone can explain the sharing of perceptually grounded categories in a population and even less so the universal sharing of colour categories across populations. What all this means for human colour categorisation remains a matter of debate. We do not claim that inductive learning on real world environments could not potentially yield the basic human colour categories, perhaps with much more constraints on embodiment, with much greater exposure to a variety of environments, etc., but it does not seem so straightforward as often assumed.

We do claim however that the experiments allow a clear conclusion for the design of artificial agents: It would be risky to rely only on embodiment constraints and statistical clustering for forming the repertoire of perceptually grounded categories for use in communication. Inevitable variation in hardware, camera calibration, sampled data, colour appearance model, and arbitrary choices during clustering, would lead to important categorical variation between the agents or between agents exposed to different environments. It is also unlikely that (artificial) genetic evolution without integrating communication in the fitness function would work to sufficiently coordinate perceptually grounded categories. Given that we have a very straightforward and effective mechanism to coordinate categories through language (as shown in section 4), it would be irrational not to use it.

6 Conclusions

This paper examined the question how a perceptually grounded categorical repertoire can become sufficiently shared among the members of a population to allow successful communication, using colour categorisation as a case study. The paper did not introduce new empirical data but examined through formal models the consequences of adopting certain approaches which were all inspired from the study of human categorisation and naming. We explored in particular three positions: (1) All human beings are born with the same perceptually grounded categories (Nativism). So when children learn a language, their categorical repertoire is already shared with that of caregivers and they only have to learn the names of these categories. (2) All human beings share the same learning mechanisms, so given sufficiently similar environmental stimuli they will arrive at the same perceptually grounded categories which reflects the statistical structure of the real world (Empiricism). Hence the acquisition of language is again a matter of learning labels for already known shared categories and there is no strong influence of language on category formation. (3) Although learning mechanisms and environments are shared, there are still important degrees of freedom left. Language communication (or other forms of social interaction where perceptual categories play a role) helps to coordinate perceptual categorisation by providing feedback on how others conceptualise the world (Culturalism). So language now plays an important causal role in conceptual development.

As stated several times, our motivation for these investigations is to find the best way for designing agents that are able to develop a repertoire of perceptually grounded categories that is sufficiently shared to allow communication. But we believe that these results are relevant to a much broader audience of cognitive scientists who have been puzzling over the same question.

The first contribution of the paper is to introduce concrete models so that a comparison of the different positions is possible. The models have been defined in enough detail and precision to allow computer simulation. Most of the time debates on categorisation and naming have assumed particular mechanisms (for example for acquiring categories or for associating names with categories) without specifying exactly how these mechanisms were supposed to work. This has made it difficult to formulate clear arguments for or against certain positions.

The second contribution of the paper is to establish some important properties for each model: First, we have shown that the coupling of category formation with language leads to the coordination of perceptually grounded categories (both in the case of genetic evolution and in cultural evolution with learning of language), even if there is no statistical structure in the data. Second, we have confirmed that although clustering algorithms (and neural networks that embody them) are sensitive to the statistical structure of real world data, it is not so obvious that this alone can explain how perceptually grounded categories can become shared.

The models presented here could be made more complex and more realistic, integrating more constraints based on what is known about human physiology, neurological processing, brain development, genetics, language, real world environments, ecology, etc. but this complexity would be more of a hinder than a help because it would obscure the contribution of the dynamics. On the other hand, integrating all these additional constraints will be necessary to explain the kinds of cross-cultural trends that have been observed in colour naming (Kay et al., 1991; Kay and Regier, 2003) or why certain cultures have adopted particular categorical repertoires and not others.

Acknowledgements

This work was conducted at the VUB Artificial Intelligence Laboratory supported by a GOA grant from the Belgian Government. It rests on prior research by several lab members, including Paul Vogt, Joris Van Looveren, and Edwin de Jong, and by work of Frédéric Kaplan at the Sony Computer Science Laboratory in Paris. Tony Belpaeme is a postdoctoral fellow of the Fund for Scientific Research - Flanders (Belgium). Additional funds have come from the European Science Foundation OMLL project, and the CNRS OHLL project on computer modeling of the evolution of language, as well as from the EU-FET ECAgents project. The initial impetus of the work came from a seminar on colour organised in 1999 by Luc Steels at the VUB AI Lab, in which Erik Myin played an important role. Data for the natural and urban distributions were collected by Joris Bleys. We are indebted to Jules Davidoff, Serguei Yendrikhovskij, Jean-Christophe Baillie and our colleagues for their comments on this work. Also, this article has benefited to a great extent from the comments of its anonymous reviewers.

