BEHAVIORAL AND BRAIN SCIENCES (2005) **28,** 469–529
Printed in the United States of America

Coordinating perceptually grounded categories through language: A case study for colour

Luc Steels
Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Pleinlaan 2 – 1050 Brussels; SONY Computer Science Laboratory, 75005 Paris, France.
steels@arti.vub.ac.be

Tony Belpaeme
Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Pleinlaan 2 – 1050 Brussels.

Abstract: This article proposes a number of models to examine through which 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 we take no stance on which position is to be accepted as final truth with respect to human categorisation and naming, we do point to theoretical constraints that make each position more or less likely and we make clear suggestions on what the best engineering solution would be. Specifically, we argue that the collective choice of a shared repertoire must integrate multiple constraints, including constraints coming from communication.

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

1. Introduction

This target article considers how a perceptually grounded categorical repertoire can become sufficiently shared among the members of a population to allow successful communication. For example, how do colour categories like "red" or "purple" 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 goal is entirely practical. We want to find out how to design artificial embodied agents (robots) that are able to do this task. Although the artificial agents might end up with a quite different categorical repertoire compared to the repertoires of human beings, it is intriguing and challenging to investigate under what circumstances the artificial agents would arrive at humanlike solutions, as this would enable communication with humans. Because the agents will be considered to be autonomous and distributed, we cannot assume telepathy or 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 natural system achieving shared perceptually grounded categorisation and communication based on a rich open-ended repertoire of categories. Moreover, if we

Luc Steels is Professor of Artificial Intelligence at the University of Brussels (VUB, Belgium), since 1983, and director of the Sony Computer Science Laboratory in Paris since 1996. His research has spanned a broad range of topics in Artificial Intelligence, focusing most recently on the origins of language and meaning by doing computational and robotic experiments. After detailed studies of lexicon formation in coevolution with category acquisition, he is now developing the Fluid Construction Grammar framework for Experiments in the emergence of grammar.

Tony Belpaeme obtained his Ph.D. in 2002 from the Vrije Universiteit Brussel (VUB), Belgium, where he then worked as a postdoctoral researcher at the Artificial Intelligence Laboratory. In 2005, he took up the position of Senior Lecturer in Intelligent Systems at the University of Plymouth (U.K.). His current research interests include cognitive robotics, the evolution of language, and the application of linguistic cognition to concept acquistion and intelligent systems in general.

© 2005 Cambridge University Press 0140-525X/05 $12.50

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

can generate categorical repertoires that are similar to those of humans, then communication between humans and artificial agents will be more feasible. The question of how a population might coordinate its 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 (Briscoe 2002; Cangelosi & Parisi 2001) and which 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 are 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 categorization must be genetically determined and the relevant genes must have developed through evolution by natural selection. This position is historically associated with rationalism (Fodor 1983) and is often found explicitly or implicitly in evolutionary psychology (Durham 1991; Pinker & Bloom 1990; Shepard 1994). Adopting this position for the design of artificial agents means that we must simulate genetic evolution (Fogel 1999; Goldberg 1989; Holland 1975; Koza 1992). 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 & 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 repertoires 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., Bowerman & Levinson 2001; Gentner & Goldin-Meadow 2003; Gumperz & Levinson 1996).

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 given not only the 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 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 & 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 to and independently of language (Harnad 1990) without direct linguistic or categorical feedback (Bloom 2000). So the engineer is not given a clear choice of 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, e.g., olfactory or gustatory experience) and because colour is well understood as a physical phenomenon (Wyszecki & Stiles 1982/2000). 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 & Sharpe 1999). In recent years, it has become clear that colour perception is perhaps more variable within normal subjects than previously thought (e.g., see Bimler et al. 2004). Results from molecular genetics show that there are several allelic variants of opsin genes, and that 15 to 20% of Caucasian females have the genetic potential to be tetra-

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

chromatic instead of trichromatic (Mollon et al. 2003; Neitz et al. 1993; Sharpe et al. 1999; Winderickx et al. 1992).

The impact of the variation of the neural substrate on colour perception, colour categorisation, and colour naming is still being investigated. But it is a reason from an engineering viewpoint that indicates 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 our understanding of 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 prelanguage children (Bornstein 1975; Bornstein et al. 1976; Davies & Franklin 2002) and in experiments where individuals from one culture learn the colour categories of another culture (Roberson et al. 2000; Rosch-Heider & Olivier 1972).

Anthropologists have also tried to collect empirical data about 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 ten 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 & Kay 1969; Davidoff et al. 1999; Kay et al. 2003; MacLaury 1997; Rosch-Heider & Olivier 1972), as first introduced by Lenneberg and Roberts (1956). The studies provided the following results: (i) The naming experiments, requiring informants to point to the best example for one of the "basic" colour words in their language, consistently showed 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 & Kay 1969). (ii) The memory experiments, which require informants to pick out a colour sample seen earlier, show that samples which are closer to focal points are better remembered than those closer to the boundaries (Rosch-Heider 1971; 1972).

Based on data of naming experiments and memory experiments, Berlin and Kay (1969) 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 crosslinguistic 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 Davidoff et al. (1999), Roberson et al. (2000), and Davidoff (2001) have presented evidence through the same sort of memory and naming tests that the focal points of English and the language of the Berinmo (a Papua New Guinea tribe) are substantially different and that Rosch-Heider's data have 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., Lucy 1997; Sampson 1997; Saunders & van Brakel 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 by, among others, Berlin and Kay (1969), Kay et al. (1991), Shepard (1992), Pinker (1994), and Kay and Maffi (1999), based on the identification of universal trends in colour categorisation. Language plays no role in this. As Pinker (1994) put it: "The way we see colors determines how we learn words for them, not vice versa" (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 2001b). The culturalist view on colour categorisation and colour naming has its own defendants; see, for example, Lucy and Shweder (1979), Gellatly (1995), Davies and Corbett (1997), Davies (1998), Dedrick (1998), and Jameson and Alvarado (2003), among others.

1.2. Objectives

The present article 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 is usually believed (see the references given in sect. 1.1, para. 2), 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 (e.g., to calibrate perception to context), and so forth. Moreover, there are also constraints from the kinds of neural processes that are used for categorization 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 into artificial vision systems. We will do so in all the experiments reported in this paper.

2. *Constraints coming from the world.* Although there is

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

significant variation in the environments in which human beings find themselves (compare growing up at the North Pole versus 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. This adaptation 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 (such as Mars) they 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 are 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, another one may decide to drive on the right side. Speakers of English have agreed to call a particular hue "blue" but they could just as well have called it "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. All of which 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. But this can only be when not only physiological constraints (such as those due to the retinal receptors) but also the colour categories themselves are genetically determined – in other words, 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, the empiricists 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 roles. 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, thereby making 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 needed 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 article are very common in many sciences but still surprisingly controversial for psychologists. For example, there is now a large body of game theoretic models that have revolutionized 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; they may show, for example, 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; nevertheless, a lot can be learned about the possibility of certain outcomes or their plausibility. These models might help us to infer the effects of certain consumer behaviours on specific business models, without evidence of whether consumers actually exhibit these behaviours. Similar theoretical approaches are now widespread in biology, where, 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 target article is in the same line and uses the same methodological tools. The verbal interactions between the agents are modeled as multiagent 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. In each of our models we use the same neural networks for categorisation (radial basis function networks). 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 of our models incorporate the 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 we examine whether colour categories become shared

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

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 whether 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 multispectral camera. The use of computer simulations for examining the behaviour of complex systems is common in all the sciences of complexity, including nonlinear physics (Nicolis & Prigogine 1989) and 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 are 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 (sects. 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 forces the question of 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 in coordinating the 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 helps 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 conclude 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 target article (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 nonverbal gestural feedback (pointing). The agent population is an example of a distributed multiagent 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 the 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), they are embodied, because the apparatus and subsequent processing reflects human physiology (Kaiser & Boynton 1996), they are situated, because the games are embedded in the context of communicative acts in a shared real-world setting (Suchman 1987), and they are 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 (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 (Fogel 1999; Goldberg 1989; Holland 1975; Koza 1992).

Individualistic learning is modelled by a process in which the categorical repertoires and lexicons of the agents

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

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 1,269 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). The simulations do not use monochrome colour samples or random values in RGB or another colour space; instead, they start from realistic colour data. In each game, the agents 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 nonedible mushrooms. Mushroom eaters will therefore need to acquire the ability to distinguish these fine shades of orange. If the distinction is much clearer (e.g., because all edible mushrooms are orange and all nonedible 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 and is discussed by 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:

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

$\bar{x}(\lambda)$, $\bar{y}(\lambda)$, 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 formulas (Wyszecki & Stiles 1982/2000, p. 166).

Obviously the realism of this model can be improved. For example, Lammens (1994) started from the neural response functions proposed by De Valois and De Valois (1975) and 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 & 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:

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

Rather than 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 equation (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 & Papert 1969), where w_j is a weight factor with a range between 0 and 1:

$$y_k(\mathbf{x}) = \Sigma w_j z_j(\mathbf{x}). \quad (3)$$

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:

$$\forall c \in C : y_b(\mathbf{x}) \geq y_c(\mathbf{x}). \quad (4)$$

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 L. One word form can be associated with several categories (because the agent must be able to main-

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

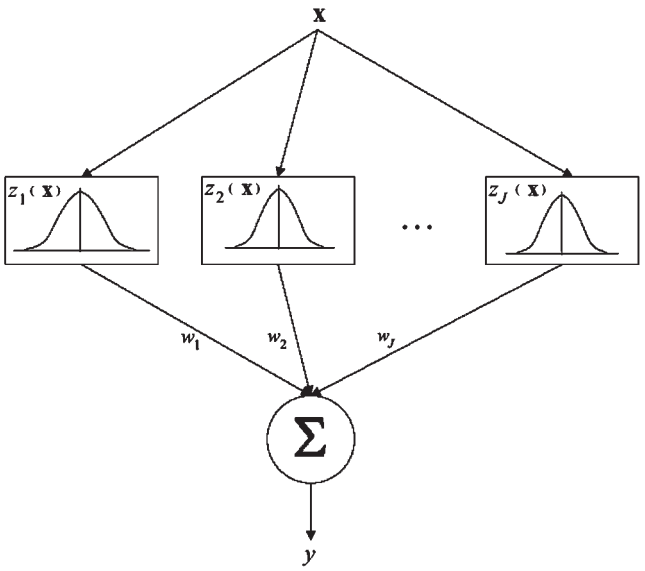


Figure 1. Categorisation is performed by an adaptive network consisting of locally reactive units fully connected to a summing output unit. Each such network corresponds to one colour category.

tain multiple hypotheses about the meaning 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 C of n categories and a set F of m word forms, this network consists of $n \times m$ relations, each having a strength $s \in [0.0, 1.0]$, so that $L = C \times F \times [0.0, 1.0]$. 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 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 that 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 roles 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 & Kaplan 1999; Steels et al. 2002), including experiments on autonomous mobile robots (Steels 2001a; Steels & Kaplan 2002; Vogt 2003).

2.4.1. The discrimination game. The discrimination game has been chosen so as to introduce the ecological dimension into the models. As already mentioned, suppose that various types of mushrooms have similar form and shape and are distinguishable only by their colour, and further suppose that 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 in which the topic is the edible mushroom and the other objects in the context are the nonedible mushrooms. So ecology is concretised through which 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, and so forth. 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 (which 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 $O = \{o_1, \ldots, o_N\}$ 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 $O = \{o_1, \ldots, o_N\}$ and the topic $o_t \in O$ are presented to the agent.
2. The agent perceives each object oi and produces a sensory representation for each object: $S_{o_i} = \{s_1{}^{o_i}, \ldots, s_N{}^{o_i}\}$. 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

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

where 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:

$$\text{count}\left(\left\{c_{S_{o_1}}, \ldots, c_{S_{o_N}}\right\}, c_{S_{o_t}}\right) = 1. \tag{6}$$

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 at a 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) and says, for example, "Could you give me the red one," whereupon the hearer has to apply the category that is the meaning of "red" to the objects in the context and identify the desired fruit. The meaning of "red" is the category that discriminates the topic from the other objects in this context. So the guessing game implies a discrimination game.

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

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 previously 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, the hearer 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 the speaker had chosen as the topic, the game is successful. If not, the speaker identifies the topic and the hearer adapts its categorical network and its lexicon as in equation (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 used in anthropological research (Berlin & Kay 1969; Lantz & Steffire 1964; Lenneberg & Roberts 1956; Kay et al. 1991; 1997; MacLaury 1997; Rosch-Heider 1972) 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. It would be desirable for anthropological experiments to be made more realistic by asking subjects to name topics within ecologically valid contexts (see also Jameson & Alvarado 2003). Presenting all the Munsell chips at once is obviously an unusual problem setting for human subjects; it is 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 (described in sect. 3) assumes that there is no causal role of language in concept formation, so agents only play discrimination games. In section 4, guessing games are used 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 examine colour samples drawn from real-world data where a clear chromatic structure is present.

3. Learning without language

We saw earlier that there could be two approaches to the problem of how concepts are acquired: either they are learned or they are innately present, the latter implying that they have evolved through genetic evolution. The two possibilities are explored in sections 3.2 and 3.3, respectively. The discrimination context is the same for both experiments and consists of four stimuli chosen from a total of 1,269 Munsell chips. In the learning case, the agents adapt their categorical networks during their lifetime in the spirit of connectionist learning systems (Churchland & Sejnowski 1992). In the genetic evolution case, the agents have a fixed network and change takes place only 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: (i) If the colour sample is uniquely categorised, then agent A is capable of discriminating the topic from the other colour samples, and the discriminative success for game i is $ds^A_i = 1$. (ii) If no unique category is found for the topic, the discrimination game has failed; $ds^A_i = 0$.

The discriminative success of the agent for a specific environment ideally reaches 100%. 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

$$DS^A_j = \frac{\sum ds^A_i}{n}. \tag{7}$$

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

$$DS_j = \frac{\sum DS^A_j}{m}. \tag{8}$$

The category variance cv between the categorical repertoires of the different agents is measured by computing the

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

cumulated distance between the categories of the agents of a population $P = \{A_1, \ldots, A_n\}$, as in

$$cv(P) = \frac{1}{\frac{1}{2}n(n-1)} \sum_{i=2}^{n} \sum_{j=1}^{i-1} D(A_i, A_j), \quad (9)$$

where $D(A_i,A_j)$ 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 a function of the success of the play.

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

$$w_i = w_i + \beta z_i(S_{o_i}) \quad (10)$$

where $z_i(S_{o_i})$ is the output of unit i for the topic S_{o_i}, 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 ($C = \varnothing$); 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 ($\mathbf{m} = S_{ot}$). 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 by a small factor. The weight decay, a learning rule standard in the literature (Krogh & Hertz 1995; Rumelhart & McClelland 1986), is defined as

$$w_j = \alpha w_j, \quad (11)$$

where $\alpha \leq 1$ is a nonnegative 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.

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 (Fig. 2), a population of 10 agents plays a series of 1,000 discrimination games. The context of a game contains four colour stimuli chosen randomly from the complete set of more than 1,269 Munsell chips, of which one stimulus 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

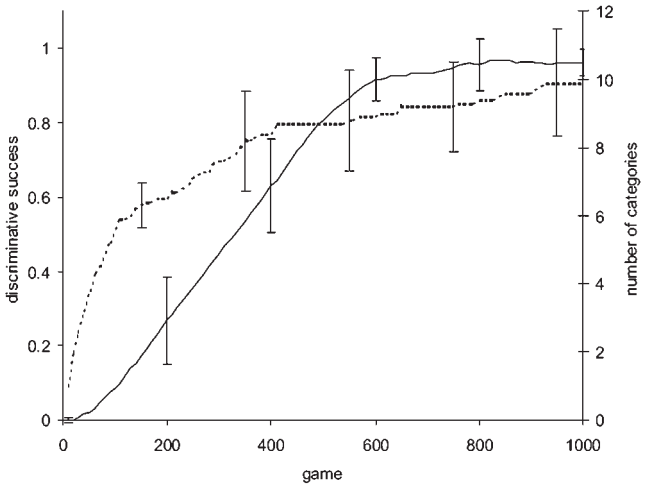


Figure 2. Average discriminative success *DS* and average number of categories (dotted line) of 10 agents playing discrimination games.

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 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 progresses, the similarity is not absolute (if all categories were similar, the category variance would be zero). This demonstrates that the constraints that 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

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

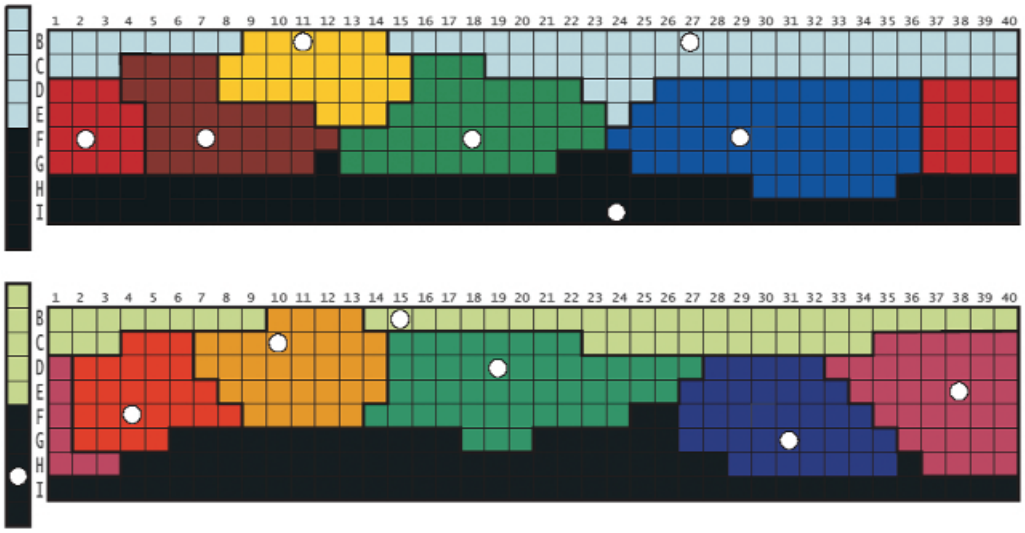


Figure 3. The maximum (white circle) and the extent (colour coding) of the categories of two agents after playing 1,000 discrimination games. The chart consists of saturated Munsell chips, following Berlin and Kay (1969). Observe how categories are distributed across the Munsell chart, and how both agents end up with different categories.

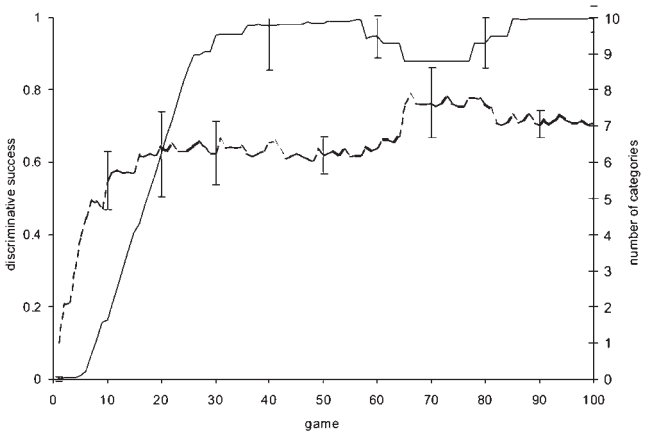


Figure 4. Average discriminative success *DS* and average number of categories (dotted line) for a population of 10 agents that *learn* colour categories. In the first 50 games the context is chosen from a simple stimuli set; after 50 games the set of stimuli is extended to increase complexity. The graph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

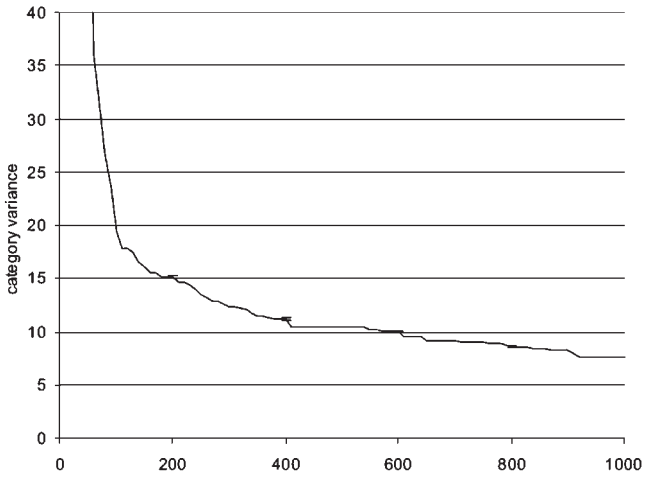


Figure 5. The category variance of a population of 10 agents playing discrimination games (for the same simulation as reported in Fig. 2). The graph shows how the categories of all agents start to resemble each other due to ecological pressure, but do not become equal.

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, the results hint why it has not been possible to explain basic colour categories based on physiological constraints alone (e.g., Gellatly 1995; Jameson & D'Andrade 1997; Saunders & van Brakel 1997). If different populations exposed to different environmental stimuli and ecological challenges were to be compared, the repertoires of the agents in the population would be even more different.

Table 1 shows the interpopulation 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'. The number of agents in the populations P and P' are n and m, respectively, which are assumed to be equal for all populations being compared:

$$cv'(P,P') = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} D(A_i, A_j'). \qquad (12)$$

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' decreases. Populations where all individuals have identical categories have $cv' = 0$.

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.