References

Belpaeme, T. (2001). 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. (2002). *Factors influencing the origins of colour categories.* PhD thesis, Vrije Universiteit Brussel, Artificial Intelligence Laboratory.

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

Bimler, D., Kirkland, J., and Jameson, K. (2004). Quantifying variations in personal color spaces: Are there sex differences in color perception? *COLOR Research and application*, 29(2):128–134.

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

Bloom, P. (2000). *How children learn the meanings of words.* The MIT Press, Cambridge, MA.

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

Bornstein, M. H. (1985). On the development of color naming in young children. *Brain and Language*, 26:72–93.

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.

Bowerman, M. and Levinson, S. C., editors (2001). *Language Acquisition and Conceptual Development.* Cambridge University Press, Cambridge.

Briscoe, T., editor (2002). *Linguistic evolution through language acquisition: formal and computational models.* Cambridge University Press, Cambridge.

Burton, G. and Moorhead, I. R. (1987). Color and spatial structure in natural scenes. *Applied Optics*, 26(1):157–170.

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

Cangelosi, A. (2001). Evolution of communication and language: using signals, symbols and words. *IEEE Transactions in Evolution Computation*, 5:93–101.

Cangelosi, A. and Parisi, D., editors (2001). *Simulating the Evolution of Language*. Springer Verlag, London.

Churchland, P. S. and Sejnowski, T. J. (1992). *The computational brain*. The MIT Press, 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.

Davies, I. and Franklin, A. (2002). Categorical perception may affect colour pop-out in infants after all. *British Journal of Developmental Psychology*, 20:185–203.

Davies, I. R. (1998). A study of colour grouping in three languages: A test of the linguistic relativity hypothesis. *British Journal of Psychology*, 98:433–452.

Davies, I. R. and Corbett, G. (1997). A cross-cultural study of colour grouping: Evidence for weak linguistic relativity. *British Journal of Psychology*, 88:493–517.

Dawkins, R. (1976). *The Selfish Gene*. Oxford University Press, Oxford.

de Boysson-Bardies, B. (1999). *How Language Comes to Children: From Birth to Two Years*. The MIT Press, Cambridge, MA.

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.

De Valois, R. L. and De Valois, K. K. (1975). Neural coding of color. In Carterette, E. C. and Friedman, M. P., editors, *Handbook of Perception, Volume V: Seeing*, pages 117–166. Academic Press, 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.

Durham, W. H. (1991). *Coevolution: Genes, Culture and Human Diversity*. Stanford University Press, Stanford.

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.

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. (1983). *The Modularity of Mind*. The MIT Press, Cambridge, MA.

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. and Sharpe, L. T., editors (1999). *Color Vision: From Genes to Perception*. Cambridge University Press, New York.

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.

Gentner, D. and Goldin-Meadow, S., editors (2003). *Language in mind*. The MIT Press, Cambridge, MA.

Gerhardstein, P., Renner, P., and Rovee-Collier, C. (1999). The roles of perceptual and categorical similarity in colour pop-out in infants. *British Journal of Developmental Psychology*, 17:403–420.

Gibbons, R. (1992). *Game Theory for Applied Economists*. Princeton University Press, Princeton, NJ.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley, Reading, MA.

Gumperz, J. J. and Levinson, S. C. (1996). *Rethinking Linguistic Relativity.* Studies in the Social and Cultural Foundations of Language 17. Cambridge University Press, Cambridge.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42:335–346.

Hassoun, M. (1995). *Fundamentals of Artificial Neural Networks.* The MIT Press, Cambridge, MA.

Hendley, C. D. and Hecht, S. (1949). The colors of natural objects and terrains, and their relation to visual color deficiency. *Journal of the Optical Society of America*, 39(10):870–873.

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.

Howard, C. M. and Burnidge, J. A. (1994). Colors in natural landscapes. *Journal of the Society of Information Display*, 2(1):47–55.

Hurford, J. R. (1989). Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua*, 77(2):187–222.

Jameson, K. and D'Andrade, R. (1997). It's not really red, green, yellow, blue: an inquiry into perceptual color space. In Hardin, C. and Maffi, L., editors, *Color categories in thought and language*, pages 295–319. Cambridge University Press, Cambridge.

Jameson, K. A. and Alvarado, N. (2003). The relational correspondence between category exemplars and names. *Philosophical psychology*, 10(1):25–49.

Kaiser, P. and Boynton, R. (1996). *Human Color Vision.* Optical Society of America, Washington DC.

Kay, P., Berlin, B., Maffi, L., and Merrifield, W. (1997). Color naming across languages. In Hardin, C. and Maffi, L., editors, *Color Categories in Thought and Language.* Cambridge University Press, Cambridge.