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

Table 1. *Interpopulation category variance* cv′ *of five populations for which the categories have been learned under identical experimental settings, except for the initial random seed*

cv′	A_1	A_2	A_3	A_4	A_5
A_1	9.29				
A_2	10.14	9.38			
A_3	10.62	10.51	9.62		
A_4	10.84	11.25	10.94	9.22	
A_5	10.89	11.14	10.31	11.21	9.83

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" that 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 is used to determine 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 equation (13). The centre of the added locally reactive unit is randomly chosen from a normal distribution with mean $\mathbf{c}$ and standard deviation σ.

$$\mathbf{c} = \frac{\sum w_i \mathbf{m}_{c,i}}{\sum w_i}. \qquad (13)$$

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 four stimuli chosen from a total of 1,269 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 Fig. 4), genetic evolution generates more colour categories, and after a number of generations there is again an adequate repertoire (Fig. 7). Figure 8 shows the focus and extent of the categories of two agents plotted on the two-dimensional Munsell colour chart.

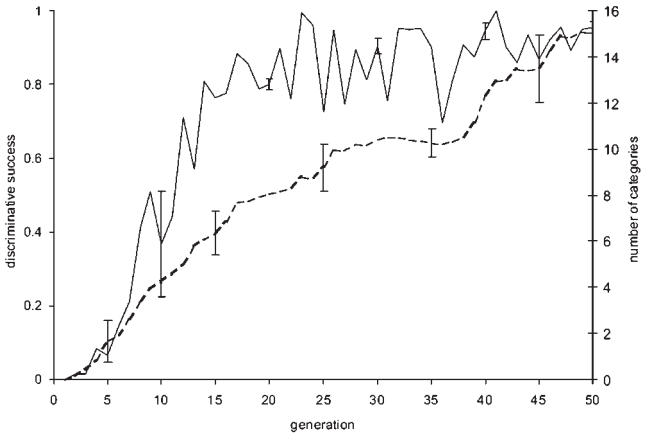


Figure 6. Average discriminative success *DS* and average number of categories (dotted line) for a population of 10 agents for which the colour categories are *evolving* in a genetic fashion.

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 repertoire within their lifetime, and adapt to environmental changes (caused, e.g., by the availability of new dyes) also within their lifetime. In the

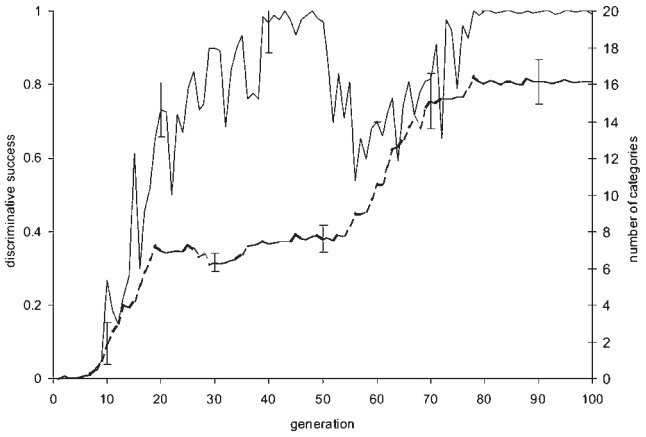


Figure 7. Average discriminative success *DS* and average number of genetically evolved categories (dotted line) for a population of 10 agents. In the first 50 games the context is chosen from a simple stimuli set; after 50 games the set of stimuli is extended to increase complexity.

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

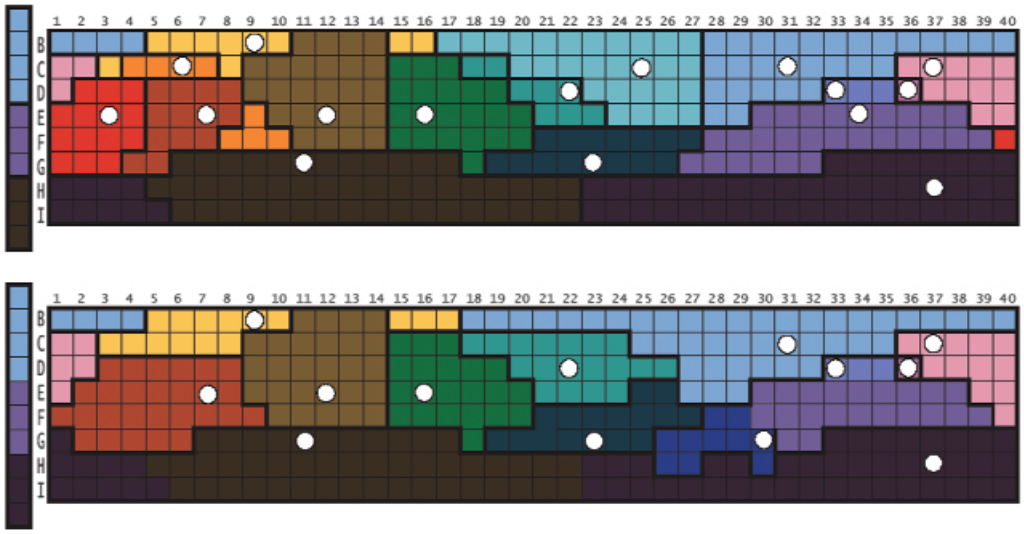


Figure 8. The maximum (white circle) and the extent (colour coding) of the categories of two agents with genetically evolved categories. Because of the dynamics of the evolutionary process, most categories of both agents are identical.

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 were applied to an equally small population of 10 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 because they are 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 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 that 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, the two repertoires are very different. This is shown in Table 2, which shows that the variation within a population is almost nonexistent ($\leq$0.40) but across populations the variation is considerable. With different ecological and environmental constraints the variation would be even more dramatic.

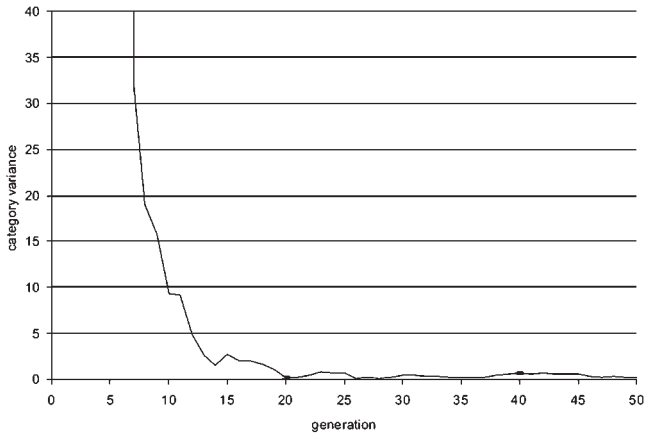


Figure 9. The category variance of a population of 20 agents after evolving for more than 50 generations (under the same conditions as in Fig. 6). It can be seen that there is hardly any variation between the categories of the agents.

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.

Table 2. *Interpopulation category variance of 5 populations for which the categories have been evolved using the discriminative success as fitness measure*

cv'	A_1	A_2	A_3	A_4	A_5
A_1	0.40				
A_2	4.91	0.40			
A_3	3.98	5.75	0.05		
A_4	3.67	4.54	4.64	0.20	
A_5	5.60	6.26	6.10	5.55	0.27

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

4. Learning with language

The previous section compared individualistic learning with genetic evolution. Both are capable of explaining how categories may be acquired by individuals, but only genetic evolution can also explain how colour concepts could become shared. In the next series of experiments, we study the impact of language (and therefore 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 in modelling both cases: learning (sect. 4.2) and genetic evolution (sect. 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: (i) The topic pointed at by the hearer is equal to the topic chosen by the speaker; therefore, the game i is a success for both agents A: communicative success $cs^A_i = 1$. (ii) 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: $cs^A_i = 0$. (iii) The game got stuck somewhere halfway, either because the speaker or the hearer did not have a discriminating category, or because 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: $cs^A_i = 0$.

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

$$CS^A_j = \frac{\sum cs^A_i}{n}. \tag{14}$$

The cumulative success CS_j for a population of m agents A for the last series of n games at game j is defined as

$$CS_j = \frac{\sum CS^A_j}{m}. \tag{15}$$

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 (Bloom 2000; de Boysson-Bardies 1999). Nevertheless, a number of mathematical and computational models show that genetic evolution can in principle do the job (Cangelosi 2001; Nowak & Krakauer 1999). 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 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 $\{f_1, \ldots, f_m\}$ 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.

(a) If the communication is 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*).

(b) If the communication is unsuccessful, the speaker decreases the strength s_{kj} by δ_{dec}.

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

(a) If the communication is successful, the hearer increases the strength s_{kj} by δ_{inc} and decreases the strength of competing words associated with the same category by δ_{inh}.

(b) If the communication is 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, the speaker creates a new word form (by generating a random combination of syllables from a prespecified 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 it 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 nonlinear dynamical systems theory (Nicolis & Prigogine 1989; Stengers & 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 who also deposit pheromone, and so there is a a positive feedback loop that 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 that use 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 nonlinear 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

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

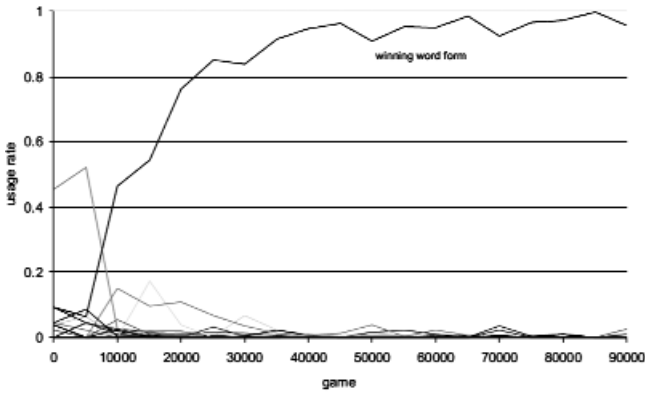


Figure 10. This graph plots the usage rate of all possible words for the same meaning in a consecutive series of language games. Initially, many words are competing until one dominates due to a winner-take-all effect.

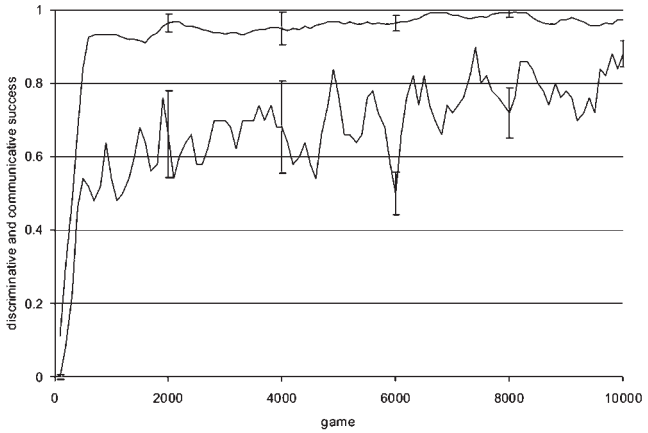


Figure 12. Average discriminative success *DS* (top line) and average communicative success *CS* (bottom line) for a population of 10 agents for which the colour categories are learned under influence of linguistic communication.

brightest object is also the one that is uniquely yellow. (This relates to Quine's [1960] well-known puzzle in which 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 is disentangled when situations arise where two meanings are incompatible – as when, for example, a bright object is blue. This type of disentanglement is also observed with the mechanisms described here; see Figure 11 (taken from Steels & Kaplan 1999), which considers this "semiotic dynamics" in more detail.

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 reenforced by increasing the weights of its network according to equation 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 & 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 the generation of a new category by the discrimination game leads to the creation of a new word. The discrimination game itself provides feedback about whether a particular category is successful and so it embodies environmental and ecological constraints. The language game provides feedback about whether the category worked in the communication, and so it exercises a cultural constraint.

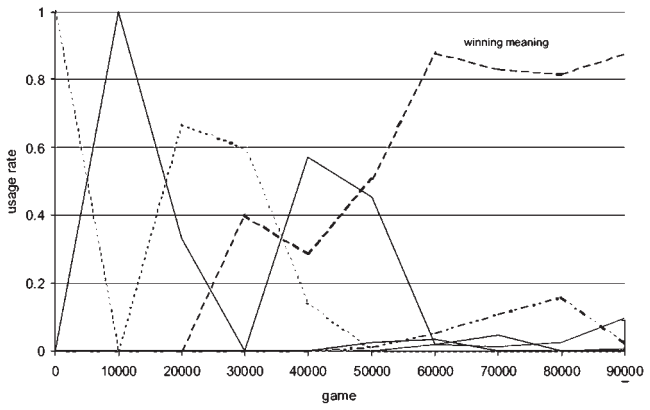


Figure 11. This graph plots (on the *y*-axis) the usage percentage of different meanings associated with the same word. Different meanings may coexist until a situation arises that disentangles them.

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%. 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 13 looks at the similarity between the categorical repertoires of the agents. We see that now the agents do have similar repertoires – in contrast to the experiment in individualistic learning (sect. 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 that 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 com-

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

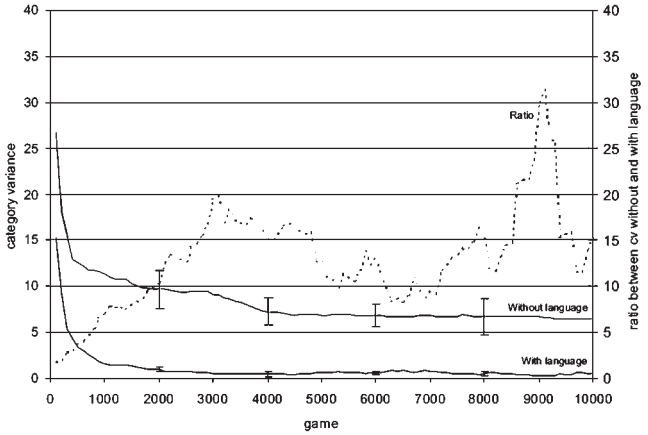


Figure 13. The category variance for the simulation described in Figure 12. To show the influence of language, the category variance is plotted as well for exactly the same circumstances but now without language. The ratio between the two clearly demonstrates how the similarity of the colour categories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

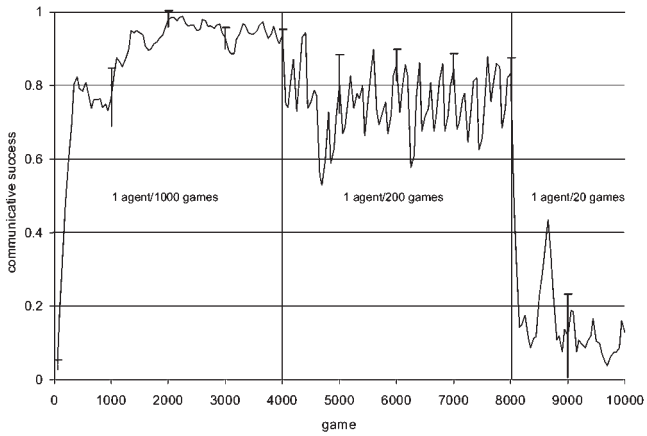


Figure 14. Illustration of memetic evolution in a population of five agents. In each game the context consists of four stimuli chosen from the complete Munsell set. A flux is introduction by replacing an agent after n games (where n = 1,000, 200, and 20, respectively). Too high a flux destabilises the communicative success.

munication is a very effective way for a population of agents to go the final stretch in arriving at a shared categorical repertoire.

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 sect. 3.2) their categories diverge too much to support communication later. So both must be learned at the same time in a coevolutionary dynamics. This shows that the Sapir-Whorf thesis is not only feasible but is 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 interpopulation coherence between agents in one population is high but between populations it is far lower.

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 or 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 get 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 is possible (Blackmore 1999; Dawkins 1976).

Table 3. *Interpopulation category variance of five populations for which the categories are* learned under linguistic pressure

cv'	A_1	A_2	A_3	A_4	A_5
A_1	0.30				
A_2	4.29	0.45			
A_3	3.83	4.52	0.36		
A_4	5.09	5.60	5.31	0.51	
A_5	5.26	5.80	5.37	6.08	0.55

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 is 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

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

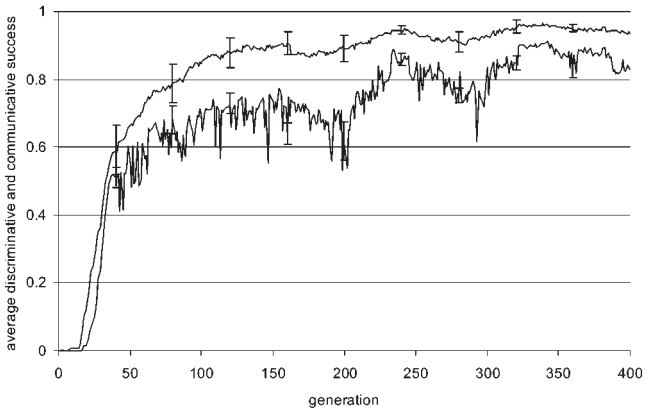


Figure 15. Average discriminative success *DS* (top line) and average communicative success *CS* (bottom line) for a population of 20 agents for which the colour categories are *evolved*. The fitness of the agents is based on success in the guessing game.

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.

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 proves our models satisfy at least minimal working conditions.

1. The self-organisation of a shared lexicon in a population has been 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 (sect. 4.3).
2. The formation of a repertoire of colour categories has been shown to occur through consecutive discrimination games, both for individual learning (sect. 3.2) and for genetic evolution (sect. 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 (sect. 4.3) and in the case of genetic evolution (sect. 4.2), including if the fitness function integrates communicative success (sect. 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 (sect. 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 (sect. 3.3).
3. On the other hand, sharing *across* populations did not occur for genetic evolution or 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 (sect. 4.4). Learning adapts even faster to ecological and cultural circumstances and as soon as these circumstances diverge, colour categories diverge as well (sects. 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 or 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 microcircuits (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.

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).

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

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., Burton & Moorhead 1987; Hendley & Hecht 1949; Howard & 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 (2001b) has investigated how colour categories can be extracted from the statistics of natural images. He used a clustering algorithm to extract colour categories from a sample of natural colours and concluded 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 that more or less correspond to the evolutionary order as proposed by Berlin and Kay (1969). 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 from urban environments. The natural data set contains 25,000 pixels drawn randomly from photographs of animals, plants, and landscapes; the urban data set 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 far fewer highly saturated colours (as already observed by Hendley & 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.

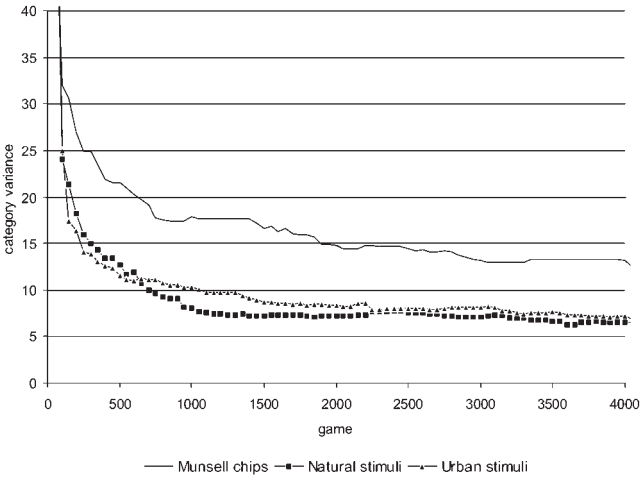


Figure 17. Results of experiments in statistical learning of colour categories for random data. Note how categories are spread out over the colour space.

Results of discrimination games with these data are shown in Figure 16, where the left side shows the focal points of 10 agents for natural environments and the right side displays the same for urban environments. Agents were left to play discrimination games until they each reached, on average, 11 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 foci are shown as well in all diagrams (Sturges & 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 members of the population nor among different populations. There are several reasons why this is the case. Although the natural and

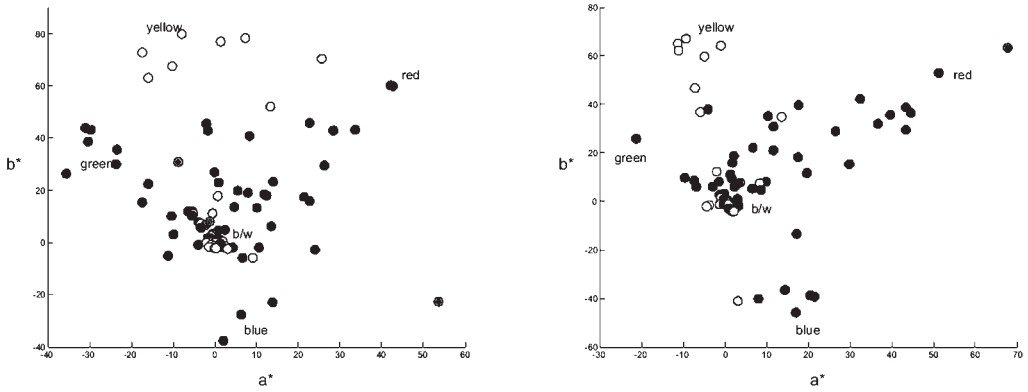


Figure 16. Results of discrimination game experiments for natural (left) and urban data (right). The centroids of all colour categories of 10 agents are plotted. Agents arrive at focal points that are more constrained than for random data but not sufficiently to explain sharing.

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

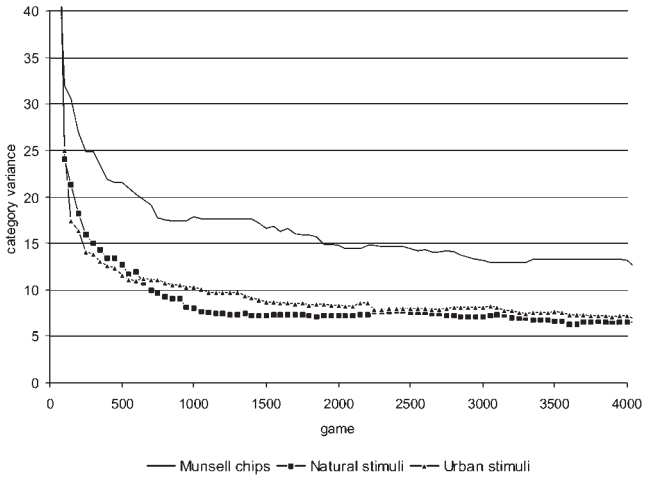


Figure 18. Category variance for three runs of 10 agents playing discrimination games. The agents have in each run been offered different kinds of colour stimuli: Munsell stimuli, having a uniform statistical distribution, and natural and urban colour stimuli, having a nonuniform distribution. The statistical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

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 (cf. 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 (2001b) 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 approximately 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.

Yendrikhovskij (2001b) used the CIE $L^*u^*v^*$ colour appearance model instead of the CIE $L^*a^*b^*$ model used in this article, and so we compared the outcome for both colour appearance models (Fig. 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 occur in humans (Gegenfurtner & Sharpe 1999; Neitz et al. 2002), drive a purely empiricist acquisition of colour categories to diverging results.

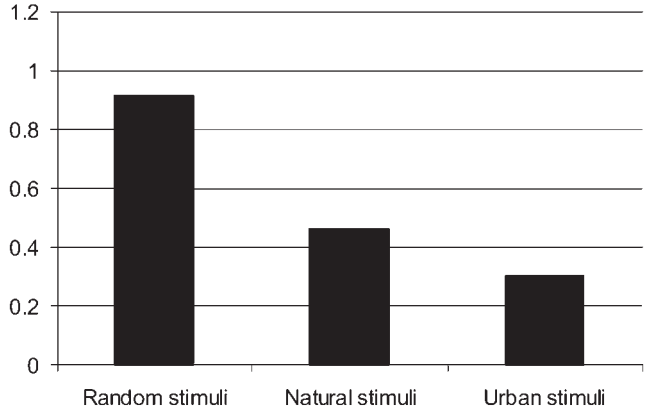


Figure 19. Category variance for categories extracted from three different types of chromatic stimuli: random, natural, and urban stimuli. The agents now used a clustering algorithm instead of discrimination games and each agent extracts 11 categories.

Perhaps even more important, 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 & Whitfield 1995). This is evident from Table 4, which shows the correlation[13] between (a) categories extracted by the clustering algorithm[14] and (b) human colour categories (as measured by Sturges & 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

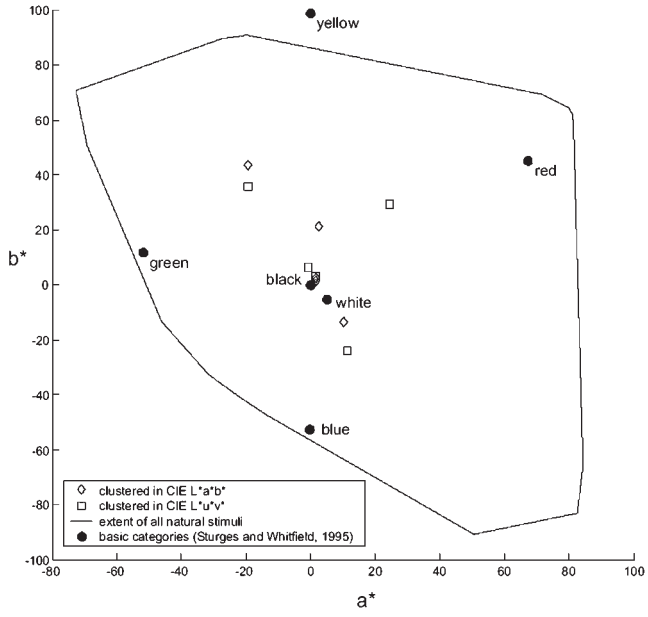


Figure 20. Clusters extracted from natural chromatic data. Five clusters are extracted in the CIE $L^*a^*b^*$ space (diamonds) and five are extracted in the CIE $L^*u^*v^*$, but then mapped onto and displayed in the $L^*a^*b^*$ space (squares). The clusters differ to a large extent, demonstrating how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

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

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*

	human	natural	urban	random
human	1	0.615	0.580	0.562
natural		1	0.593	0.622
urban			1	0.462
random				1

is approximately equal. This demonstrates that the nonuniform 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.

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 many more constraints on embodiment, with much greater exposure to a variety of environments, and so forth, 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 of coordinating categories through language (as shown in sect. 4), it would be irrational not to use it.

6. Conclusions

This target article has examined the question of 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 article did not introduce new empirical data but examined through formal models the consequences of adopting certain approaches that were all inspired from the study of human categorisation and naming. We explored in particular three positions: (i) 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. (ii) 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. (iii) 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 was 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 this target article 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 (e.g., 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 this article 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, and so forth, but this complexity would be more of a hindrance 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 & Regier 2003; Kay et al. 1991) or why certain cultures have adopted particular categorical repertoires and not others.

ACKNOWLEDGMENTS

This work was conducted at the VUB Artificial Intelligence Laboratory and 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 on the 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 proj-

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

ect, and the CNRS OHLL project on computer modeling of the evolution of language, as well as from the EU-FET ECAgents and the EU-FET Cogniron 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.

NOTES

Tony Belpaeme is currently at the following affiliation: School of Computing, Communications and Electronics, University of Plymouth, Plymouth, United Kingdom.

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 (Fairchild 1998; Wyszecki & Stiles 1982/2000), which is also intended to be an equidistant colour model, meaning that colours can be compared using a simple distance function (something that is not possible in other colour spaces such as CIE XYZ or RGB, the last one being 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, which 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 occurs by measuring the distance between the stimulus and the exemplars belonging to each category.

5. For the category variance measure, equations (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 $B = \{b_1, \ldots\}$,

$$d(A,B) = \frac{\sum_{a\in A} \min_{b\in B} \|a-b\| + \sum_{b\in B} \min_{a\in A} \|a-b\|}{|A|\cdot|B|}. \qquad (16)$$

This distance metric d has the following properties: (i) The distance between two identical sets is zero, $d(A,A) = 0$. (ii) The distance is symmetrical, $d(A,B) = d(B,A)$. (iii) The distance is non-negative, $d(A,B) \geq 0$. (iv) 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

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

with

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

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

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

$$D(A,A') = \frac{\sum_{c\in A} \min_{c'\in A'} d_{\text{category}}(c,c') + \sum_{c'\in A'} \min_{c\in A} d_{\text{category}}(c,c')}{|A|.|A'|}, \qquad (18)$$

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 $D(A,A')$ 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 $\beta = 1$. The rate 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 so that it balances the decay rate α of weights in equation (11). α takes care of a slow forgetting of categories, and is set by default to $\alpha = 0.1$.

7. The baseline discriminative success – that is, the chance success that agents would achieve by randomly creating categories – 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 1,000.

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, and 5 R/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/size of context – that is, 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 arguing over the colour of one's shirt, the agents do not always agree on what category a stimulus belongs to.

11. Berlin and Kay (1969) said that there are 11 basic colour categories, but other than that there is no 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 equation (9), in which D is now defined as equation 16. Note that the category variance reported in Figure 19 cannot 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 nonparametric test and does require the data to have a normal distribution. The test returns values between −1 and 1. A value of 1 indicates that the correlation is perfect, and a value of −1 that the correlation perfect but inverse. Values 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 (2001b). It extracts 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

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

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 & Stiles 1982/2000). Each correlation reported in Table 4 is the mean of these five correlations.

Open Peer Commentary

Intimations of optimality: Extensions of simulation testing of color-language hypotheses

David Bimler

Health and Human Development, Massey University, Palmerston North, New Zealand 5331. **d.bimler@massey.ac.nz**

Abstract: By emphasizing that color categories are the collective achievement of a language community, the methodology of Steels & Belpaeme (S&B) suggests a number of corollaries. It focuses attention on whether a system of categories is optimized to match color experience. If a hypothesis can be operationalized about the nature of the optimality – about how color language becomes standardized – it becomes testable.

It was not the intention of Steels and Belpaeme (S&B) to advance one or another of the rival nativist, empiricist, or culturalist positions on color language. Even so, the interest of many readers will center on the prospect of support or disconfirmation from their methodology for specific assumptions about color language (if not from the present report, then from its potential extensions). I would like to make five observations, intended not to reject S&B's results, but to note possible omissions and to expand on insights that follow.

1. In the pure empiricist position as defined by S&B, an observer's color lexicon verbalizes a system of color categories that matches regularities in the distribution of colors in the visual environment. Each observer derives this optimized category system independently. S&B also consider a combination of the empiricist and culturalist approaches in which category systems, derived to match a structured color environment, are compared and harmonized among agents (Fig. 18). The *pure* position is undermined by evidence in section 5 of S&B. First: minor differences in the metric of color dissimilarity (a prerequisite for any algorithm of clustering or category formation) lead to significant departures between clustering solutions (Fig. 20). But individuals certainly do vary in their dissimilarity metrics, that is, in color perception.

This point becomes more salient when the extremes of variation are considered. Observers with forms of color deficiency (dichromacy, monochromacy, and even complete blindness), whose experiences of color perception are grossly aberrant, nevertheless acquire category systems that are organized in ways not so far removed from normal (Marmor 1972; Shepard & Cooper 1992). Presumably they internalize the standard category organization in the course of language acquisition. Clearly, no picture is complete if it omits a component of culturalism. Again, Jameson and Hurvich (1978) showed that dichromats were intellectually aware that "red" and "green" are distinct categories, and indeed diametrically opposite. This awareness was not derived from personal experience, because it did not deter the same dichromats from rating red and green stimuli as subjectively most similar (when sequencing them by similarity), even though lightness cues allowed them to identify the red and the green stimulus within each most-similar pair.

2. Citing these reports also serves to highlight a shortcoming in S&B's "guessing game," compared to natural language. This game couples the separately-derived category systems among agents in simulations. It certainly captures one channel of color conversation, when speakers agree or disagree about their verbalized categories. But natural language also contains a second channel, one conveying information about category *relationships* – whether two categories are adjacent or extreme opposites – in an abstract sense or in the context of actual examples. This embedded or implicit information is part of language acquisition, utilized by the color-deficient observers discussed earlier. It is not immediately obvious how the guessing game might be extended to enable simulations to model this second channel.

3. A second problem with the pure empiricist position is that not all observers experience identical color distributions. The category systems created by clustering the distributions for rural and urban environments are significantly different (Fig. 16). Variations in color distribution among human habitats, and even between seasons for the same habitat, are well-documented (Mizokami et al. 2003; Webster & Mollon 1997). Thus, empiricism alone cannot account for the observed cross-cultural consensus on where the boundaries and the foci of color categories should be drawn.

The World Color Survey (Kay et al. 1997) has confirmed earlier observations that despite wide separation and long histories of divergent development, languages from a range of cultures have converged on similar ways of partitioning the gamut of perceived colors (though specific languages deviate from this consensus, and speakers within a language community are never unanimous).

A pure *nativist* position cannot solve this problem either, though it may conceal it. According to strong nativism, Basic Color Terms (BCTs) consistently appear across many languages because they have been "hard-wired" by evolution into the neural substrate of visual processing, having conferred a selective advantage on our ancestors. This begs the question as to the nature of that selective advantage. One might argue that the BCTs provide an optimal match to the visual environment, but this is simply projecting the empiricist position into the distant past.

The development of color lexicons over historical time is thought to be progressive ("Languages are infrequently or never observed to lose basic color terms": Kay & Maffi 1999, p. 744). Thus, languages verbalizing all 11 BCTs are historically recent (Stage-VII languages, in the terminology of Berlin & Kay 1969). Any advantage bestowed by hard-wired BCTs could hardly be one of improved communication, if they were not all used by early languages.

4. Griffin (2004) used a triad test, similar to S&B's "discrimination game," to quantify the optimality of different category systems. Triads of stimuli were categorized by color, predicting an odd-one-out within each triad, and the accuracy of this prediction was found by comparing it to the actual odd-one-out, known on the basis of object identity. Stimuli in Griffin's study were images of actual objects (using the WWW as an image database), so their color statistics were considerably more structured than a sparse selection of points in color space, or even pixels of real environmental scenes. Intriguingly, the BCTs provided more correct predictions than plausible alternative categories.

5. The structure within color experience that drives multiple populations to converge on a single optimal system of categories might come from the sensory apparatus of the agents. Notably, the color gamut is not a featureless circle, any more than the "color solid" in three dimensions is spherical; there are cusps and protrusions in the Commission Internationale de l'Eclairage spaces (Fig. 20), providing natural anchors for color categories. See also the Interpoint Distance Model (Jameson & D'Andrade 1997). If such ultimately physiological sources of structure can be operationalized for testing with S&B's methodology, combined with a

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

realistic color environment and coupling among agents, it may be that they lead to convergence. It is not clear whether to classify this as a nativist or empirical position.

Implications for memetics

Susan Blackmore

Department of Psychology, University of the West of England, Coldharbour Lane, Bristol BS16 1QY, United Kingdom.
susan.blackmore@blueyonder.co.uk **www.susanblackmore.co.uk**

Abstract: The implications that Steels & Belpaeme's (S&B's) models have for memetics are discussed. The results demonstrate the power of memes (in this case colour words) to influence both concept formation, and the creation of innate concepts. They provide further evidence for the memetic drive hypothesis, with implications for the evolution of the human brain and for group differences in categorisation.

Steels & Belpaeme's (S&B) results are, as they point out, among the first computer simulations to show "how the memetic evolution of language and meaning are possible." They do not explore the further implications of this for memetics, and I propose to do so here.

The basic principle underlying memetics is that memes (including words) are replicators; they can compete with each other and with other replicators in both memetic evolution and meme-gene coevolution. This contrasts with some other theories of cultural evolution in which, as Wilson puts it, the genes will always keep culture on a leash (Lumsden & Wilson 1981). For memetics there is no obvious leash; a replicator can take on either the role of dog or owner under different circumstances. These interactions have previously been modelled (e.g., Bull et al. 2000; Kendal & Laland 2000) and are modelled in new ways by S&B.

The critical experiment for meme-gene coevolution is in section 4.4 where the authors explore the influence of language on the genetic evolution of colour concepts. In their model, not only do word forms compete to describe the colour space, but agents' concepts evolve "genetically." In this key experiment, communicative success of the agents determines fitness, so that agents with the best communicative skills are used to make mutated copies for the next generation.

There are two processes here that are highly relevant to memetics. First (sect. 4.3), when the simulation is run many times the successful memes (colour words) are different each time, which in turn influences the colour concepts that the agents adopt (the Sapir-Whorf thesis). This shows the power of memetic evolution to influence concept formation. Second (sect. 4.4), when communicative success determines fitness, the adopted concepts become genetically assimilated. This is the process that I have previously called memetic drive (Blackmore 1999). It implies that the direction taken by memetic evolution (in this case, the winning words) drives the direction taken by genetic evolution (in this case, innate colour categorisation). In other words, the vagaries of memetic success end up influencing the genetically encoded colour categories.

Although S&B do not mention this, it seems likely that as the simulation proceeds, the mutated agents will increasingly start with a fitness advantage over agents like those who started the simulation, because their innate colour concepts map more closely onto the memetically evolved colour words in use in that population. In other words, outsiders would be at a disadvantage in learning the colour words and so (in this model) would be less likely to become good communicators and survive to the next generation. This would be another reason why, when cultural or memetic factors play a role in fitness, the divergence between populations becomes more pronounced.

If this process occurs in human evolution, there are two significant implications. First, our brains could have been shaped by the results of memetic evolution. That is, the words that happened to evolve in the past (and they might easily have evolved differently) have influenced the ways in which we innately categorise the world. Second, it implies that differences between populations could be greater, or form more quickly, than is assumed in purely genetic models or in models of cultural evolution that do not treat their cultural units as replicators.

Is this plausible? I think so. There is plenty of evidence that, in human mate selection, being articulate, artistic, and creative (S&B's "communicative success") is highly prized. Miller (2000) interprets this in terms of runaway sexual selection, but the models used here demonstrate the memetic alternative. Although it is generally assumed that people from any ethnic background are equally capable of learning any human language (Pinker 1994), there may still be differences to be found if we knew what to look for. The methods used here would allow the relevant variables, such as population size and degree of isolation, to be modelled, and specific predictions made.

S&B have confined their models to colour concepts and words, and to some extent have generalised their findings to all of language. The memetic drive hypothesis can be extended well beyond this to the idea that many aspects of brain design are the way they are because of the history of memetic evolution. For example, the way religious memes evolved in the past (including rituals, or concepts of gods and spirits) could have shaped our peculiarly religious natures (Blackmore 1999; Dawkins 1989), and could thus explain the persistence of religious concepts even in highly educated societies. The way that musical memes happened to evolve could have designed our musical abilities (Dennett 1999), thus explaining a skill that Pinker (1997) describes as being biologically "useless."

More controversially, the process of memetic drive might have implications for understanding group differences in cognitive ability. Indeed, this troublesome issue might usefully be reframed, building on S&B's work, in terms of group differences in innate categorisation. Making plausible assumptions about human population sizes, degree of isolation, and time scale, the methods developed here could be used to model human gene-meme coevolution and find out whether we should expect to see existing human populations that differ in their innate ways of categorising the world because of differences in their past memetic evolution. In these and other ways, S&B's work should prove valuable for testing many memetic hypotheses.

Language, ecological structure, and across-population sharing

Alexa Bódog,[a] Gábor P. Háden,[b] Zoltán Jakab,[c] and Zsolt Palatinus[b]

[a]*Department of General Linguistics, University of Szeged, 6724 Szeged, Hungary;* [b]*Department of Psychology, University of Szeged, 6722 Szeged, Hungary;* [c]*Department of Cognitive Science, Budapest University of Technology and Economics, 1111 Budapest, Hungary.*
alexa.weirdling@gmail.com **robag.nedah@gmail.com**
zjakab@cogsci.bme.hu **palatoki@yahoo.com**

Abstract: We propose a way to achieve across-population sharing within the authors' model in a way that is plausibly in accordance with human evolution, and also a simple way to capture ecological structure. Finally, we briefly reflect on the model's scope and limits in modeling linguistic communication.

We found the authors' approach fascinating and their results very interesting. In this commentary we wish to propose a few amendments that could be implemented using their model. Then we will comment on some interpretations of the results offered in the target article.

Modeling across-population sharing. According to the au-

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

thors' findings, genetic evolution and language can both lead to within-population sharing of color categories, whereas comparable across-population sharing was not obtained in any simulated scenario, not even using structured environments. However, consider the following simulation experiment. (1) Start with languageless genetic evolution (as in sect. 3.3) in a single larger population, and wait until a shared set of categories evolves. (2) Divide the larger population into a few subpopulations, and start a linguistic evolution in each subpopulation. This should go as in section 4.3, but let all agents start this stage with their previously evolved color categories. At this point, the following dilemma arises: (i) When evolution stops, should color categories freeze, that is, not be further modified by word learning, or (ii) alternatively, should learning take over and continue to shape color categories once evolution stopped?

Point (i) of the dilemma will trivially result in between-(sub)population sharing of color categories; it might still be interesting to see how color words evolve here. In point (ii), initial between-subpopulation sharing may fade if, at the initial stage of linguistic evolution, communication failures resulting from uncoordinated individual vocabularies tend to largely alter previously evolved categories, and eventually result in different category systems (as language develops) shared only within populations (Sapir-Whorf effect). Also, marked environmental differences between subpopulations at stage (2) may contribute to the elimination of between-subpopulation sharing.

Note that something like this happened in human evolution. Phylogenesis of trichromacy preceded the occurrence of language.[1] Instead of a dilemma, the authors could vary the extent to which color categories, evolved at stage (1), can be changed by word learning at stage (2).[2] Here is the constraint we would like to see at work, followed by motivating considerations.

As we understand, univeralism about basic color categories implies that linguistic evolution in humans cannot alter some previously evolved (basic) color categories that are grounded in physiology. (However, discrimination and communication efforts can still refine these categories, introducing narrower subcategories within their range.) Our basic color categories are inseparable from the dimensions of color space. The two also evolve together. Trichromat color space is inhomogeneous. It essentially includes a partititoning into color categories, which, in turn, are coupled with its dimensions. For example, unique reds correspond to one side of one chromatic dimension (red-green) and at the same time the midpoint (zero perceptual value) along the other (yellow-blue) one. Bluish greens correspond to perceptions comprising the so-called negative end of both chromatic dimensions, and so on.

In the authors' model, however, dimensions of the CIELAB space (L*, a*, and b*) have no impact on category formation. The categories are unrelated to the dimensions – they arise from an independent system, the adaptive networks. These networks can partition the L*a*b* space in any arbitrary way; they are not constrained by the three dimensions. In addition, in the genetic model, categories evolve, whereas the dimensions of the sensory (L*a*b*) space are the same for every agent and never change.

Incorporating the strong relationship of dimensions and basic color categories in the authors' model seems technically easy. Agents could start with a sensory (e.g., L*a*b*) representation plus a few (4 to 8) basic color categories whose focal points and extents are oriented toward the chromatic axes of the sensory space. For example, locally reactive units of one additive network could have central values in the $a^* > 0$, $b^* = 0$ range, and narrow extents – this would be the category called unique red. Another adaptive net could have local units with central values in the $a^* > 0$, $b^* > 0$ range, and wider extents, corresponding to the category orange, and so on. It could be stipulated that unsuccessful discrimination games do not eliminate these basic categories, but only introduce new narrower subcategories within their range.

Next, how could such a system evolve? In the program, genes could code the spectral sensitivities of the three cone types as well as the linear combination coefficients (nonlinear exponents?) of the cone outputs. As in the opponent processing model, one such combination corresponds to one dimension of color space. The number of combinations could also vary genetically, or be fixed at three. On the other hand, a set of basic color categories should always remain anchored to the dimensions, the number of dimensions determining that of basic color categories. This idea could be included in stage (1) of the simulation suggested previously. Such a simulation, we think, could account for physiology-based between-population sharing, in a way faithful to what we know about human evolution.