Kay, P., Berlin, B., Maffi, L., and Merrifield, W. R. (2003). *The World Color Survey.* Center for the Study of Language and Information, Stanford.

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 Maffi, L. (1999). Color appearance and the emergence and evolution of basic color lexicons. *American Anthropologist*, 101(4):743–760.

Kay, P. and McDaniel, C. (1978). The linguistic significance of the meanings of basic color terms. *Language*, 54(3):610–646.

Kay, P. and Regier, T. (2003). Resolving the question of color naming universals. *Proceedings of the National Academy of Sciences*, 100(15):9085–9089.

Komatsu, H., Ideura, Y., Kaji, S., and Yamane, S. (1992). Color selectivity of neurons in the inferior temporal cortex of the awake macaque monkey. *Journal of Neuroscience*, 12(2):408–424.

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.

Lammens, J. M. (1994). *A computational model of color perception and color naming.* PhD thesis, State University of New York.

Langton, C. G., editor (1995). *Artificial Life: An Overview.* The MIT Press, A Bradford Book, Cambridge, MA.

Lantz, D. and Stefflre, V. (1964). Language and cognition revisited. *Journal of Abnormal and Social Psychology*, 69(5):472–481.

Lehky, S. R. and Sejnowksi, T. J. (1999). Seeing white: Qualia in the context of decoding population codes. *Neural Computation*, 11:1261–1280.

Lenneberg, E. H. and Roberts, J. M. (1956). The language of experience: A study in methodology. *International Journal of American Linguistics*, memoir 13.

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. (1997). *Color and Cognition in Mesoamerica.* University of Texas Press, Austin.

Maturana, H. and Varela, F. (1998). *The Tree of Knowledge (revised edition).* Shambhala Press, Boston.

May, R. (1986). When two and two do not make four: nonlinear phenomena in ecology. *Proceedings of the Royal Society of London B*, 228:241–266.

Maynard Smith, J. (1982). *Evolution and the theory of games.* Cambridge University Press, Cambridge, MA.

Medgassy, P. (1961). *Decomposition of Superposition of Distributed Functions.* Hungarian Academy of Sciences, Budapest.

Minsky, M. and Papert, S. (1969). *Perceptrons.* The MIT Press, Cambridge, MA.

Mitchell, T. (1997). *Machine Learning.* McGraw-Hill, New York.

Mollon, J. D., Pokorny, J., and Knoblauch, K. (2003). *Normal and defective colour vision.* Oxford University Press, Oxford, UK.

Munsell (1976). *Munsell book of color, matte finish collection.* Munsell Color Company, Baltimore, MD.

Neitz, J., Carroll, J., Yamauchi, Y., Neitz, M., and Williams, D. (2002). Color perception is mediated by a plastic neural mechanism that is adjustable in adults. *Neuron*, 35(4):783–792.

Neitz, J., Neitz, M., and Jacobs, G. H. (1993). More than three different cone pigments among people with normal colour vision. *Vision Research*, 33(1):117–122.

Nicolis, G. and Prigogine, I. (1989). *Exploring Complexity: An Introduction.* W.H. Freeman, New York.

Nowak, M. A. and Krakauer, D. (1999). The evolution of language. *Proceedings of the National Academy of Science*, 96(14):8028–8033.

Oliphant, M. (1996). The dilemma of Saussurean communication. *BioSystems*, 37(1-2):31–38.

Parkkinen, J., Hallikainen, J., and Jaaskelainen, T. (1989). Characteristic spectra of Munsell colors. *Journal of the Optical Society of America*, 6(2):318–322.

Pinker, S. (1994). *The language instinct: How the mind creates language.* W. Morrow, New York.

Pinker, S. and Bloom, P. (1990). Natural languages and natural selection. *Behavioral and Brain Sciences*, 13:707–784.

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.

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-Heider, E. (1971). 'Focal' color areas and the development of names. *Developmental Psychology*, 4:447–455.

Rosch-Heider, E. (1972). Universals in color naming and memory. *Journal of Experimental Psychology*, 93:10–20.

Rosch-Heider, E. and Olivier, D. (1972). The structure of the color space in naming and memory for two languages. *Cognitive Psychology*, 3:337–354.

Rumelhart, D. and McClelland, J. (1986). *Parallel Distributed Processing: Exploration in the microstructure of cognition.* The MIT Press, Cambridge, MA. Volume 1 and 2.

Sampson, G. (1997). *Educating Eve: The 'Language Instinct' Debate.* Cassell, London.