Language versus nonlinguistic communication. Although the authors use learned lexicons in their simulation, this is the only similarity between natural languages and the the agents' communication. The use of signs occurs in prelinguistic animals as well (typical examples are call signs: food, dominance, alarm, predator, etc.). Signs can also be learned or invented; plenty of examples of this can be found among birds (e.g., Hauser 1996, p. 274). What the authors' model does show is the feasibility, in principle, of sharing color categories via prelinguistic communication, without dealing with the complexity of human language. Without the abstraction, combinatorial complexity, and other crucial features of human language in the model (Hauser 1996; Hauser & Fitch 2003; Hauser et al. 2002; Hockett 1960; Pinker & Jackendoff 2004), the analogies drawn between category acquisition on the one hand, and linguistic relativism, or meme evolution, on the other, appear a bit farfetched. For example, the phenomenon that the majority forces their category system upon new members (sect. 4.3) is not, in itself, an interesting parallel with memetics. A more important phenomenon of meme evolution occurs when one individual comes up with an attractive idea that, in turn, quickly spreads among others. We cannot see how the authors' model could produce such an effect.

Ecological structure. Although ecological factors are frequently mentioned in the article, they are not examined in the simulation experiments. However, there is a straightforward way to capture ecological structure in the authors' model. Environmental structure was characterized by the frequency distribution of the samples; ecological significance could easily be modeled by rendering different weights of relevance to the samples. (An infrequently occurring stimulus can be highly relevant for the organism, whereas some frequent ones can be less so. Say, the yellow shade of a deadly snake is highly relevant even though infrequent.) Successfully discriminated topics could contribute their relevance ratings to the agent's fitness score. We are curious as to how relevance ratings (in addition to statistical structure) would affect the formation of category repertoires as well as vocabulary acquisition. It would also be interesting to see how the simulation results thus obtained relate to instances of animal categorization and communication.

NOTES

1. The Old World and New World monkeys diverged about 30 million years ago (Mollon 2000, p16). It is believed that the transition from dichromacy to human-like trichromacy occurred at an early stage of catarrhine (Old-World-monkey) evolution (op. cit. p. 20). (Humans are descendants of Old World monkeys.) By any evidence, language is a much more recent achievement. (Simpler or more complex forms of it arose sometime between 3 million and 100,000 years ago.)

2. That is, categories successful in communication could be strengthened according to Eq. 10 (p. 477), whereas unsuccessful ones could fade according to Eq. 11.

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

How to learn a conceptual space

Antonio Chella

Dipartimento di Ingegneria Informatica, Università di Palermo, 90128 Palermo, Italy. **chella@unipa.it** **http://www.csai.unipa.it/chella/**

Abstract: The experiments proposed in the article by Steels & Belpaeme (S&B) can be considered as a starting point toward a general methodology for the automatic learning of conceptual spaces.

In recent years, several frameworks for cognitive robotics have been proposed that take into account a level that is intermediate between the "subsymbolic" low level, directly linked to the external sensors, and the "linguistic" high level, oriented toward symbolic inferences.

A cognitive intermediate level of this kind has been proposed by Gärdenfors (2000). Different from other proposals, Gärdenfors introduces an intermediate level, based on "conceptual spaces," with a precise geometric structure. Briefly, a conceptual space is a metric space whose dimensions are related to the quantities processed by the agent sensors. Examples of dimensions could be color, pitch, volume, and spatial coordinates. Dimensions do not depend on any specific linguistic description: A generic conceptual space comes before any symbolic-propositional characterization of cognitive phenomena.

A point in a conceptual space is the epistemologically primitive perceptive element at the considered level of analysis. Chella et al. (1997; 2000) describe a robot vision system based on conceptual spaces in which each point corresponds to a *geon*-like 3-dimensional geometric primitive (Biederman 1985) perceived by the robot. Therefore, the perceived objects, like the agent itself, other agents, the surrounding obstacles, and so on, are all reconstructed by means of *geons*, and they all correspond to suitable sets of points in the agent's conceptual space. A related conceptual space has been proposed by Edelman (1999), which also proposes an implementation based on Radial Basis Functions (RBF) neural networks. Song and Bruza (2003) adopted a conceptual space framework for information retrieval applications, and Aisbett and Gibbon (2001a) propose a suitable conceptual space for clinical diagnosis applications. From a theoretical point of view, Gärdenfors and Williams (2001) discuss the conceptual space approach for generating nonmonotonic logic inferences, and Chella et al. (2004) discuss conceptual spaces in the framework of the anchoring problem in robotics. Balkenius (1998) proposes a more realistic implementation of a conceptual space, from an empiric point of view, by a set of RBF units, and Aisbett and Gibbon (2001b) discuss a related implementation based on voltage maps.

One of the problems with all of the previously cited approaches is that the structure of the adopted conceptual spaces are a priori defined by the designer according to the addressed problem, in the sense that the designer has to define how many axes are necessary for a correct representation of the problem at hand, what is the meaning of the axes and the corresponding type and range of values, what are the separable and the integral dimensions, and so on. No general methodology has been adopted or proposed to allow the machine to inductively learn a conceptual space, with the exception of the multidimensional scaling algorithm (Shepard 1962a; 1962b) proposed by Gärdenfors, which is generally not suitable for real world robotic applications.

Analyzing the article by Steels and Belpaeme (S&B) from the point of view of the conceptual space theory, the described agents effectively build a conceptual space in order to represent the perceived colors. A "category," implemented by a RBF neural network, identifies a subspace of integral dimensions of colors, because each RBF unit defines a color subdimension, whereas different categories correspond to separable subspaces of colors. Therefore, the color conceptual space of the agent is generated by the union of all the subspaces of integral dimensions of colors corresponding to all the agent categories. The agent inner representation of a color is therefore given by the collection of the responses of all the RBF units built by the agent, that is, by the components of the conceptual space dimensions, in agreement with the conceptual space theory. It should be noted that each color subspace is implemented by a RBF neural network, along the lines of the approaches by Edelman and by Balkenius.

The new and important point brought forth in the S&B experiments is that the agent conceptual space is not defined a priori by the system designer, but it is learned by the agent itself according to its inner and external constraints, as fully described in the target article. Therefore, the strategy adopted by S&B is effectively able to address the previously described problem of how to learn a conceptual space. Interestingly, the conceptual space is generated not only by means of the agent perceptions, but also by the linguistic interactions among agents, that is, by means of the agent actions.

Along this line, it would be interesting to investigate the possibility for an agent to have more powerful representation capabilities that allow the agent to infer the conceptual spaces of other agents, through, for example, a sort of higher-order guessing game. In this way, the problem of sharing categories among populations could be correctly addressed, in the sense that an agent belonging to a population A_x may build an inner representation of the conceptual space of another agent belonging to a population A_y to acquire all the needed capabilities to "translate" its own color categories to the color categories of the other agent.

In conclusion, the S&B article is a seminal starting point for the investigation of a general methodology for inferential learning of conceptual spaces from an agent's external perceptions, its inner and external constraints, and its actions.

Color categories in biological evolution: Broadening the palette

Wayne D. Christensen[a] and Luca Tommasi[b]

[a]*Department of Philosophy, University Kwazulu-Natal, Durban 4041, South Africa;* [b]*Konrad Lorenz Institute for Evolution & Cognition Research, Altenberg 3422, Austria.* **Christensen@ukzn.ac.za**
http://www.kli.ac.at/personal/christensen/homepage.html
luca.tommasi@kli.ac.at
http://www.kli.ac.at/institute-b.html?personal/tommasi

Abstract: The general structure of Steels & Belpaeme's (S&B's) central premise is appealing. Theoretical stances that focus on one type of mechanism miss the fact that multiple mechanisms acting in concert can provide convergent constraints for a more robust capacity than any individual mechanism might achieve acting in isolation. However, highlighting the significance of complex constraint interactions raises the possibility that some of the relevant constraints may have been left out of S&B's own models. Although abstract modeling can help clarify issues, it also runs the risk of oversimplification and misframing. A more subtle implication of the significance of interacting constraints is that it calls for a close relationship between theoretical and empirical research.

Steels & Belpaeme's (S&B's) study attempts to combine research objectives for robotics and human science. But, although using human communication as a model may be a useful starting point for robotics, the radical differences in physical constraints between robots and humans makes it unclear how much overlap there need be between the two areas. The evolution of human communication abilities occurred in a specific biological context, with perceptual, motor, cognitive, social, and ecological constraints that don't apply to robots. Exotic abilities like direct sharing of perceptual information are possible for robots, and ultimately the most effective robot communication systems may be no more similar to human verbal communication than human communication is to that of honeybees or dolphins. This is not to suggest that there will be no important commonalities, but rather to point out that divergent specific constraints can generate very different possibilities.

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

The study is at a sufficiently high level of abstraction that such differences enter the picture only minimally. In taking inspiration from categorization and naming by humans, S&B only focus on the fact that humans are capable of open ended generation of socially shared names, without attempting to model any specific biological or psychological mechanisms that may be involved. Such a strategy offers the potential for generality, but it also faces challenges in demonstrating that the results will be robust against departures from S&B's assumptions, and that the models are informative despite the lack of realism. In their conclusion, S&B suggest that additional realism would only obscure the dynamics, but this very much depends on assuming that unmodeled constraints do not contribute to the dynamics.

Their primary hypothesis is 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." However, this sounds suspiciously like its addressing an ill-formed problem: What counts as "sufficient sharing" and "not enough" may well be sensitive to a variety of factors and vary in different contexts. For example, focal color categorization is present in birds and might have been selected for because it simplifies the cognitive demands of discriminating multiple items in an array, be they landmarks in the environment (Tommasi & Vallortigara 2004), potential mates (Bennett et al. 1997), or the incentive value of different types of food (Gamberale-Stille & Tullberg 2001). Thus, task complexity is one dimension in which departures from S&B's assumptions could have a significant impact: The complex task demands faced by birds may have favored focal color categorization independent of any social referencing constraints. The general problem is that it is very hard to know in advance where and how such interrelations arise, and hence, it can be hard to evaluate whether an abstract model has aptly represented the issues.

Comparative research provides a strategy for disentangling some of these kinds of complexities. In the case of color categorization, birds and mammals are an informative contrast, given the evolutionary radiation that separated these amniote groups from a common ancestor, and the highly sophisticated visual abilities of birds (Güntürkün 2000; Vallortigara 2004). Color perception is more refined in avian species from the level of retinal photoreceptors, because the presence of double cones, oil droplets, and tetrachromacy provide for earlier color-opponency processes than those found in mammals (Vorobyev et al. 1998).

Birds are thus endowed with the structural and functional features necessary to perceive, discriminate, and generalize color stimuli. Selective pressures undoubtedly shaped color spaces in the direction of those sensory aspects that are ecologically relevant to the species. Some behavioral responses (e.g., feeding behavior) are genetically biased in the direction of specific colors (Roper & Marples 1997), but, not surprisingly, the development of chromatic perception is dependent on the statistical structure of the colors experienced in the environment, because rearing newly hatched birds in abnormally colored environments results in alterations of the spectral range to which the birds respond when compared with control animals in color discrimination tests (Miklósi et al. 2002).

As noted, birds have been shown not only to discriminate and generalize colors, but also to categorize the color continuum in discrete regions centerd around focal points, as found in humans (Jones et al. 2001). Even more strikingly, birds have exhibited spontaneous emergence of vocalizations akin to color naming. Manabe et al. (1995) trained budgerigars to emit a high pitch call in case of the presentation of one color and to emit a low pitch call in case of the presentation of another color. Once this association was learned, spontaneous differential vocalizations were observed in response to forms when some new association of forms to colors was being established, as if the birds were anticipating the presentation of a color by its learned vocal label. Research on parrot's chattering has provided evidence of color referencing mediated by vocal communication and apparently equally depending on both parrot–parrot and parrot–human observational learning (Pepperberg & Wilcox 2000). Birds share a basic neural architecture that is substantially different from that of mammals, and yet their evolution independently achieved functions strikingly similar to humans, with the potential for categorization of color in referential communication.

Several points can be drawn from this. Comparing similar traits across diverse phyla is a useful strategy for casting light on evolutionary processes and biological mechanisms, and can help disentangle the general from the specific. However, the differences also make such comparisons fraught; analogies (e.g., describing a particular bird behavior as "naming") must be drawn very carefully. Similar problems of interpretation face theoretical modeling research. If a bird species ever evolves language that involves color naming, do S&B's results imply that the color categories must be socially shaped? We suggest that this is far from clear because it isn't clear how well the assumptions of S&B's models will map to the constraints operative in the particular evolutionary process. Given the difficulty of predicting a priori which constraints, or even which kinds of constraints, may prove relevant in evolutionary and cultural processes, there is reason to try to develop a close coupling between empirical and theoretical research so that the respective strengths and weaknesses can be balanced against each other. In arguing this, we are not seeking to dismiss S&B's work. It is an elegantly structured study that may provide a robust modeling platform for much productive theoretical exploration. But closer empirical links will help its development.

In the beginning: Word or deed?

Stephen J. Cowley

Department of Psychology, University of Hertfordshire, Hatfield, Hertfordshire AI10 9AB, United Kingdom and School of Psychology, University of KwaZulu-Natal, South Africa. **s.j.cowley@herts.ac.uk**

Abstract: Emphasizing that agents gain from culture-based patterns, I consider the etiology of meaning. Since the simulations show that "shared categories" are not based in learning, I challenge Steels & Belpaeme's (S&B's) folk view of language. Instead, I stress that meaning uses indexicals to set off a replicator process. Finally, I suggest that memetic patterns – not words – are the grounding of language.

Using remarkable simulations, Steels & Belpaeme (S&B) show why autonomous robots can gain from sensitivity to culture-based patterns. These can be used to supplement categories grounded in embodiment and, as a result, actions can be better coordinated. The simulations thus illuminate "how the memetic evolution of language and meaning is possible" (sect. 4.3). In this commentary, stressing that agents use indexical signs, I focus on the etiology of meaning. Language itself, I suggest, may depend on how grounded categories interact with memetic patterns or indexical signs.

Although "sharing" develops in genetic simulations, as for color, this may be gross. Equally, as with herring gull chicks, it may depend on "relational signs" that arise in the niche (Tinbergen 1953). Further, the simulations show that shared categories will not arise from individual learning. Mapping a word-form to a color is, for this reason, beyond autonomous devices that lack sensitivity to culture-based patterns. It becomes possible, however, provided that an encultured pattern is consistently coupled with what sensors can detect. Given learning, coupling can give a population grounded relational categories.

S&B draw on a folk view of language. Taking shared categories for granted, they assume that a lexicon maps words onto meanings. Accordingly, they adopt what has been called the "fundamental assumption of linguistics" – the view that, "in every speech community, some utterances are alike in form and meaning" (Bloomfield 1935, p. 78; for critique, see Love 2003). In spite of

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

what simulations show, form–meaning relations are treated as sacrosanct. Challenging this, Chomsky emphasizes that internal language gives thought that is independent of "the sensory field" (2002, p. 75). In parallel, Wittgenstein's later writings show why form–meaning models cannot explain even what is before our eyes. Using careful consideration of utterances like "I know that is a tree" (1969, p. 467), a major goal of *On Certainty*, is to show that, strictly, the claim makes no sense. For Wittgenstein, understanding what we see often rests – not on what is known – but rather a *sui generis* kind of judgment. Uttering "that's a tree" can make sense only to those familiar with how [tri:] ("tree") syllables can be integrated with social activity.

Despite appeal to "shared categories," S&B's results are consistent with philosophical investigation. In color "naming," agents relate patterns to *current* sensory impressions. Given consistent cooccurrence of the inputs, learning captures a relation. As for herring gulls, what matters is the pattern's use. This, moreover, is consistent with finding that "shared categories" are rendered impossible by minor differences in ecology or vision. As in S&B's simulations, relational events enable agents to combine ecological and use-based resources. Shared meaning is thus merely a (valuable) fiction. Further, if use-based patterns modify embodied categories, this fits Dennett's (1995) view that "memes" exploit neither natural nor neural languages (p. 354). Given benefits associated with relational signs, moreover, use-based patterns will be able to spread. In a population, natural or artificial selection will favor games where agents benefit from memetic patterns. Applied to humans, this suggests that the rise of indexical patterns may underpin a "culturally-based replicator process" (Aunger 2000). Further, this conforms to how Wittgenstein sees his garden game. For the philosophers, utterances of "that's a tree" depend – not on shared categories – but on relations. Specifically, they integrate pointing with gazing at objects and, in the same time-scale, uttering ("that is a tree"). The capacity uses neither a lexicon nor shared meanings, but what the *Philosophical Investigations* call a "natural history" (Wittgenstein 1958, p. 415) that engenders human "agreement" (1958, p. 241).

Many animals call to each other and human embryos sensitize to particular voices. It is perhaps to be expected, therefore, that memetic patterns use perceptually grounded discrimination. Not only do many species assess grounded sound-patterns, but, especially in birds and mammals, individuals often use indexicals to manage others (Owings & Morton 1998). Whales can sing and many birds duet; by three months, human infants find what they do shaped by caregiver vocalizations (Cowley et al. 2004). Within this frame, one can generalize S&B's simulations. If consistent use of indexicals prompts cognitive effects, our nervous systems may resemble those of herring-gull chicks in relying on *relational* signs. Whether exploiting physics (e.g,. blue light), affect (e.g., pain) and/or social relations (e.g., dominance), memetic patterns may trigger cognitive change. Mere exposure to use-based dynamics may conceivably influence activity and, as Tomasello (1999) suggests, directing attention may be a basic function of language. Further, indexicals do shape conversational events (Spurrett & Cowley 2004). Even in broad terms, speakers of languages like French, isiZulu, and Thai draw on similar syllable-based rhythms to achieve interpersonal goals. Is this not memetic? Are these perhaps carriers of public emotions that serve, above all, to protect agents from social dilemmas (Ross & Dumouchel 2004)? If so, the simulations show how "strategic signals" can take on semantic and social roles. In humans, the basis of meaning may not only be memetic, but, strikingly, many of its cognitive effects may arise from events *within* cultural and ecological boundaries.

The simulations raise many issues. How do animals get caught up with memetic patterns? How do individual consumers use cognition enhancing resources? As S&B imply, such issues can be investigated through strategic games. In a "trumping" game, for example, credit could be given to agents who anticipated actions or, echoing sexual selection, a "charming" game might credit musical performance. After all, it is striking that no less a figure than Darwin (1871) posited that the evolutionary history of language is to be traced – not to words – but to song. Further, given cultural sensitivity, autonomous robots might simulate how patterns of life codevelop with increasing variability in strategic signaling. Indexicals could be used, at least in principle, by culturally-sensitive machines that coordinated not only inner categories but also interactional events. Given capacities to self-organize, memetic control could thus come to exert an influence on both perception and action. Such agents would *do* what, following Goethe (1959), Faust and Wittgenstein came to believe in. Culturally-sensitive robots, would ground capacities for meaning (and communicating) – not in words – but in patterned deeds.

Language impairment and colour categories

Jules Davidoff[a] and Claudio Luzzatti[b]
[a]*Department of Psychology, Goldsmiths' College, London, United Kingdom;*
[b]*Department of Psychology, University of Milan, Bicocca, Italy.*
pss01jd@gold.ac.uk **claudio.luzzatti@unimib.it**

Abstract: Goldstein (1948) reported multiple cases of failure to categorise colours in patients that he termed amnesic or anomic aphasics. These patients have a particular difficulty in producing perceptual categories in the absence of other aphasic impairments. We hold that neuropsychological evidence supports the view that the task of colour categorisation is logically impossible without labels.

Our commentary contains no fundamental disagreement with the position put forward by Steels & Belpaeme (S&B). We too hold that language, at least in the form of commonly held labels, is a necessary condition for colour categorisation. Furthermore, we hold that as those labels differ between languages, consequently the content of categories will also differ. Our support for S&B comes initially from the cross-lingual evidence quoted in the target article (Davidoff et al. 1999; Roberson et al. 2000). Similar recent data (Roberson et al. 2004) gives further evidence for the Whorfian (Whorf 1956) stance on colour categories; it showed that children who have no colour terms perform colour similarity tasks based solely on perceptual discrimination. Categorical discrimination, as opposed to perceptual discrimination, was evident only after label acquisition. Thus, before label acquisition, both English and Himba (a tribe whose language contains only five colour terms) confuse navy blue with purple – a perceptually close colour. After label acquisition, Himba children confuse navy blue with black; these colours have the same label in Himba. They do not confuse navy blue with royal blue that has a different label in Himba, whereas the English children do make that confusion. Neither group of children continue to confuse navy blue with purple as they have a different label in both languages.

There is, however, an aspect of the S&B argument about which we have comment. The comment derives from the neuropsychological literature, which is a strong source of support to their argument. Goldstein (1948) reported multiple cases of failure to categorise colours in patients that he termed amnesic or anomic aphasics. These patients, as part of their language disturbance, had lost the ability to name or point to named colours. Despite normal colour vision, they found the task of colour categorisation completely bewildering. On being asked to sort colours, they predominantly adopted one of two strategies (see also the case of LEW: Davidoff & Roberson 2004; Roberson et al. 1999; 2002). Either they declared that there were very few (even only one category) or there were many categories (even as many as there were different colours). For example, LEW would choose a colour sample and look round for one that was identical. As none were identical, he would reluctantly place the colour with the one perceptually closest. The procedure was repeated and could even continue until all the samples were in the same group. In LEW, at least, the extraordinary behaviour cannot be attributed to a failure

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

in all categorisation tasks, as he very promptly divided pictures of animals into British versus foreign (Roberson et al. 1999). Nor, in general, could it be attributed to the patients' poor language skills. For example, the patient AV investigated by Luzzatti and Davidoff (1994) showed only minimal aphasic disturbance, but nevertheless could not sort colours.

The performance of these anomic aphasics lends support to the argument that agreed upon labels are essential for the production of colour categories. However, it might ask for a reconsideration of the agent's performance in its individualistic (empiricist) simulation. S&B show that agents who cannot communicate with each other arrive at solutions sufficiently different that they could not form the basis of a common colour categorisation. Nevertheless, if we inspect Figure 3, we see that each agent arrives at reliable categories, even ones that at first glance look somewhat similar. The somewhat similar categories are produced because S&B give the agents a categorisation device based on a prototype. If such a device were available to the patients, one might have predicted, at least, idiosynchratic categories after their language loss. If the patients had, premorbidly, the usual category prototypes, one might even predict normal categorisation. Yet, the patients are unable to produce proper categories. In our view, allowing the agents a categorisation device considerably underestimates the importance of the labels.

In our view, the task of colour categorisation is logically impossible without labels (Dummett 1975; Roberson et al. 1999). The agent in the natural world is confronted with coloured sensations that vary on a perceptual continuum. Importantly, the continuum has no natural discontinuities (or prototypes) that would allow obvious division into categories. The agent is confronted with displays that conform to what is known as the Sorites paradox. Imagine the situation where the stimuli in the discrimination game are not widespread, as in their example, but vary by so little that the perceptual discrimination mechanism cannot tell them apart. Let's say that the samples are light sources of 620 nm, 619.9 nm, 619.8 nm, and 619.7 nm. As they cannot be told apart, they can all be given the same label – let's call it red. If they were then replaced by four further colours, each decreasing by 0.1 nm, none of these could be distinguished from the 619.7 nm sample. So, the agent could call all the new four red, as well. The procedure could be repeated over and over again until paradoxically colours at the other (blue) end of the spectrum would also be called red. This nonsensical situation (Sorites paradox) is resolved by making a nonperceptual (e.g., verbal label) discontinuity in the continuum (Dummett 1975). As we see in the behaviour of the anomic aphasics, they behave as if the Sorites paradox is more than a philosophical speculation. So, there is a reason to doubt whether any agent relying solely on its own feedback in the real world would produce categorical performance of any reliability.

The neuropsychological evidence adds support to the argument for the importance of labels in category formation. Only if the world was constrained to present examples that were sufficiently far apart perceptually to allow discontinuities in colour space, could the agent form any useful categories. Those discontinuities would not arise from the changes in perceptual input that denote differences between urban and rural environments. And, as for the discontinuities arriving genetically, we still wait for any convincing evidence that the human visual cortex contains neurons that selectively act as detectors for red, green, blue, or any other colour.

Realistic constraints on brain color perception and category learning

Stephen Grossberg

Department of Cognitive and Neural Systems, Boston University, Boston, MA 02215. **steve@bu.edu** **http://www.cns.bu.edu/Profiles/Grossberg**

Abstract: Steels & Belpaeme (S&B) ask how autonomous agents can derive perceptually grounded categories for successful communication, using color categorization as an example. Their comparison of nativism, empiricism, and culturalism, although interesting, does not include key biological and technological constraints for seeing color or learning color categories in realistic environments. Other neural models have successfully included these constraints.

Steels & Belpaeme (S&B) approach the symbol-grounding problem using learning algorithms that "take as much inspiration as possible from categorisation and naming by humans" (sect. 1, para. 3). They nicely summarize three traditional positions about this problem, but then test these positions using models that do not simulate biological data and are tested on toy problems that do not demonstrate superior performance against the best models available. Talking biology without data and talking technology without comparative benchmarks does not satisfy the demands of either biology or technology.

At least two processes subserve color categorization: color preprocessing and recognition learning to categorize preprocessed data. The authors used a small stimulus set: "four stimuli chosen from a total of 1,269 Munsell chips" (sect. 3), measured their reflected spectral energy, and claimed that they "start from realistic colour data" (sect. 2.2) and use a "reasonable model of human lightness perception" (sect. 2.3.1). However, perceived colors cannot be derived from local measurements. Lighting conditions and surface context dramatically influence color perception. Preprocessing requires discounting the illuminant, perceptual grouping, surface filling-in and anchoring processes that have been simulated by other neural models, including lightness data (e.g., Grossberg & Todorovic 1988; Hong & Grossberg 2004; Mingolla et al. 1999; Pessoa et al. 1995). The authors' studies of color category learning and naming used inputs that do not represent challenges that autonomous robots would meet in the real world.

The authors' neural categorization models are also inadequate. A realistic model must realize stable incremental learning of both small and large categories in response to changing environmental statistics. Concerning stability: Its dot products (equation 2) and winner-take-all choices (equation 3) are well-known in competitive learning and self-organizing map models (e.g., Grossberg 1976a; Kohonen 1984; Rumelhart & Zipser 1986). These models suffer from a problem common to most learning models. I called the problem the stability-plasticity dilemma; some others call it catastrophic forgetting (Carpenter 2001; Grossberg 1980; Page 2000); namely, the memory of learned categories can be rapidly erased in response to changing environmental statistics. Thus, the authors' categorization models cannot work well in a realistic incremental learning setting. It has been mathematically proven that catastrophic forgetting can be overcome by an appropriate top-down matching and attentional focusing mechanism (Carpenter & Grossberg 1987; 1991) that has been embodied in Adaptive Resonance Theory, or ART, algorithms (Grossberg 1999) and supported recently by many behavioral and neurobiological experiments, reviewed in Raizada and Grossberg (2003). Indeed, ART was introduced as an attempt to solve the catastrophic forgetting problem that was identified in competitive learning and self-organizing map models (Grossberg 1976b).

Concerning adaptive control of category size: Unsupervised clustering is sensitive only to the statistics of the input environment. Supervision enables categories of variable size to be learned that are also sensitive to cultural and language constraints. The claimed ability of Eskimos to perceive and name unusually subtle shades of blue may combine both properties. Realistic models

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

must automatically switch between incremental unsupervised clustering, when predictive feedback is unavailable, and supervised clustering, when predictive feedback is available, without a loss of memory stability, to learn categories that are sensitive to both environmental statistics and cultural constraints. ARTMAP algorithms achieved this goal by combining top-down matching and attention with vigilance or sensitivity control that together maximize generalization while minimizing predictive error using a patented concept called match tracking (Carpenter & Grossberg 1991). In this regard, S&B discuss distinguishing between edible and nonedible mushrooms, notably the need to distinguish "fine shades of orange" (sect. 2.2) in mushroom databases. However, they do not simulate classical mushroom databases (Schlimmer 1987). ARTMAP has been benchmarked on the mushroom database with a 99.8% accuracy during on-line learning (Carpenter et al. 1991).

S&B also note the importance of studying "categorization and naming by humans" (sect. 1), but do not do model human performance. ARTMAP has simulated the set of thirty human categorization experiments, called the 5–4 category structure (Smith & Minda 2000), which is a standard benchmark for human categorization (Ersoy et al. 2002). Whereas traditional cognitive models can fit these data, they do so without learning the categories and without describing underlying brain dynamics. ARTMAP learns the categories and fits the data at least as well as cognitive models, and also proposes how to settle the classical exemplar/prototype debate concerning whether exemplars or prototypes are stored in memory. ARTMAP predicts that critical feature patterns to which humans learn to pay attention are stored in memory. Under language/cultural supervision, these prototypes can be either specific ("exemplars"; Estes 1994; Medin & Smith, 1981; Medin et al. 1983) or general ("prototypes"; Posner & Keele 1970; Smith & Minda 1998; 2000; Smith et al. 1997). Typically, both specific and general information will be learned ("rule-plus-exceptions"; Nosofsky 1984; 1987; Nosofsky et al. 1992; Palmeri et al. 1994).

ARTMAP is a standard tool for learning complex categorical relationships from high-dimensional input vectors that include color among other visual features, while autonomously discovering hierarchical knowledge relationships among the categories (Carpenter et al. 2004a; 2004b; Parsons & Carpenter 2003).

S&B summarize familiar features of neural models of supervised learning using nomenclature about games. Although these games sound novel, they actually embody well-known neural modeling concepts, including memory search or hypothesis testing to create new categories, the use of predictive success to culturally constrain learned naming, and the need to control category size. All of these properties are unified and proceed automatically in ARTMAP algorithms. It remains for S&B to demonstrate, through comparative benchmarks, that their models can cope with the categorical challenges that this alternative approach has already handled.

Modeling category coordination: Comments and complications

James A. Hampton

Psychology Department, City University, Northampton Square, London EC1V OHB, United Kingdom. **hampton@city.ac.uk**
www.staff.city.ac.uk/hampton

Abstract: Consideration of color alone can give a misleading impression of the three approaches to category coordination: the nativist, empiricist and culturalist models. Empiricist models can benefit from a wider range of correlational information in the environment. Also, all three approaches may explain a set of perceptual categories within the human repertoire. Finally, a suggestion is offered for supplementing the naming game by varying the social status of agents.

The broad conclusion drawn by Steels & Belpaeme (S&B) on the basis of their explorations of three general models for category name coordination is that, whereas the genetic and the cultural/language-based models can lead to coordinated categorization and naming practices within populations, the statistical structure available for colors in the immediate environment is insufficient to allow the empiricist model to achieve the same level of performance.

The latter claim is critically dependent on the choice of environmental structure provided to the empiricist model. S&B explored only one source of structure – a random sampling of pixels taken from photos of environmental scenes. Before drawing any firm conclusion about the possibility of achieving full coordination of categories simply from statistical covariation in the world, a more realistic characterization of the environment is surely necessary. In particular, colors are not seen by individuals as independent pixels, but as reflectances of the surfaces of objects and parts of the visual scene, which can be tracked through space and time as the individual moves through the scene. Other visual properties such as shape and size of the color patch, where it is located relative to objects in the scene, and sensory properties from other modalities all provide rich sources of correlational structure which establish a categorization of the world. Coordination of color categories may then benefit from the association of colors with object classes (oranges are typically orange, lemons typically yellow, the sky typically blue, blood red, and so forth). To take one of their examples, if coordination of color categories is important for the detection of poisonous mushrooms, then it is unrealistic to suppose that the morphology, size, smell, and habitat of the mushroom will not also play a role in categorization – and hence provide crucial evidence about where to draw the color category boundary in this instance. It is therefore an empirical question whether a richer modeling of the statistical structure in the environment would be sufficient to allow a purely empiricist model to develop coordinated categories as efficiently as the other two model types.

Presenting the three approaches to the problem as mutually contrasting accounts may also be misleading. Human categorizers (and human cultures that develop category systems) form and name categories on the basis of a wide range of sources of information. It is easy to find prima facie candidates for perceptual categories that are grounded in each of the three models explored – genetic, empirical, or cultural. Coordination of names for basic emotions such as happiness and grief, or bodily states such as hunger, thirst, or fatigue is presumably based on our common genetics. Coordination of names for biological classes probably relies on the fact that the similarity structure of biological classes at an intermediate level (e.g., elephant, tiger) contains clearly defined clusters with high within-cluster similarity and low between-cluster similarity, giving relatively universal taxonomic systems across different cultures at this level (Lopez et al. 1997). Artifact classification at the basic level may similarly rely solely on high levels of distinctiveness (Rosch et al. 1976). Other categories that depend more on language may be found in culturally-specific categories relating to social practices. For example, classification of ceramics, painting, or music in terms of different artistic styles, or notions of good and bad taste in clothing or decoration are perceptually grounded, but may depend heavily on language for their coordination. It is only the fact of having the concept in the language that leads the language learner to attend to the relevant perceptual cues and construct the necessary prototype representations. A wider view of perceptual categories suggests therefore that the three approaches considered by S&B – nativism, empiricism, and culturalism – all have their place in explaining the rich repertoire of human concepts.

My final comment relates to the cultural model itself. S&B's model assumes a fully cooperative pair of individuals in the language game. Each is willing to adapt his/her categorization and usage of language in the service of improving communication. In actual human societies, the degree of cooperation may be less

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

evident. The right to determine how things are categorized and named is not so evenly distributed. Children are expected to conform to adult norms. In Western cultures, social classes whose education has given them greater access to the elaborated use of language (Bernstein 1981) may determine that there are right and wrong ways in which to classify and name the world. For an increasing number of domains, the correct use of a word is the province of an expert – Putnam's Division of Linguistic Labor (Putnam 1975) – to whom other language users are inclined (and expected) to defer (Kalish 1995). It would be an interesting exercise for S&B to consider introducing differing levels of social status in the naming game. Agents of lower status would be willing to adapt their representations rapidly, whereas agents of higher status would hold on to their beliefs longer in the face of disagreement, particularly if the interlocutor was of lower status than themselves. It would be fascinating to know if introducing this dimension into the game leads to quicker coordination of categories throughout the community, and whether the higher or lower status players end up showing greater or less variance as a subclass of agents.

Language and the game of life

Stevan Harnad

Centre de Neuroscience de la Cognition, Université du Québec à Montréal, Montréal, Québec H3C 3P8, Canada. **harnad@uqam.ca**
http://www.ecs.soton.ac.uk/~harnad/

Abstract: Steels & Belpaeme's (S&B's) simulations contain all the right components, but they are put together wrongly. Color categories are unrepresentative of categories in general and language is not merely naming. Language evolved because it provided a powerful new way to acquire categories (through instruction, rather than just the old way of other species, through trial-and-error experience). It did not evolve so that multiple agents looking at the same objects could let one another know which of the objects they had in mind, co-coining names for them on the fly.

Contra Wittgenstein (1953), language is not a game. (Maynard-Smith [1982] would no doubt plead *nolo contendere*.) The game is *life*, and language evolved (and continues to perform) in life's service – although it has since gained a certain measure of autonomy too.

So are Steels & Belpaeme (S&B) inquiring into the functional role for which language evolved, the supplementary roles for which it has since been coopted, or merely the role something possibly resembling language might play in robotics (another supplement to our lives)?

For if S&B are studying the functional role for which language evolved, that role is almost certainly absent from the experimental conditions that they are simulating. Their computer simulations do not capture the ecological conditions under which, and for which, language began. The tasks and environments set for S&B's simulated creatures were not those that faced any human or prehuman ancestor, nor would they have led to the evolution of language had they been. On the contrary, the tasks faced by our prelinguistic ancestors (as well as our nonlinguistic contemporaries) are precisely the ones *left out* of S&B's simulations.

S&B do make two fleeting references to a world in which foragers need to learn to recognize and sort mushrooms by *kind* – with color possibly serving as one of the features on the basis of which they sort. But a task like learning to sort mushrooms by kind is not what S&B simulate here. They simulate the task of sorting colors, and not by *kind*, but by a kind of "odd man out" exercise called the "discrimination game." In this game, the agent sees a number of different colors (the "context"), of which one (the "topic") is the one that must be discriminated from the rest. If this is done by two agents, it is called the "guessing game," with the speaker both discriminating and naming the topic-color, and the hearer having to guess which of the visible context-colors the speaker named. Both agents see all the context-colors.

Now the first thing we must ask is: (i) Were any of our prelinguistic ancestors ever faced with a task anything like this? (ii) And if they had been, would that have led to the evolution of language? (iii) Indeed, is what is going on in S&B's task *language* at all?

I would like to suggest that the answer to all three questions is no. S&B's is not an ecologically valid task; it is not a canonical problem that our prelinguistic ancestors encountered, for which language evolved as the solution. And even if we trained contemporary animals to do something like it (as some comparative psychologists have done, e.g., Leavens et al. 1996), it would not be a linguistic task – indeed it would hardly even be a categorization task, but more like a joint multiple-choice task requiring some "mind-reading" (Premack & Woodruff 1978; Tomasello 1999) plus some coordination (Fussell & Krauss 1992; Markman & Makin 1998).

On the other hand, there is no doubt that our own ancestors, once language had evolved, *did* face tasks like this, and that language helped them perform such tasks. But language helps us perform many tasks (even learning to ride a bicycle or to do synchronized swimming) for which language is not necessary, for which it did not evolve, and which are not themselves linguistic tasks. (This is S&B's "chicken/egg" problem, but in a slightly different key.)

Let's now turn to something that *is* ecologically valid. Our prelinguistic ancestors (and their nonlinguistic contemporaries, as well as our own) did face the problem of categorization and category learning. They did have to know or learn what to do with different *kinds* of things, in order to survive and reproduce: what to eat or not eat, what to approach or avoid, what kind of thing to do with what kind of thing, and so forth. But categorizing is not the same as discriminating (Harnad 1987). We discriminate things that are present simultaneously, or in close succession; hence, discrimination is a *relative* judgment, not an *absolute* one. You don't have to *identify* what the things are in order to be able to discern whether two things are the same thing or different things, or whether this thing is more like that thing or that thing. Categorization, in contrast, calls for an absolute judgment of a thing in isolation: "What kind of thing is this?" And the identification need not be a name; it can simply be *doing* the kind of thing that you need to do with that kind of thing (flee from it, mate with it, or gather and save it for a rainy day).

So categorization tasks have not only ecological validity, but cognitive universality (Harnad 2004). None of our fancier cognitive capacities would be possible if we could not categorize. In particular, if we could not categorize, we could not name. To be able to identify a thing correctly, in isolation, with its name, I need to be able to discriminate it absolutely, not just relatively – that is, not just from the alternatives that happen to be copresent with it at the time (S&B's "context"), but from all other things I encounter, past, present, and (one hopes) future, with which it could be *confused*. (Categorization is not necessarily exact and infallible. I may be able to name things correctly based on what I have sampled to date, but tomorrow I may encounter an example that I not only cannot categorize correctly, but that shows that all my categorization to date has been merely approximate too.)

Notice that I said categorize *correctly*. That is the other element missing from S&B's analyses. For S&B, there are three ways in which things can be categorized: (N) innately ("nativism"), (E) experientially ("empiricism"), and (C) culturally ("culturalism," although one wonders why S&B consider cultural effects nonempirical!). To be fair, the way S&B put it is that these are the three ways in which categories can come to be *shared* – but clearly one must *have* categories before one can share them (the chicken/egg problem again!).

Where do the S&B agents' color categories come from? They seem to think that categories come from the "statistical structure" of the things in the world, such as how much things resemble one another physically, how frequently they occur and cooccur, and how this is reflected in their effects on our sensorimotor transducers. This is the gist of S&B's factor E, empiricism. Where the statistical structure has been picked up by evolution (another em-

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

pirical process) rather than experience, this is factor N, nativism. But then what are we to make of factor C, culturalism? I think that what S&B really have in mind here is what others have called "constructionism." With factors N and E, categories are derived from the structure of the world; with factor C they are somehow "constructed" by cultural practices and conventions. It is in this light that S&B introduce the "Whorf Hypothesis" (Whorf 1956), according to which our view of reality depends on our language and culture. But the Whorf Hypothesis fell on especially hard times with color categories, and S&B unfortunately inherit those hardships in using colors as their mainstay.

There are many ways in which color categories are unrepresentative of categories in general. First, they are of low dimensionality (mainly electromagnetic wave frequency, but also intensity and saturation). Second, they have a known and heavy innate component. We are born with sensory equipment that prepares us to sort (and name) colors the way we do with incomparably higher probability than the way we sort the categories named by most of the other nouns and adjectives in our (respective) dictionaries. Nor are most of the categories named by the words in our dictionaries variants on prototypes in a continuum, as colors are.

Yes, there are variations in color vision, color experience, and color naming that can modulate color categories a little; but let's admit it: not much! Moreover, color categories are hardly decomposable. With the possible exception of chromatographers, most of us cannot replace a color's name with a description – unlike with most other categories, where descriptions work so well that we usually don't even bother to lexicalize the category with a category-name and dictionary-entry at all. Even "the color of the sea" is only a one-step description, parasitic on the fact that you know the sea's color. Compare that with all the different descriptions that you could substitute for "chair."

Why does describability matter? Because it gets much closer to what language really is, and what it is really for (Cangelosi & Harnad 2001). Language is not just a category taxonomy. We use words (category names) in combination to *describe* other categories, and to *define* other words, which makes it possible to acquire categories via *instruction* rather than merely the old, prelinguistic way, via direct experience or imitation. S&B think naming's main use is to tell you which object I have in mind, out of many we are both looking at now. (It seems that good old pointing would have been enough to solve that problem, if that had really been what language was about and for.)

But not only are color categories unrepresentative of categories in general, and the joint discrimination game unrepresentative of what language evolved and is used for, but categories do not derive merely or primarily from the passive correlational structure of objects (whether picked up via species evolution or via individual experience). It is not the object/object or input/input correlations that matter, but the *effects* of what we *do* with objects: the input/output correlations, and especially the *corrective feedback arising from their consequences*. What S&B's model misses, focusing as it does on discrimination and guessing games instead of the game of life, is that categories are acquired through feedback from *mis*categorization. We have this in a realistic mushroom foraging paradigm, but not in a hypothetical discrimination/guessing game (except if we gerrymander the game so that successful discriminating/guessing becomes the name of the game by fiat, and then that is fed back in the form of error-correcting consequences).

Yet all the right elements do seem to be there in S&B's simulations. They are simply not put together in a realistic and instructive way. The task of mind-reading in context seems premature. Every categorization in fact has *two* contexts. First, there is its *context of acquisition*, in which the category is first learned (whether evolutionarily via N or experientially via E) by trial-and-error, with corrective feedback provided by the consequences of miscategorization. The acquisition context is the series of examples of category members and nonmembers that is sampled during the learning (the "training set" in machine learning terms). Until language evolves, categories can only be learned and marked on the basis of an *instrumental* "category-name" (approaching, avoiding, manipulating, eating, mating). With language, there is the new option of marking the category with an arbitrary name, picked by (cultural) convention.