Sapir, E. (1921). *Language: An introduction to the study of speech.* Harcourt, Brace and Co., New York.

Saunders, B. and van Brakel, J. (1997). Are there nontrivial constraints on colour categorization? *Behavioral and Brain Sciences*, 20(2):167–228.

Sharpe, L. T., Stockman, A., Jägle, H., and Nathans, J. (1999). Opsin genes, cone photopigments, color vision, and color blindness. In Gegenfurtner, K. R. and Sharpe, L. T., editors, *Color Vision: From genes to perception.* Cambridge University Press, New York.

Shepard, R. N. (1992). The perceptual organization of colors: An adaptation to regularities of the terrestrial world? In Barkow, J., Cosmides, L., and Tooby, J., editors, *Adapted Mind*, pages 495–532. Oxford University Press, Oxford.

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).

Sperber, D. (1996). *Explaining culture: a naturalistic approach.* Blackwell Publishers, Oxford.

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. and Shimohara, T., editors, *Proceedings of the Conference on Artificial Life V (Alife V) (Nara, Japan)*, Cambridge, MA. The MIT Press.

Steels, L. (1997). The synthetic modeling of language origins. *Evolution of Communication*, 1(1):1–34.

Steels, L. (2001a). Language games for autonomous robots. *IEEE Intelligent Systems*, sept-oct 2001:17–22.

Steels, L. (2001b). 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. 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. (1999). 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):3–32.

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*, pages 252–271. Oxford University Press, Oxford, UK.

Stengers, I. and Prigogine, I. (1986). *Order Out of Chaos.* Bantam Books, New York.

Sturges, J. and Whitfield, T. A. (1995). Locating basic colours in the munsell space. *COLOR Research and Application*, 20(6):364–376.

Suchman, L. (1987). *Plans and Situated Actions.* Cambridge University Press, Cambridge, MA.

Teller, D. (1998). Spatial and temporal aspects of infant color vision. *Vision Research*, 38:3275–3282.

Tomasello, M. (1999). *The cultural origins of human cognition.* Harvard University Press, Cambridge, MA.

Van Wijk, H. (1959). A cross-cultural theory of colour and brightness nomenclature. *Bijdragen tot de taal-, land- en volkenkunde*, 115:113–137.

Vogt, P. (2003). Anchoring of semiotic symbols. *Robotics and Autonomous Systems*, 43(2):109–120.

Whorf, B. L. (1956). *Language, Thought and Reality: selected writings of Benjamin Lee Whorf.* The MIT Press, Cambridge, MA. Edited by Carrol, J.B.

Winderickx, J., Lindsey, D., Sanocki, E., Teller, D., Motulsky, A., and Deeb, S. (1992). Polymorphism in red photopigment underlies variation in color matching. *Nature*, 356:431–433.

Wittgenstein, L. (1953). *Philosophical Investigations.* Macmillan, New York.

Worden, R. (1995). A speed limit for evolution. *Journal of Theoretical Biology*, 176:137–152.

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.

Yendrikhovskij, S. N. (2001). Computing color categories from statistics of natural images. *Journal of Imaging Science and Technology*, 45(5):409–417.

Notes

[1] k is a normalising constant, the colour spaces use relative colorimetry with $k = 0.00946300$, which is based on the standard CIE illuminant called "D65": 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.

[2] An alternative to the CIE $L^*a^*b^*$ space is the CIE $L^*u^*v^*$ space (Wyszecki and Stiles, 1982; Fairchild, 1998), which is also intended to be an equidistant colour model, meaning that colours can be compared using a simple distance function (something which is not possible in other colour spaces such as CIE XYZ or RGB, the last one being a the technical colour representation used in colour display devices such as television and computer monitors).

[3] The results are not very sensitive to different values of σ within a certain range. In the simulations reported here, σ is fixed to 10. The adaptive networks do not share locally reactive units, however this does not mean that they cannot have units sensitive to the same region in the colour space.

[4] Alternatives could be considered for the representation of the colour categories. One possibility would be to implement categories as single points in colour space. In addition with a distance metric, this representation would exhibit most properties associated with perceptual categories. However, categories would have a spherical membership function in the colour space, this is an assumption we would not like to make. Another alternative, which avoids this, uses k-nearest neighbour classification (Mitchell, 1997). Here a category is made up of several examples of colour stimuli, and classification of a stimulus happens through measuring the distance between the stimulus and the exemplars belonging to each category.