When a category has already been learned instrumentally, adding an arbitrary name is a relatively trivial further step (and nonlinguistic animals can do it too). But then comes the second sense of "context": the *context of application* (for an already acquired category) in which the learned arbitrary category-names are used for other purposes. S&B's paradigm is, in fact, just one example of the context of application (telling you which of the colors that we are both looking at I happen to have in mind), but not a very representative or instructive one. Far more informative (literally!) is a task in which it is *descriptions* that resolve the uncertainty, and the alternatives are not even present. This is not discrimination but instruction/explanation. But for that you first need real language, and not just a taxonomy of arbitrary names (Harnad 2000).

What follows from this is that a "language game" in which words and categories are jointly coined and coordinated "on the fly," as in S&B's color-naming simulations, is not a realistic model for anything that biological agents ever do or did. There is still scope for Whorfian effects, but those will come from the fact that both our respective experiential "training samples" (for all categories) and our corrective feedback (for categories about which our culture and language have a say in what's what, and hence also a hand in dictating the consequences of miscategorizing) have degrees of freedom that are not entirely fixed either by our inheritance or by the structure of the external world.

Categories are *underdetermined*, hence so are the features we use to pick them out. In machine learning theory, this is called the "credit/blame" assignment problem ("which of the many features available is responsible for my successful or unsuccessful categorization?"), which is in turn a symptom of the "frame problem" (how to anticipate all potential future contingencies from a finite training sample?) and, ultimately, the "symbol-grounding problem" (how to connect a category-name with all the things in that category, past, present, and future?) Underdetermination leaves plenty of room for Whorfian differences between agents.

A synthesis of many levels of constraints as a modern view of development

Derek Harter and Shulan Lu

Department of Psychology, Computer Science and Information Systems, Texas A&M University, Commerce, TX 75429.
Derek_Harter@tamu-commerce.edu
http://faculty.tamu-commerce.edu/dharter/
Shulan_Lu@tamu-commerce.edu

Abstract: The debate of nativisim versus empiricism is over the relative importance of evolutionary versus ontogenetic mechanisms. This is mostly seen today as a false dichotomy. The synthesis of these positions provides a modern viewpoint of grounded category formation. This combined view places equal importance on feedback between these levels in guiding development, and is more appropriately compared to culturalist positions.

Much of the debate between nativism and empiricism seems to us to echo similar debates that have been prevalent in developmental psychology and biology on the question of nature versus nurture. That is to say, how much of a role does genetic evolution play in the development of behavior in humans and animals? How much can be attributed to ontogenetic learning by the individual? Is either factor predominant and, if not, are there some areas of behavior and learning where one or the other is the main contributing factor? These debates seem, however, to have reached somewhat of a preliminary consensus, that it is neither and both

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

at the same time (see, e.g., Oyama 1985), the basic idea being that both complex adaptive systems (evolution and ontogenetic development) are at once separate, but also tightly coupled with each other in mutual feedback relations. The complex and constant feedback between these levels are what define and shape the successful behavioral strategies (from micro-cellular to societal) that such systems constantly seek out in order to survive and reproduce in their environments (Thelen & Smith 1994; Thelen et al. 2001).

So, from this we believe it may be a bit of a straw-man to compare a culturalist position to simple nativist and empiricist positions as separate from one another. A more modern viewpoint (Lewontin et al. 1984; Oyama 1985) would need to view nativism and empiricism in a synthesized manner and conclude that the complex mutual feedback between the evolutionary and ontogenetic processes is what coordinates the development of categories. Therefore, the culturalist position is mainly innovative in that it posits a new complex adaptive system, that of language and culture use, as a third factor that plays a role in the feedback among levels to coordinate the development of categories.

The authors describe the culturalist position as "viewing language . . . as a complex adaptive system that is constantly coordinated by its users" (sect. 1, para. 8). However, they go on to indicate that they believe that a consensus needs to be reached on which approach, nativist, empiricist, or culturalist, is most appropriate in explaining the grounded development of categories, and therefore most useful for an engineer in developing a mechanism to implement robust category development in an artificial system. Even if language use and social interaction are shown to be another type of system that plays an important role in the development of categories, does this really mean that genetic evolution and/or individualistic learning would be shown to play lesser roles? No, we believe, and possibly the authors would agree. All three are involved, and understanding the development of categories necessitates understanding all three systems, as well as how they interact with and feedback on one another. As the authors say, the question is really one of levels of freedom, and which levels of adaptive systems are most involved in constraining which levels of freedom.

The authors maintain neutrality on the question of nativism, empiricism, and culturalism as to which, if any, theory best explains observations and data on human performance. However, they do state a position, saying that multiple sources of constraints are present in the formation of shared categories. They list three constraints: those coming from embodiment, those from the environment, and those from culture; and they generally identify nativists as emphasizing the first category, empiricists the second, and culturalists as throwing cultural constraints into the mix. It would seem that the true position they favor, and one we would very much agree with, is that all of these sources of constraints play important roles. Emphasizing one over the others always misses an important point, that it is the interaction between these constraints at different levels that is the key component of development. This article provides important results that will help us to tease apart the contributions of these various influences, using simplified models of developmental processes. However, we feel that the authors don't go far enough in pushing a synthesized view. Some people may still be stuck in a viewpoint maintaining the primacy of one type of constraint in the developmental process, but at least in terms of genetic and environmental constraints, it is clear to many that both play important roles, interact with each other, and the interactions between the mechanisms must be studied, as well as the mechanisms themselves.

Language and social interaction, as complex adaptive systems, would seem to occupy an intermediate level, in terms of time scale, between the relatively slow processes of evolutionary development, and the quick processes of ontogenetic learning. Therefore might they represent a kind of bridging level between the long-range and short-range processes? What level of social interaction is necessary so that a population will develop a shared set of grounded categories? The experiments in this article are a type of communication, but a very simple one at that. Is it really necessary to have a human-like language, or is some much more simple type of social interaction capable of developing shared categories? For example, is simply the fact of animals being social, where they have to act together and coordinate behavior, enough to provide some type of simple semiotic symbols that would allow for the development of coordinated categories? If any type of social interaction is capable of producing shared categories, does a more full-blown human language accomplish something even more in constraining levels of freedom? What extra mileage might a human-like natural language add to the development of shared categories?

It is not evolution, but a better game would need a better agent

Christian Huyck and Ian Mitchell

Middlesex University, Hendon, London NW4 4BT, United Kingdom.
c.huyck@mdx.ac.uk **http://www.cwa.mdx.ac.uk/chris**
i.Mitchell@mdx.ac.uk **http://www.cs.mdx.ac.uk/staffpages/ianm**

Abstract: Steels & Belpaeme (S&B) refer to the neural plausibility and evolutionary plausibility of their algorithms. Although this is not central to their goal of effective artificial agents, their algorithms are not neurally or evolutionarily plausible. Their communication games are interesting, and more complex games would lead to more effective agents. However, the algorithms could be improved either by using standard subsymbolic algorithms or by algorithms that are really neurally or evolutionarily plausible.

We accept Steels & Belpaeme's (S&B) main point that communication can increase the overlap between conceptual representations of both human and artificial agents. This said, we find several faults with it, including their inconsistent use of arguments, and their poor usage of evolutionary algorithms. We also find two related areas that should be addressed: hierarchical categories and more complex games.

S&B play fast and loose with their overarching methodology. They are inspired by the main approaches to human categorisation, but are not constrained by these approaches. This is fine when they are making points about artificial agents, but they frequently make references to evolutionary, environmental, and neural arguments.

For example, they state that their evolution simulations are too slow: "the agents need . . . at least 400 years" (sect. 3.3). This is a legitimate argument against people learning categories by evolution, but it is probably not appropriate for other types of biological systems. For example, fruit flies could learn the category in 20 days. A better argument against humans learning categories by evolution is that a person born to one language group, but raised in a second, learns the second. However, this argument is entirely irrelevant to their main point about artificial agents.

They also describe the use of environmental stimuli (sect. 2.4.1). S&B go into depth about the environment and mushrooms, but, in fact, their stimuli are just a set of 3-tuples. Their simulations have very little to say about the environment.

S&B make frequent use of the word *language* (e.g., sect. 4). However, their simulations only use labels. This is clearly a different thing from what is typically referred to as language, which includes syntax, grammar, semantics, and pragmatics. At best, their simulations are dealing with a symbol-grounding problem.

We are sceptical of the biological plausibility of radial basis function networks, though individual neurons do seem to map reasonably well to neurons. Moreover, the system they model has a network topology and learning algorithm that does not seem biologically plausible.

We find real problems with the genetic evolution simulations. S&B say that a generation is formed by retaining the best half of the previous generation, and a single mutated copy of each. This

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

is a very high mutation rate, and the mutation that they use is very different from standard mutation. The standard genetic algorithm (GA) mechanism involves crossover, and surprisingly, they have none in their simulations. It is not clear why S&B do not use crossover, but it may be because they get saturation in the population. One way to fix this problem would be to use structured GAs (Dasgupta & McGregor 1992). Finally, their population sizes of ten are very small. We find it surprising that this system does as well as reported, because it does not use crossover to explore the search space; mutation, at best, gives simulated annealing, and the number of agents is small. Undoubtedly, a more traditional system would do better. It is a bit of stretch to say that standard GAs are an accurate representation of evolution, and a much greater stretch to say that S&B's system reflects evolution.

These problems are minor, and do not significantly affect the main points of the article. The main point is to develop agents that have collective shared categories. These categories could be improved by developing more complex environments, and by developing hierarchical categories.

The two games that are explained are very simple. That is, the agents that are being developed are developed to function in very simple environments. One obvious way to improve the agents is to put them in more complicated environments. That is, they could develop in more complex games, and the inputs could be more sophisticated.

The games are simple for several reasons. First, the objects are simple, being made up of 3-tuples. Moreover they are discrete in that they do not overlap. This could easily be changed by playing the game on standard categorisation tasks instead of colors (Cairns et al. 2001). Human agents could even move the categories so that the system shared the vocabulary with the agents. Second, the games are very simple. The discrimination game is entirely about categorisation and nothing else. The guessing, as S&B say, is "the most basic language game one can imagine." A range of more difficult games and environments can be easily imagined, including, for example, different objects, variable pay out, delayed response, and multiple players. These could all significantly affect the type of agent that would be best.

On simple expansion, what we are interested in is hierarchical categories. One of the works that S&B refer to is Rosch (1978) and one of the major points of her work is that people store hierarchical categories (Rosch & Mervis 1975). The best way to modify S&B's games to account for hierarchical categories is not entirely clear. A simple way to integrate hierarchy into the problem is to modify the discrimination game. If the target category is very different from the other members of the presentation set, then a high-level category would be used. If the presentation set is similar, then a subcategory would be used. This would be one way to expand the capability of artificial agents.

A host of other game modifications could be proposed: negation, short-term memory, sequences, payoff, deception, and even full-fledged language could be added. It is not clear how well S&B's agents would respond to these modifications.

As they are not dependent on biologically plausible mechanisms, they could use standard machine learning algorithms and normal programming techniques to enable their agents to have the ability to manage these new game environments. However, a better way might be to use nets that are more biologically plausible. This would enable S&B to develop agents that are more human-like and could use human-like solutions.

Dynamical categories and language

Takashi Ikegami

Department of General Systems Sciences, University of Tokyo, Komaba, Tokyo 153–8902, Japan. **ikeg@sacral.c.u-tokyo.ac.jp**
http://sacral.c.u-tokyo.ac.jp/~ikeg

Abstract: The dynamical category uses the sensory-motor coordination to do categorization. If categories are inevitably grounded in sensory-motor coordination, sharing categories may also share the same sensory-motor coordination. Concerning this aspect, we discuss the color category as a dynamical categorization. Additional to the converging effect of a category by communication, we discuss the diverging effect of communication that creates new categories.

A category (of the perceptually grounded type) can be generated easily, but is difficult for the members of a population to share. From the study of embodied agents, we know that agents can synthesize categories that inevitably reflect their own physical experiences and constraints (Pfeifer & Scheier 1999).

Because different agents have different experiences and constraints, it is not trivial for them to share categories among themselves. Fortunately, communication through language actually helps agents share a common category. Steels and Belpaeme's (S&B's) article studies this, taking the category of colors as an example.

Indeed, color seems to be a most intriguing category. As Lakoff (1987) argues in *Woman, Fire and Dangerous Things*, color is a typical example of what he calls a radial category. That is, color lacks a rigid boundary (e.g., between red and blue), as each color is only characterized by a protypical tone of color. Lakoff argues that in addition to color, the grammar of language also forms a radial category. Generally, perceptually grounded categories are best characterized as radial categories.

To study the mechanism by which radial categories develop, we simulated a mobile agent that explores a two-dimensional space. We use a unique neural architecture whereby the agent autonomously opens and closes its own sensory neural connections (Iizuka & Ikegami 2004). Using these mobile agents, we studied frequency discrimination as a case study. We examine how agents, evolving via a genetic algorithm, come to categorize lights that flash with different frequencies into likes and dislikes. When a flashing light is placed in front of an agent, the agent approaches the light if it flashes with a frequency belonging to the "favorite" category, otherwise it is avoided.

We may draw two main conclusions from this study. First, the category of likes and dislikes indeed forms a radial category. That is, no simple criterion exists for the category, such as a particular frequency above which all lights belong to the "favorite" category. Second, the category is not represented in a separate neural substrate. The neural representation of the category coexists in the same neural network which guides the agent's exploring dynamics. Thus we conclude that it is a dynamical category, which can be synthesized via motion behavior but is difficult to be analyzed analytically. The dynamical category uses the sensory-motor coordination to do categorization.

If categories are inevitably grounded on sensory-motor coordination, sharing categories means sharing the sensory-motor coordination. Therefore, S&B's article can be interpreted in terms of the communication among agents involving the agents' sensory-motor coordination.

By designing guessing and discrimination games, S&B suggest that agents develop communication in order to get the meaning of pointing and word expressions. This is also true for color. Communication helps agents to get the meaning of color. Indeed, we claim that to get the meaning of color is to share the same sensory-motor coordinations behind the color categorization. We thus understand that color categories converge via communication.

What we must next consider is the creative role of linguistic communication. We do not just share the same sensory-motor coordination but we acquire the new categorization via communica-

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

tion. A repertoire of language is apparently much larger than that of motion. Therefore, it may provide a new source for the sensory-motor coordination. We thus expect that the diversity of a color category is enhanced by using language, if language can generate more novel sensory-motor coordination of colour experience.

Finally, it is more interesting to study the category for intermediate colors (tones), not the colors themselves, because tones are more subtle than colors and are mixtures of prototypical colors. Without having to worry about such intermediate colors, communication may become easier. Interestingly, however, by using communication the intermediate colors become diverse. In other words, we feel that not only the convergence of color categories, but the increasing diversity of intermediate colors are both caused by social and linguistic communication. This is what we consider to be a creative role of communication beyond sensory-motor coordination.

Sharing perceptually grounded categories in uniform and nonuniform populations

Kimberly A. Jameson

Institute for Mathematical Behavioral Sciences, University of California – Irvine, Irvine, CA 92697-5100 and The Center for Research in Language, University of California–San Diego, La Jolla, CA 92093-0526.
kjameson@uci.edu
http://aris.ss.uci.edu/cogsci/personnel/kjameson/kjameson.html

Abstract: Steels & Belpaeme's (S&B) procedure does not model much of the important variation that occurs across human color categorizers. Human perceptual variation and its corollary consequences impact real-world color categorization. Because of this, investigators with the primary aim of understanding color categorization and naming across cultures should exercise some caution extending these findings to explain how different human societies lexicalize color appearance space.

Steels & Belpaeme (S&B) clearly state that their simulated "perceptually grounded color categories" do not strive to model human categorical representation, but, rather, are practical models of color categorization by artificial embodied agents, or "robots." Their aim is to clarify the conditions under which robot category repertoires make feasible robot communication with human color categorizers. Their approach to color categorization is powerful and refreshingly comprehensive. They synthesize different constraints and contributing factors – biology, psychology, and culture – that are typically pitted against each other in the color cognition and categorization literature.

Despite the authors' statements concerning the work's limited applicability to the behavior of living categorizers, readers of the article are likely to extend these findings to other forms of color categorization phenomena, including human color categorization across cultures. For this reason, discussion is needed of the findings' implications for category processing within and between humans.

S&B state "the agent's architecture is intended to model what we know today about human colour perception, categorization, and naming." (sect. 2.1). Within a population "all agents are assumed to have exactly the same perceptual process." (sect. 2.3.1). And agents base their categorization task decisions (sect. 2.4) on sensory input "So" from a standard uniform perceptual representation (i.e., CIE [Commission Internationale de l'Éclairage, or International Commission of Illumination] L*a*b*). S&B vary spectral distributions that are sampled, but the stimuli are always first converted from spectral distributions to CIE tristimulus values, and then to CIE L*a*b* values before agents engage in any categorization games. Thus agents make all categorization decisions on CIE translations of spectra, rather than on actual spectra (eq. 5). S&B make the following related assumptions:

(A) All agents embody a CIE standard model.

(B) All agents in a population uniformly replicate the same perceptual process.

It should be noted that the above-mentioned details are important considerations if one is seeking an independent neural network or computational modeling verification of human color categories (Cf., Yendrikhovskij 2001a), because these details impose consequences on the artificial network that will influence the network solutions obtained.

Assumption (A) places undeniable constraints on the shared category solutions obtainable by a population of agents. It fixes the dimensionality, metameric class relations and the gamut of the stimulus space to be equivalent with the CIE standard observer. This is a good idea when engineering a robot that strictly aims to perceive spectra with a standardized human eye. Under such circumstances one would minimally expect agents to categorize stimuli in ways compatible with humans, because much of stimulus structure (i.e., metamer equivalences and relations, dimensional structure, and the perceptual gamut of the space) is preprogrammed.[1] Indeed, predefining metameric equivalences alone is enough to establish how agents lump spectra into equivalence classes. This preprogrammed lumping of spectra, however, would not match how most other terrestrial species sort spectral stimuli. Thus, the CIE network-input settings used by S&B differ from others that could be programmed into the agents – say, a standard observer model for male spider monkeys, turtles, or the honeybee – all of which would predictably produce category repertoires fundamentally different from those the authors obtained.[2] Consequently, in addition to S&B's clear disclaimers about the generalizability of the processes by which their robots establish category repertoires, those interested in human color categorization should note that it is also wrong to infer that their networked agents replicate human color categorization behaviors because it is purportedly an optimal species-independent way of categorizing the available terrestrial spectra (this view about optimality appears in the cognitive literature, e.g., Shepard 1994; 1997).

Because assumption (A) predetermines the dimensionality, metamers, and gamut of the categorized color space, a generalization of (A) would also accommodate subspaces of natural categories from agents possessing *fewer* dimensions and *restricted* gamuts. However, such subspaces bring S&B the additional challenge of modeling different metametric class equivalences.[3] Such network modeling adjustments for (A) would be an important step towards modeling human categorization, and bear on assumption (B).

Assumption (B) limits extending S&B's findings to human color categorization, because real human groups that develop and share categorical repertoires are not comprised of individuals with uniform perceptual processing, or uniform color processing expertise. Indeed, relatively minor perceptual variation could significantly impact the network solutions that S&B report. First, considering just perceptual processing variation across agents (i.e., differences of dichromacy and anomalous trichromacy compared to trichromacy), such subgroup processing could impact convergence rates and robustness of solutions under the simplest situations (i.e., learning without language, sect. 3). Discrimination game outcomes could vary for dichromat agents, compared with anomalous-trichromat or trichromat agents, in accord with the observation that "even small variations in colour perception . . . drive . . . colour categories to diverging results" (sect. 5.1). Interactions between actual dichromats and trichromats suggest that perceptual variation effects could extend beyond single agent processing to learning with language scenarios (sect. 4) and guessing game outcomes, making plausible the idea that agent perceptual variation could effect robustness and variance of a population's category repertoire, and, in turn, indices of discrimination success and number of converged on categories.

Human dichromats occur at different rates across ethnolinguistic societies, and, with varying degrees of effectiveness, communicate using trichromat-based lexical categories for which they have no perceptual distinctions (e.g., Jameson & Hurvich 1978;

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

Shepard & Cooper 1992). In one society where rod monochromacy commonly occurs in the population, color normal individuals share a pragmatic categorical repertoire with achromatopes who perceive a "colorless" world (Sacks 1997). In other societies, other complexities arise during processes wherein perceivers learn through social interaction to use normative linguistic codes despite perceptual differences that could undermine the code's meaning (Jameson 2005a; 2005b; Jameson et al. 2001). Thus, within populations, variation in perceptually correlated knowledge is integral to the cognitive side of learning and sharing a color repertoire, but such human variation runs counter to Assumption (B).

Addressing both (A) and (B) as suggested here would permit S&B to make useful comparisons between perceptually grounded categories shared by uniform populations and those shared by nonuniform populations.

NOTES

1. This seems to work against the suggestion that "artificial agents might end up with a quite different categorical repertoire compared to . . . human beings" (sect. 1).

2. Just as S&B demonstrate different sets of "chromatic distributions . . . do not lead to categories that are similar . . ." (sect. 5.1), so too would very different category solutions arise if initially agents were given a honey-bee observer model, and these category solutions would almost certainly bear little resemblance to the category solutions they found using their agent populations.

3. Just as dichromats are accommodated by the CIE standard observer model, but have different known metameric class relations.

Seeing and talking: Whorf wouldn't be satisfied

Boris Kotchoubey

Institute of Medical Psychology and Behavioral Neurobiology, University of Tübingen, 72074 Tübingen, Germany.
boris.kotchoubey@uni-tuebingen.de
http://www.uni-tuebingen.de/medizinischepsychologie/stuff/

Abstract: Although Steeles & Belpaeme's (S&B) results may be useful for development of technical devices, their significance for behavioral sciences is very limited. This is because the question the authors asked was "Why do people use similar words in a similar way?" rather than "How can similar words stand for similar experience?" The main problem is not shared word usage, but shared references.

> Polonius: What do you read, my lord?
> Hamlet: Words, words, words.
> —*Hamlet*, Act II, Scene II

The clarity with which the target article is written makes the critique easier. The main goal is formulated from the very beginning: To explore how colour words "may become sufficiently shared among the members of a population" (sect. 1) so that if I say "red" everybody can select a red (and not a yellow) object from a presented set. Moreover, Steels & Belpaeme (S&B) make no secret that this "goal is entirely practical . . . to design . . . robots that are able to do this task." (sect. 1) Though I am not an expert in robotics, it appears that the authors attained substantial progress in approaching their goal.

The question is, however, whether this pragmatic approach can shed light on the real mechanisms in question. I agree that the study can contribute to "designing agents that are able to develop a repertoire of . . . categories that is sufficiently shared to allow communication" (sect. 6). But I doubt that "these results are relevant to . . . an audience of cognitive scientists" (sect. 6) who are interested in the psychology of colour perception. Although the authors admit that "the artificial agents might end up with a quite different categorical repertoire compared to . . . human beings," (sect. 1) they miss a much worse peril, that their agents come to categories very similar to human categories (thereby creating the illusion of relevance), but using processing means that have nothing in common with those used by human brains.

S&B suggest that their data support the Sapir–Whorf thesis on the dependence of colour perception on language. This thesis has been formulated in rather ambitious terms, for instance, by Sapir: "We see and hear and otherwise experience very largely as we do because the language habits of our community predispose certain choices of interpretation" (cited by Whorf 1962, p. 134), or by Whorf's commentator S. Chase: "Speakers of different languages see the Cosmos differently" (ibid, p. x). Particularly, Whorf emphasised the importance, not only of verbal categories, but rather of the syntax of different languages (e.g., tenses, subject–predicate structure, use of plurals and singulars, etc.), in organisation of our basic mechanisms of perceiving and conceiving of the world.

This expected relationship to the very structure of colour experience is lacking in the target article. Not sharing perception (e.g., the fact that you see red where I also see it) but sharing word usage is the problem the entire study is pivoted around. By the way, colour may not be the best case for study interaction between sensory and cognitive factors because the sensory information can only be obtained with central vision (there are no cones on the periphery) and high luminance (cones do not work in twilight), hence one may state that we see most objects grey most of the time. But the main point is that mere agreement in verbal behavior does not prove the agents' similarity in their "segmentation of the face of nature" (Whorf 1962, p. 241).

Of course, we cannot really know another person's sensory qualia (e.g., the qualium of redness), but we can approach this knowledge by using a broad range of methods, beyond categorisation and naming. And probably the most reliable result obtained to date is that if we vary tasks, conditions, instructions, cue availability, and so forth, so also varies the role of language as a determinant of behavior. Thus, the long-assumed effect of language spatial terms, such as "on the left of" or "to the north of,"on space perception proved to be the effect of available spatial cues. Natural peoples, when tested in their natural conditions, use significantly more objective (allocentric) spatial cues than Europeans (Dutch or English) tested in the lab. Also English-speaking people, without changing their mother language, use more allocentric cues when tested outdoors as compared to being in a closed room with blinds pulled down (Li & Gleitman 2002). The availability of potentially useful information appears, therefore, to exert a stronger effect on space perception than the language itself.

Turning back to colours, the data are not very different. For example, most European languages have one basic term for blue, whereas Russian has two; a popular Russian children's song listing "the seven colors of the rainbow" mentions light-blue and dark-blue as two completely different colours, the latter being close, but not identical, to purple. Nevertheless, being presented with a large number of green and blue colour tones, Russian and English subjects did not differ in their classification; particularly, Russians did not tend to group dark and light blue separately (Davies & Corbett 1997). There is no evidence that English speakers are unable to distinguish those hues that Russian speakers do.

Kay and Kempton (1984) developed colour triads, such as one containing two green colours and one blue. One of the green colours (Green 1) was separated from the other green (Green 2) by a larger number of just noticeable differences than from Blue. When asked to choose the stimulus that looked least like the other two, subjects chose Blue. However, when asked to compare stimuli pairwise, they found Green 1 and Green 2 more different than Green 1 and Blue. The issue may be even more complicated because neuropsychological data indicate that a patient who performed like controls in this experiment (and who, therefore, could distinguish between classification and similarity judgment) was nonetheless unable to classify colours according to their names. His sorting was based on superficial perceptual similarity (Robertson et al. 1999). This may indicate that not only the presence of

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

verbal cues can substantially affect the result of classification, but also the explicit versus implicit nature of those cues.

To summmarise, the Whorfian question was formulated (Li & Gleitman 2002, p. 267) as follows: "Do the differences in how people talk create the differences in how they think?" The target article, in contrast, answered a quite different question: "Do the differences in how people learn to talk create the differences in how they subsequently talk?" It is not surprising that the answer to the latter question was positive, but this does not permit any conclusion concerning the former one.

Not all categories work the same way

Sidney R. Lehky

Computational Neuroscience Laboratory, The Salk Institute, La Jolla, CA 92037. **sidney@salk.edu**

Abstract: The relative contributions of biological and cultural factors in determining category characteristics almost certainly vary for different categories, so that the results of these simulations on color categories don't necessarily generalize. It is suggested here that categories that pick out structure in the environment of strong behavioral significance to individual agents will be predominantly biologically determined and will converge without interagent communication, whereas those categories that serve primarily to coordinate behavior in a population will require communication to converge.

The computer simulations described in Steels & Belpaeme's (S&B's) article provide an interesting example of a situation in which language communication amongst a population of agents can affect the development of color categories. Although the empirical situation regarding color categories, of course, remains to be determined, these theoretical studies will be valuable in constraining the debate about what is possible.

It seems likely, however, that the potential for cultural shaping of perceptual categories can differ sharply depending on the particular category at hand. Some categories may be more culturally dependent, others may have a stronger learning component, and finally, some may be genetically hard-wired. In other words, there may be different categories of categories, and the results of studying one sort of category won't necessarily generalize to others.

What characteristics might in general distinguish more culturally dependent perceptual categories from the more biologically dependent ones? (Here I lump together genetic evolution and individualistic learning under "biology.") I would suggest that if a category represents structure in the environment that is of strong behavioral significance to each individual agent, then that category will develop in a predominantly biology-dependent manner such that all agents in the population, without communicating, will share the same category. If a category does not directly distinguish any behaviorally critical feature of the environment, but rather serves to coordinate the behavior of agents in the population, then communication between agents will be required to ensure the convergence of category properties.

Let us see how this distinction might operate in the context of color categories. CIE (Commission Internationale de L'Éclairage, or International Commission of Illumination) color space, of whatever variant, is a continuous space. The question arises as to in which situations is it advantageous to discretize this continuum into a small set of fixed categories. Two possibilities will be given here, corresponding to the distinction made earlier.

The first is if there were a small set of special colors that flag aspects of the environment that have overriding significance to the agents behaviorally (perhaps related to mating, food selection, predator evasion, etc.). There could, in that case, be an advantage in creating color categories centered on these special colors in order to highlight them for structures associated with implementing decisions and motor responses. To the extent that the embodiment characteristics of agents within a population are essentially the same (similar sensory apparatus, motor capabilities, etc.) and they have similar behavioral repertoires, it seems a reasonable possibility that all agents will converge to these same color categories independent of interagent communication.

The second situation is if agents needed to communicate information about color to each other. In this case, discretization of the color space reflects the discrete nature of the vocabulary used to describe it. Here the color categories don't correspond to anything that is of behavioral significance to an agent operating in isolation. A different set of categories would neither enhance nor detract from the survival prospects of the isolated agent. Thus, there is little pressure for isolated agents to develop the same categories. It is only within a population that the categories acquire significance, and the categories converge through interagent communication to coordinate behavior within the population.

Without the presence of a set of ecologically "special" colors and without language, the "discrimination game" described in the target article could probably be implemented in a robot by setting receiver-operating characteristics within a signal detection model, without fixed categories. Although statistical clustering of natural inputs can lead to the creation of color categories, it is not clear what benefits arise from building a robot with categories derived in this manner, other than perhaps somewhat more efficient encoding of sensory inputs (in an information-theoretic sense, Simoncelli & Olshausen 2001). It is also possible that characteristics of the sensory apparatus may lead to biases in color category formation in noncommunicating agents, so that there is some degree of correlation in the categories formed by them (as indeed we saw in the simulations in this article). However, if these embodiment-specific effects are confined to the input (sensory) stages of the system and do not translate to something behaviorally meaningful, as one considers the agent in its sensorimotor entirety, they may not provide a sufficient drive to strongly coordinate the color categories of noncommunicating agents (again, as we saw in this article).

Moving away from color categorization, consider more abstract categories, such as animals versus non-animals, or food objects versus non-food objects. Membership in these categories can rapidly be determined visually by both humans and nonhuman primates (Fabre-Thorpe 2003). These are examples of perceptual categories that are of strong behavioral significance to individual organisms, perhaps more so than the color categories formed by humans. The expectation here is that individuals undergoing unsupervised learning in their natural environment will be able to converge to the same visual categorization of food versus non-food items (for example), without any communication amongst themselves, to a greater degree than color categories will converge for non-communicating agents.

On sticking labels

Jan Pieter M. A. Maes

Department of Psychiatry, University Hospital Antwerp, 2650 Edegem, Belgium. **janpietermaes@hotmail.com.**

Abstract: Steels & Belpaeme (S&B) are clearly interested in the possible test their models may provide for human language theories. However, they only superficially address the assumptions underlying their own agent architecture, while these are of crucial relevance to the topic of human language. These assumptions fit an Augustinian picture of language, which Wittgenstein challenges in his *Philosophical Investigations*. It is too early to draw conclusions regarding human language evolution from such models.

> Could a machine think? – Could it be in pain? – Well, is the human body to be called such a machine? It surely comes as close as possible to being such a machine.
> —Wittgenstein (1953)

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

Apart from their repeatedly stated practical and pragmatic goals, Steels & Belpaeme (S&B) are clearly interested in the possible test their models may provide for theories about human language and concept formation. Seeing the models in action can bring out hidden assumptions underlying the different approaches, that is, if we assume that the approaches are appropriately modeled in the first place (see later discussion). In discussing their results from this point of view, they emphasize the causal influence of communication and language on category formation and find in it an argument regarding the feasibility of the Sapir–Whorf hypothesis. However, they only superficially address the assumptions underlying their own agent architecture and conceptualization of language, and these are of crucial relevance to this topic. This creates a constant tension in the article regarding the extent to which conclusions regarding the human situation can be drawn. Need their agents be human, or are we humans like their agents?

It is not clear why the authors conclude from their results that language must have a causal role in category formation. If sharing a categorical repertoire is the goal, a genetic scenario without language seems to do the trick just fine in a simpler way. Taking into account real world chromatic distributions increases sharing, as probably a combined genetic and individual learning model – which is not used – would also do. I would rather interpret these results as minimizing the role for language, that is, if the sharing of categories is our only goal. But is it? Why, for example, would we not imagine animals (or agents) and humans developing some shared categories? For both of us, it is not good to eat poisonous mushrooms that are only distinguishable from the good ones by color. Only animals do not talk. There are some fundamental differences between animal and human categories that are not addressed in the article. An animal cannot consider or ask itself what colors are; we can. From this point of view it would be nice to see how the models handle some form of colorblindness (which can be programmed into the constraints coming from embodiment). Even if it is not obvious to imagine what it would be like to be colorblind, it is something we can talk and inquire about. Even if there clearly is no sharing of categories, there still is useful communication. We can see from this that the authors' conceptualization of language and concept formation misses an important aspect of the human situation.

In the learning with language experiments, agents are given the ability to play language games and engage in joint attentional interaction (pointing) to start with. The authors don't discuss how these abilities themselves emerge and their relevance to language and concept formation. They claim that their approach resonates with the philosophical work of Wittgenstein. In this regard, it is interesting to see to what purpose Wittgenstein (1953) invented his language games. In the opening paragraphs of his *Philosophical Investigations* he draws attention to an Augustinian picture of language in which words are learned by ostensive teaching (pointing) alone. He likens this to learning a foreign language when you already have a language of your own. You then only have to learn to correctly stick new labels to already mastered concepts by means of feedback. Only someone who already knows how to do something with it can significantly ask a name of something. Wittgenstein shows that if you do not presuppose such knowledge in language games, pointing would be inherently ambiguous. There would be no shared reference. There is more to language and concept formation than labeling.

In discussing the nativist and empiricist approaches, the authors say that in these theories, the learning of language is just a matter of learning labels for already acquired categories. They do this to highlight the contrast with the culturalist approach in which they reserve an additional, causal role for language. But if we take into account the Augustinian way (by assuming the ability to engage in language games) in which the authors model this approach, it becomes clear that the contrast isn't so great anymore. In the culturalist model, agents are just better at sticking on labels than their counterparts in the other models would be. The authors stay trapped in a referential theory of meaning by coupling names and categories by means of associative networks. It is true that in this model language can have a causal role in category formation. I only want to argue that this way of modeling dangerously simplifies the human situation, if we do not take into account our assumptions when doing so. As a consequence, one cannot draw conclusions regarding the feasibility of the Sapir–Whorf hypothesis based on model mechanics alone. The authors are correct to highlight the fundamental differences between the three approaches, but the way they model them does away with these differences.

The culturalist approach is fundamentally different. In these theories, the "gift of sharing attention and achieving a workable intersubjectivity" (Bruner 2000), together with a nurturing social context serves as a scaffolding for the child's acquisition of language. Just as with language games, there is shared reference and successful communication before there is a fully developed language. With the cultural psychologists like Vygotsky (1934) we can say that these abilities together with developing linguistic abilities are already influencing each other and changing from the beginning. Again, assuming joint attentional interaction and then just building a naming module on top of it is not an appropriate way to model this.

I think the authors have used a novel and interesting approach to the evolution of communication in artificial agents. Maybe one day robots will communicate with each other and human beings using technology based on their models. However tempting, I think it is too early to draw conclusions regarding human language from these kind of models.

Is color perception really categorical?

Mohan Matthen

Department of Philosophy, University of British Columbia, Vancouver, British Columbia V6T 1Z1, Canada. **mohan.matthen@ubc.ca**
http://www.philosophy.ubc.ca/matthen

Abstract: Are color categories the evolutionary product of their usefulness in communication, or is this an accidental benefit they give us? It is argued here that embodiment constraints on color categorization suggest that communication is an add-on at best. Thus, the Steels & Belpaeme (S&B) model may be important in explaining coordination, but only at the margin. Furthermore, the concentration on discrimination is questionable: coclassification is at least as important.

The categoricity of color perception is useful for communication (telling you that my car is green is a lot easier than communicating the exact shade; usually the latter is informationally superfluous anyway), also as a help to color constancy (it appears green in a wide range of illumination conditions, but it looks different shades in these conditions), as well as for retaining eidetic memories within certain boundaries (I will be able visually to recall its greenness long after I have lost the ability correctly to recall its shade). All this tells us little about the origins of color categoricity. Are color categories accidental (albeit useful) side-effects of other features of color processing (e.g., of opponent processing) or did they emerge as a direct result of their usefulness in the previously described ways?

The phonemes constitute a clear case of categorical perception that emerged as a direct result of usefulness – at least if the motor theory of speech perception (Liberman & Mattingly 1985; Liberman et al. 1967) is correct. The motor theory maintains that speech perception is concerned not with acoustic patterns but with the "articulatory gestures" of the human speech production system. The acoustic patterns by which a /b/ is transmitted might be very similar to those that characterize a /d/; nevertheless, these phonemes are produced by different articulatory gestures. Consequently, as a sound pattern gets close to the boundary between these two syllables, it is still heard as an instance of one or another

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

of /b/ or /d/, until, after a relatively narrow boundary zone of ambiguity, the perceived sound flips over to the other. Human speech perception seeks to decode clearly differentiated acoustic gestures; it *uses* acoustic patterns to do this. This is why the boundary between closely adjacent acoustic patterns is heard not as a continuum, but as a boundary. Rather contentiously, I shall call this *paradigmatic* categorical perception.

Color perception is not paradigmatically categorical. There are, of course, basic color terms, as Berlin and Kay (1969) discovered, and these are highly cross-cultural, even if not universal. But, as Steels & Belpaeme (S&B) clearly recognize, the boundary effects here are not anything like those observed in the case of phonemes. There is substantial cross-cultural agreement with respect to the identification of "focal instances" of terms like "black," "white," "red," "blue," and so forth. However, there is not much agreement about the boundaries of these terms. In the case of color, the perceptual "flip" as one moves from blue to green is not marked. This is the exact opposite of the case with phonemes, where focal instances do not exist but the boundaries are sharp (though not cross-cultural, because the learning of specific languages seems to erase some phonemes that are perceptually available at birth, see Werker & Tees 1984).

The empiricist theory of concept-building has been encapsulated in a model of *paradigm* and *foils* (Quine 1969). The paradigm is the central instance of a concept, generally identified by a convention-creating act of stipulation; the foils are supposed to be *counter*instances that lie just beyond the concept boundary, again identified conventionally. A concept is defined as comprising everything more similar to the paradigm than the foils. Obviously, the model will work only for "perceptually grounded" concepts involving a single submodality such as color, because similarity is well-defined only within the similarity spaces of such qualities. For things outside the context of such similarity spaces, similarity is notoriously illogical (Tversky 1977), and cannot be made to perform consistently even in the most stringent learning regimes.

Now, one possibility is that human color categorization emerges from the paradigm–foil method constrained not by stipulation and convention, but by the salience of the so-called unique colours, that is, those colors that consist of a single Hering primary. These primaries are themselves subject to a certain amount of interpersonal variation (Kuehni 2004), which suggests biological but not interpersonal constraint. However, the boundaries of color categories, that is, the foils, might be roughly determined by equal distance measures between foci. Thus, the S&B model of interpersonal feedback might prove useful, but only after the biological embodiment of color processing in a given individual has done its work. This would indicate a fuzzy concept centered on clear but variant foci, which, as we have seen, is, in fact, what we find (Kay & McDaniel 1978).

In such an account, embodiment shapes the concepts only by creating natural paradigms and through the innate similarity space of color, but not by giving us whole concepts ready-made. Thus, the nativism of this account is tempered by a kind of developmental dynamic. We might have as many categories as we do, simply because the basic colors fall out of a topologically natural way of carving up color space (Jameson 2005b). Culture or environment may intrude to *reduce* the saliency of some of these factors, as they do in the case of phonetic perception. (The cultural reduction of salience may well be sufficient to explain what variance in categories there is across populations, and individual variations in color perception can explain variance within populations.) Unlike the boundaries between phonemes, color boundaries do not represent anything other than themselves. Thus, the colors might well be a natural system, largely explained by embodiment constraints on the empiricist model of learning. Though such a system would assist us in communicative tasks, it is hard to be sure that this is nonaccidental.

S&B make *discrimination* the determinant of categorization, in other words, they use this task to replace the constrained traditional empiricist model outlined earlier. There are two points to be made here.

First, discrimination is, of course, not the only thing that we do with color. We also use color to coclassify things (i.e., to put them in the same rather than different categories) in order to mark them for object reidentification and association with other characteristics (we assume, for example, that *mutatis mutandis*, two fruit of very similar color are equally ripe). One might think that coclassification is actually more fundamental to understanding categorization than discrimination. (On this point, see Matthen 2005, especially Chaps. 1 and 11.) It is hard to know how coclassification influences the shape and number of color categories. Does it interfere with or reinforce the categories spawned by the discrimination game?

Second, in the case of human development, the existence of distinguished points in color space (e.g., the unique hues) could have ensured that there were focal points for categories in place before any discrimination tasks were run. This would also account for the sharing of categories across populations. In other words, the addition of distinguished points in color space, as suggested earlier, might well reduce the importance of the simulations run here.

How culture might constrain color categories

Debi Roberson and Catherine O'Hanlon

Department of Psychology, University of Essex, Colchester, Essex, CO4 3SQ, United Kingdom. **robedd@essex.ac.uk**
http://www.essex.ac.uk/psychology/psychology/CLIENTS/debiRoberson/debiRoberson.html **cgohan@essex.ac.uk**
http://www.essex.ac.uk/psychology/psychology/PhDstudents/OHanlon.html

Abstract: If language is crucial to the development of shared colour categories, how might cultural constraints influence the development of divergent category sets? We propose that communities arrive at different sets of categories because the tendency to group by perceptual similarity interacts with environmental factors (differential access to dying and printing technologies), to make different systems optimal for communication in different situations.

Steels & Belpaeme's (S&B's) study introduces a novel and ingenious method for comparing nativist, empiricist, and culturalist accounts of colour categorisation. Their findings suggest that language is crucial to the development of a shared categorical repertoire. They make the point, also made by Roberson et al. (2000), that acceptance of a causal role for language does not imply that colour categories are free to vary arbitrarily, because physiological and environmental factors also affect the process, but they comment that "it is less obvious by what kind of process cultural constraints could play a role" (sect. 1.2). We here suggest two additional constraints (over and above those imposed by physiology and the visual environment) that we believe operate on the range of potential category sets and a process by which they might interact with environmental constraints (the context from which stimuli must be discriminated). They are, first, a universal tendency to group by similarity, and second, the differential need to communicate successfully experienced by different cultures about a shared set of categories.