[5] For the category variance measure (eq. 9 and 12) a distance metric D between two category sets is needed. For this we first define a distance metric d between two point sets $A = \{a_1, \ldots\}$ and $B = \{b_1, \ldots\}$,

$$d\left(A,B\right)=\frac{\sum\limits_{a\in A}\min\limits_{b\in B}\left\|a-b\right\|+\sum\limits_{b\in B}\min\limits_{a\in A}\left\|a-b\right\|}{\left|A\right|.\left|B\right|} \tag{16}$$

This distance metric d has the following properties. (1) The distance between two identical sets is zero, $d(A,A)=0$. (2) The distance is symmetrical, $d(A,B)=d(B,A)$. (3) The distance is non-negative, $d(A,B)\geq 0$. (4) The sets need not have the same number of elements.

Recall that a category consists of locally reactive units with a central value $\mathbf{m}$ and a weight w. The distance between two categories c and c' can be computed as the weighted distance between the central values of the locally reactive units. We define the distance between two categories as

$$d_{\text{category}}(c,c')=d(\{\mathbf{m}_1,\ldots,\mathbf{m}_n\},\{\mathbf{m'}_1,\ldots,\mathbf{m'}_m\}) \tag{17}$$

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

where n and m are the number of locally reactive units in respectively category c and c'.

An agent has a *set* of categories, the distance D between two category *sets* of agent A and agent A' is defined as

$$D\left(A,A'\right)=\frac{\sum\limits_{c\in A}\min\limits_{c'\in A'}d_{\text{category}}(c,c')+\sum\limits_{c'\in A'}\min\limits_{c\in A}d_{\text{category}}(c,c')}{\left|A\right|.\left|A'\right|} \tag{18}$$

where $|A|$ and $|A'|$ are the number of categories of agent A and A' respectively. Note that the distance measure is sensitive to the number of categories: more categories result in a lower $D\left(A,A'\right)$ value. The category variance is therefore necessarily a relative measure —to be interpreted by comparing it to other category variances— rather than an absolute measure.

[6] The learning rate β is a positive value and is by default $\beta=1$. It determines how fast weights of the locally reactive units increase in reaction to the successful use of the category. β is not critical to the results attained, but should be set such

that it balances the decay rate α of weights in eq. 11. α takes care of a slow forgetting of categories, and is set by default to $\alpha = 0.1$.

[7] The baseline discriminative success –i.e. the chance success that agents would achieve by randomly creating categories– depends is proportional with the number of categories of an agent and inversely proportional with the size of the context. The baseline discriminative success can be estimated numerically; in this particular example it is 0.26 at game 1000.

[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.

[10] The baseline average communicative success is always lower than the average discriminative success. When agents do discriminate the stimuli in the context perfectly and when they are able to interpret the communicated words, the baseline communicative success will never be lower than $1/\mathrm{size\,of\,context}$, i.e. the hearer's success of randomly guessing the topic. Communicative success in most circumstances never reaches 100%: some topics are located just between two categories, and subsequently two agents might classify the topic with categories having different colour terms, which makes the guessing game fail. Just like argueing over the colour of ones shirt, the agents do not always agree on what category a stimulus belongs to.

[11] Berlin and Kay (1969) say that there are eleven basic colour categories, but other than that there is not specific reason why we let the agents play discrimination games until they have on average 11 categories.

[12] The category variance for categories extracted with a clustering algorithm is computed in the same way as the category variance for adaptive networks; see eq. 9, in which D is now defined as eq. 16. Note that the category variance reported in figure 19 can not be compared to category variance values elsewhere, as the distance measure is different in this case.

[13] The correlation measure used is the Kendall's Tau-b correlation. We chose this

measure as it is a non-parametric test and does require the data to have a normal distribution. The test returns values between -1 and 1. A value of 1 indicating that the correlation is perfect, and a value of -1 that the correlation perfect but inverse. Values in-between -1 and 1 indicate a correlation to a lesser degree, with 0 signifying that is no correlation between the data.

[14] The cluster algorithm used here is the k-nearest neighbour algorithm (Mitchell, 1997), as also used by Yendrikhovskij:2001. It extract k clusters from a set of values using an iterative optimisation method. First, $k = 11$ clusters were extracted from each data set (the nature data set, the urban data set and a data set containing random colours). Then, the extracted centroids of these clusters were taken to compute the correlation with human colour categories. For this, the centroids needed to be matched with the human colour categories: this was done by an exhaustive search to find the optimal match. Next, correlations were computed in the L^*, a^*, b^*, C^*_{ab} and H_{ab} dimensions (with C^*_{ab} and H_{ab} being the chroma and hue of the CIE $L^*a^*b^*$ space, see Wyszecki and Stiles, 1982). Each correlation reported in table 4 is the mean of these five correlations.