The tendency to group by similarity is pervasive, both across cultures and across cognitive domains. Colour cognition is no exception to this and no culture or language has yet been reported that violates this principle by grouping together two areas of colour space (e.g., yellow and blue) in a category that excludes the intermediate area (e.g., green). Young children (from two very different cultures) group colours on the basis of perceptual similarity before they acquire any colour categories (Roberson et al. 2004), as does an adult patient with colour anomia, who had lost the ability to categorise colours explicitly (Roberson et al. 1999). Drawing children's attention to the relative similarity of colours,

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

through linguistic contrast, also promotes faster category learning (Au & Laframboise 1990; O'Hanlon & Roberson 2004).

If categories are initially formed based on the relative similarity of stimuli, as Dedrick (1996) and Roberson et al. (2000) have argued, then both the range of available stimuli in the environment and variability in the need to communicate about colour should affect the eventual set that a community arrives at. Communicative need varies widely across populations (Heider & Olivier 1972; Jameson & Alvarado 2003a; Kuschel & Monberg 1974; Levinson 1997; MacLaury 1987) and different sets of categories may be optimal for colour communication in Western societies compared to traditional cultures lacking printing or dyeing technologies (Davidoff et al. 1999; Roberson et al. 2005). In a recent study, Roberson and Agrillo (under review) used a human communication task similar to that modeled here. For English speakers, the central (prototypical) exemplars of their eleven basic colour categories were communicated more successfully than other colour stimuli in the context of the test array (160 Munsell samples from a standard uniform distribution). These eleven terms, common to most Western languages, may well be an optimal set for their communicative needs (Guest & Van Laar 2002), given the wide range of available colours in their environment. However, the usefulness of the name in distinguishing individual items of a set is attenuated by context. Where all mushrooms are orange, the term "orange" is less useful.

Many traditional cultures, however, have fewer than eleven categories, each containing a wide range of exemplars, extending to very desaturated colours, and with little interindividual agreement on where the best examples of categories are located (Heider & Olivier 1972; MacLaury 1987; Roberson et al. 2000; 2005). Without the full range of saturated stimuli that can be artificially produced, traditional communities may have no need of the finer categorical distinctions required when a wider variety is available, and thus lack the motivation to refine their colour lexicon further. At this stage, "different ecological and cultural circumstances" (sect. 3.3) increase the divergence of colour categories in these communities. The finding that "statistical extraction of categories from natural colour data" (sect. 5.1) does not deliver the eleven basic categories found in most Western languages supports this hypothesis. However, as S&B show, even for categories formed from a data set of stimuli from natural surroundings, language is still "crucial for the convergence of colour categories" (sect. 4.3). The reduced variance of the stimulus set may, however, result in the formation of fewer categories.

What then might provide the impetus for convergence of colour lexicons as contact between communities increases? Industrialisation may encourage the introduction of new terms (and categories) to increase both communicative power and discriminability. MacKeigan's work with Mayan weavers suggests that the introduction of new colours of thread directly relates to changing colour vocabulary, perhaps by altering the discriminability of colors within an available set (MacKeigan 2004). Because discriminability is a property of a colour stimulus "in the context of a particular array" (Lucy 1992, p. 165), the introduction of a new range of colours reduces the communicative success of a previously optimal set of categories, and thus motivates change.

This combination of decreased discriminability (because of an increase in the range of available stimuli) and the need for communicative success, combined with a shared tendency to group by similarity, could yield convergence of colour categories and color lexicons across populations without a requirement for a genetically determined set. As S&B have shown, shared categories emerge fastest where there is "strong structural coupling between concept acquisition and lexicon formation" (sect. 5).

It takes a(n) (agent-based) village

Teresa Satterfield

Department of Romance Languages & Center of Study for Complex Systems, University of Michigan, Ann Arbor, MI 48109-1275.
tsatter@umich.edu **http://www.umich.edu/~tsatter**

Abstract: Steels & Belpaeme (S&B) take technical and conceptual shortcuts that have significant negative consequences on simulation implementation and agent behavior. Justifiably, their model represents a proof of concept of the role of culturalism in category formation; nevertheless, the absence of detailed information concerning the embodiment of agents and the superficial implementation of learning approaches make their results less relevant than they could be.

One advantage of computational modeling over empirical research is the ability to abstract away layers of the real world to focus on aspects considered most relevant for the problem to be solved. However, in Steels & Belpaeme's (S&B's) attempt to address the "acquisition problem" in terms of human color perception, categorization, and naming across a population, perhaps they go too far in paring down the object of inquiry. They do not specify, or they eliminate altogether, a number of "details" that they deem unnecessary to the learning process, contrary to empirical data and current theoretical formulations. In all theoretical approaches simulated by S&B, the agent populations reach final states reflecting successful acquisition of color categories and corresponding lexicons for the given environment and ecology, which subsequently are shared across a speech community. However, to obtain these results, S&B take technical and conceptual shortcuts that have significant negative consequences on simulation implementation and agent behavior. Justifiably, the model represents a proof of concept of the role of culturalism in category formation; nevertheless, the absence of detailed information concerning the embodiment (internal attributes and states) of agents and the superficial implementation of learning approaches make S&B's simulation results less relevant than they could be, either for comparison of the different theoretical positions or for understanding the human phenomena investigated.

Conceptually, the key problem is S&B's drastic simplification of the learning/acquisition approaches explored. Nativism is sketched as "all humans could be born with the same perceptually-grounded categories as part of their 'mentalese'." By representing innateness (formalized as a genetic algorithm) as a fixed network throughout the agent's lifetime, S&B equate nativism to an adult-like, fully-operational repertoire of categories springing forth at birth, requiring no further experience or maturation/development. Change occurs in each successive generation through genetic mutation. Yet, empirical observations indicate that despite our genetic endowment, a child raised without exposure to any human language will never come to speak one (Lenneberg 1967). Nativists acknowledge both experience-dependent mechanisms and maturational constraints in the learning/acquisition process (e.g., Chomsky 1980). Furthermore, S&B fail to keep the critical distinction between acquisition and evolution. Phylogeny (how perceptually-grounded categories evolved in the species over time) does not necessarily beget ontogeny (developmental properties of day-to-day learning by the child), and it is too simplistic to suggest that the underlying mechanisms for these different domains are identical within a nativist approach. S&B alternatively claim in empiricist accounts (formalized as connectionist models) that "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" (sect. 1). This overgeneralization is not only at odds with attested maturational factors accounting for first (L1) versus second language (L2) learning (Newport 1990; 1991; Sorace 2003), but also with connectionist models of L1 learning in which neural networks function optimally when forced to "start small," thereby undergoing a developmental change that resembles the incremental increase in working memory occurring over time in

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

children (Elman 1993). Although S&B represent empiricist learning as adaptive categorical networks during the agent's lifetime, they stop short at realistically constructing models in which agents within the same population vary developmentally.

Empirical studies pertaining to vision demonstrate similar patterns of emergence. Infants show true color vision when they are able to discriminate between two stimuli of different wavelength, but equal luminance. At two to four months of age, infants can discern chromatic differences fundamentally at adult isoluminance (Teller 1998). This said, color vision reaches its adult form only in early adolescent years. Relative sensitivity to varying wavelengths is said to change between infancy and adolescence, with the red–green mechanism appearing to develop before the yellow–blue mechanism (Teller 1998). It is also hypothesized that "appropriate color naming depends on maturation and integration of specific cortical neurological structures" (Bornstein 1985). S&B claim to capture the "prototypical nature of colour categorisation, as demonstrated by the naming and memory experiments" (sect. 1.3) through the use of neural nets; however it is difficult to know whether they accommodate both adult and child learners.

The second problem is a technical shortcoming of S&B's models, which, in fairness, falls out somewhat from the conceptual defects. Absent in S&B's simulations is the role of maturational/developmental factors in constraining learning/acquisition. Measures such as, "all agents are assumed to have exactly the same perceptual process," sidestep the point that infant perceptual processes may be sufficiently different from adults. If so, S&B's predictions are potentially invalid for the discrimination game, where the "best" category is found through adult sensory representation (computed CIE L*a*b* values). Success in the discrimination task is crucial, as it is a prerequisite for ongoing communication. Agents are initialized with zero categories acquired, yet all possess relatively powerful capabilities of perception, associative memory, and a pregiven repertoire/alphabet of syllables. One speculates that these experiments depict a homogeneous group of agents who are either "wunderkinds," or cognitively challenged adults (both possibilities represent individuals with mature/adult-like capacities in language and vision, with immature repertoires of color categories). Regardless, either population is anomalous. One benefit of agent-based models is that we are able to design multiple varieties of agents, whereby each agent retains its profile of internal characteristics across a designated lifespan (Epstein & Axtell 1996; Ferber 1998). Parameters including age, working memory, attention, lexicon, and perception, and so forth, can be used in a distributed population of heterogeneous child and adult agents (Satterfield 2001; 2005). These properties are easily integrated as instance variables in the models, and can further constrain agent interaction (culturalism). Lastly, to fully exploit the power of agent-based models in generating complex structures from the "bottom-up," the initial attributes of each agent in the population must be made explicit. S&B are vague about initial states and specific attributes are not ascribed, nor is further elaboration given on the agents' basic architecture for perception, categorization, and naming, beyond "all agents have . . . unique associated information structures, representing its repertoire of categories and its lexicon" (sect. 2.1) Moreover, the logic of learning/acquisition theories dictates that the initial state of the learner be outlined, in order to make informed evaluations with respect to the learner's final state.

Colour is a culturalist category

J. van Brakel

Department of Philosophy, Catholic University of Leuven, 3000 Leuven, Belgium. **Jaap.vanBrakel@hiw.kuleuven.ac.be**

Abstract: Extrapolation of Steels & Belpaeme's (S&B) results show that colour is a *culturalist* category. Populations will *only* share the category of colour if it is built into the system. If "left to themselves" different populations may or may not stumble on the colour category. Populations that do not share a colour category may still be able to communicate in a wide variety of environments.

Although Steels & Belpaeme's (S&B) agents are "grounded," "embodied," "situated," and "cultural," nevertheless their target article is a beautiful example of agents acting in a block world, that is, in a world consisting of "1269 matte finished Munsell colour chips" (sect. 2.2) or "25,000 pixels drawn randomly from photographs of animals, plants, and landscapes" (sect. 5.1). S&B's theoretical models and simulations bring out many points relevant to the debate between innatists, empiricists, and culturalists, as this debate has been going on for more than a century. But this debate is staged on the wrong assumptions.

Although S&B may agree that "1,269 matte-finished Munsell colour chips" is a block world, they may reply that making the models "more complex and more realistic" will "be more of a hindrance than a help because it would obscure the contribution of the dynamics" (sect. 6). However, the kind of complexity and realism they are thinking of would only make their block world, *as already defined*, more complex and realistic. It wouldn't address the hidden presupposition shared by innatists, empiricists, culturalists, and S&B alike, which can be succinctly expressed by the rhetorical question: "What if colour is a culturalist category?" There is no evidence that shows beyond any doubt that *colour* is not a culturalist category. Note that it would have to be shown beyond any doubt, because if it is a possibility, the rhetorical question needs to be addressed; in particular as S&B are keen to bring out hidden assumptions of various positions. Moreover, *that* colour is a culturalist category is supported by extrapolating the arguments and simulations S&B present. Sections 3.2, 3.3, 4.3, and 4.4 all conclude that "colour categories are not shared across populations." The only reason why these populations are sharing colour categories is because that constraint ("defining" the world as a Carnapian colour chip/pixel world) has been built into the design of the particular hardware and software used, the environments provided, as well as all the knowledge that went into these designs and the (meta)language used to speak about what the agents are doing when playing their games.

Imagine a population of agents that is using six colour words (trained/developed in any of the models of S&B) meeting a population of agents that also use six words. For convenience sake I will translate the vocabulary of the latter population as follows: very bright, medium bright, dark, unripe, ripe, overripe. What will happen? I surmise that for a wide range of environments the two populations would quite easily learn to communicate successfully (in terms of S&B's criteria) and over time through more interactions this will improve further. The fact that one population is using "colour words" and the other is using labels for what I have glossed as brightness and ripeness is quite irrelevant for the degree of convergence in their communicative success.

The model I sketch (and which would need to be simulated) is somewhat similar to Quine's gavagai example of radical translation (which S&B discuss). But it is not the case, as they suggest, that the different (meta)categories will be disentangled in one unique way when the "multicultural" population is presented with a greater variety of environments. Of course (the typical situation when human cultures interact), one of the two populations may dominate the shared form of life and after a few generations all agents may communicate in terms of brightness and ripeness and colour may have been "forgotten." The reason it may *seem* that "incoherence is disentangled when situations arise where two

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

meanings are incompatible" (sect. 4.2), is because one knows or can easily stipulate a vocabulary in terms of which we (the experimenters) can disentangle the situation. Of course, in specific cases, something we might call disentanglement may occur. But there is no unique way for disentanglement to go. This is also the point of Quine's radical translation thought experiment.

The arguments and simulations of S&B that support their conclusion that different populations will have different colour categories will simply repeat themselves at a higher level of abstraction. Different "multicultural" encounters will converge to different metacategories. Populations will only share the category of colour (early and late), if it has been built into the system as an absolute feature of the block worlds they live in. Language learning is not only "crucial for the convergence of colour categories" (sect. 4.3), but also for convergence on the category of colour. Which metacategories will survive "depends on environmental, ecological, and physiological constraints, but there are multiple solutions" (sect. 3.3). The chicken and egg problem (sect. 1.2, 6.2) applies not only to naming colour *categories*, but also to calling "it" a *colour* category.

As S&B correctly point out – a point which cannot be emphasised sufficiently – no culturalist has ever said that the choice of colour categories is arbitrary. A culturalist doesn't deny that there are "cross-cultural trends that have been observed in colour naming" (sect. 6). This is self-evident, virtually a priori, given that the only data studied are *colour naming* data. Similarly, no culturalist will ever say that the category of colour is arbitrary. As we know from science fiction literature, both the innate hardware (of human beings or other agents) and the physical environment place strict limitations on what categories are possible, that is, which categories are afforded by the (innatists') hardware and the (empiricists') environment. But, as S&B point out as well, these severe constraints still leave open innumerable options. By recognising this openness, including the theoretical recognition that colour is a culturalist category, there will be more options for survival and more options for developing interesting artificial intelligence (AI) models.

AI could make a real contribution, if it would study more complex models, not in the sense of better approaching, say, colour perception of twenty-first-century humans, but in the sense of dropping more and more culturalist assumptions. S&B's correct advice that we should not "program into the agents a specific repertoire of categories" (sect. 1) applies to the *specific* category of colour as well.

A categorial mutation

Oscar Vilarroya

Unitat de Recerca en Neurociència Cognitiva, Departament de Psiquatria i Medicina Legal, Universitat Autònoma de Barcelona, 08035 Barcelona, Spain. **oscar.vilarroya@uab.es**

Abstract: The proposal of Steels & Belpaeme (S&B) is on the right track to solve the nativist/empiricist/culturalist controversy. However, their nativist model of colour categorization does not correspond to a proper genetic model. Colour perception is the outcome of a complex process of development. A direct correspondence between genes and colour categories cannot be the right approach to the problem.

The proposal of Steels & Belpaeme (S&B) is a breath of fresh air. The idea that such an old issue as the nativist/empiricist/culturalist controversy may be explored empirically must be welcomed. I believe that the question of concept formation and shared categories will not be resolved until we have found the underlying developmental and learning mechanisms.

The fact that the authors use artificial theoretical models to test the hypothesis does not diminish the relevance of their proposal. Some may be tempted to claim otherwise. However, one can argue that learnability assumptions must be theoretically grounded before any empirical data can be fully interpreted. A successful theory of category acquisition must prove that the learning mechanisms proposed by the theorists not only account for all the data under consideration (descriptive adequacy), but also for all the possible categories that can be acquired, given the type of data and cognitive resources that are typical of human conceptualization (explanatory adequacy). If the authors show that category sharing can be attained exclusively by: (a) genetically determined pathways present in a population; (b) extracting information from the environment; or (c) cultural constraints in a community, then the arguments against the feasibility of any of the positions in question loses force. If the outcome is negative, as has been the case in the empiricist model, the supporters of such a position must counter with models that are able to attain such a competence, or abandon the hypothesis altogether.

Despite my support for S&B's overall proposal, I do have a serious objection. This concerns the nature of one of the models in the comparison. S&B have slipped the label "genetic" into their nativist model of colour perception. Such a model is not an adequate characterization of how genes are related to behavioral competences. Note that I am not claiming that S&B's nativist model is inappropiate for humans, but rather that the model is simply not a genetic model of *any* cognitive ability.

In their presentation of the nativist approach, the authors propose "a model of genetic evolution capable of evolving 'genes' for focal colours" (sect. 1.3). However, in the context of relating genetics and cognition, locutions indicating that focal-colour perception has roots in the human genome does not entail that there are genes for focal colour categories. The fact of the matter is that every human cognitive capacity has a genetic basis. For example, proficiency in reading or socializing ultimately depends on the development of specific biological structures (Lewin et al. 1997; Marshall 1980; among others). Let me expand on this.

It is uncontroversial that "genes for X" assertions are susceptible to the reading that a gene (or a set of genes) could, in a very literal sense, encode a behavioral faculty. A gene of this kind would be a set of instructions for the cortical representations that implement the faculty in question. Unfortunately, genes do not work this way.

The assumption behind S&B's genetic model is that phenotypical traits are somehow represented in the genome. This unfortunately turns genes into what Schaffner (1998) labels "traitunculi," that is, copies of a trait codified in certain stretches of DNA. Genes, however, have no representational resources to specify phenotypical traits (Elman et al. 1996; Nijhout 1990; Oyama 1985, among others). The way a gene is expressed as, say, a behavioral output is the result of a complex intermeshing of processes requiring many developmental levels and components. Now, the existence of these intermediate steps between genes and behaviors is not only a question of complexity and distance, or of a simple hierarchical control structure (Schlichting & Pigliucci 1998), but also of emergence at each developmental level. Although genes may be understood as coding for the production of proteins, they do not encode *how* proteins interact, much less the more distal effects, such as *how* cells and tissues communicate, or *how* the central nervous system forms. Consequently, genes do not hold a blueprint or a program for the development of an organism. Even more so, genes definitely are *not* a blueprint or program for the specific functions or behaviors that are needed for the complete developmental process.

The so-called genetic determination of a certain phenotype is, in reality, the product of the whole epigenetic process and not only that of genes. The implicit belief that a gene might control the development of a certain behavioral competence, such as colour perception stems, perhaps, from the empirical fact that certain gene mutations and deletions may change the course of development in remarkable ways. However, to paraphrase Nijhout (1990, p. 442), this is like equating the steering wheel with the driver. All that a specific connection between a given gene and a trait shows is that the gene's protein is necessary for the normal development

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

of the phenotype. Finding such a correlation is neither an end unto itself, nor is it really an answer; rather, it is, at best, only a first methodological step that can be used to manipulate and explore the developmental process at hand (Nijhout 1990).

The bottom line here is that the S&B's nativist model, despite its success, is not a good model of genetic development of colour categorization. The proof for this lies in the very evolutionary dynamics that the authors' model provides. Focal perception is a complex process. It is not only determined by genetics, but by development, neurophysiological constraints, as well as experience. The highly rapid and diverse evolutionary dynamics observed in S&B's experiment is hardly, if ever, possible for complex neurophysiological functions (Bowmaker 1998; Lickliter & Honeycutt 2003; Schlichting & Pigliucci 1998; Surridge et al. 2003; Worden 1995). This in itself casts doubts on the appropriateness of the model.

In sum, I applaud S&B's efforts to model category sharing. By the same token, I believe that they need to come up with a better model for the genetics of colour categorization. I encourage them to do so.

ACKNOWLEDGMENT

I thank Joe Hilferty for discussions of this article. Portions of this article have been extracted from a submitted manuscript by Oscar Vilarroya and Joseph Hilferty "The gene out of the bottle" and from the communications presented by the same authors at the 8th International Cognitive Linguistics Conference (Logroño, Spain, July 2003).

Learning colour words is slow: A cross-situational learning account

Paul Vogt[a,b] and Andrew D. M. Smith[a]

[a]*Language Evolution and Computation Research Unit, School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, Edinburgh EH8 9LL, United Kingdom;* [b]*Induction of Linguistic Knowledge, Computational Linguistics, Tilburg University, 5000 LE Tilburg, The Netherlands.* **paulv@ling.ed.ac.uk**
http://www.ling.ed.ac.uk/~paulv **andrew@ling.ed.ac.uk**
http://www.ling.ed.ac.uk/~andrew

Abstract Research into child language reveals that it takes a long time for children to learn the correct mapping of colour words. Steels & Belpaeme's (S&B's) guessing game, however, models fast learning of words. We discuss computational studies based on cross-situational learning, which yield results that are more consistent with the empirical child language data than those obtained by S&B.

Steels & Belpaeme (S&B) have successfully shown how computational modeling can contribute greatly to the study of the evolution of language and cognition. S&B have – in our opinion correctly – decided to write their article from an engineer's point of view. We feel, however, that their model of linguistic communication would have been more realistic, and therefore the results they obtained more robust, if they had used a model of acquiring colour categories through multiple contexts.

S&B model the communication between agents using the *guessing game* model, which is, in itself, not unreasonable. Their claim, however, that this game is "equivalent" to colour chip naming experiments carried out by anthropologists (sect. 2.4.2), is not justified, in our opinion. The guessing game is primarily a model of learning through corrective feedback, whereas colour chip naming experiments consist of an anthropologist (A) asking an informant (B) to point out, on a chip set, the focal colour of a colour term from B's language. There are three important differences between the anthropological experiments and the guessing game. First, B is not doing any learning, in fact, A is learning about B's representation of colour and about B's language. Second, A does not correct B's responses or provide any feedback about them. Finally, there is no negotiation between A and B about what the words should refer to.

This positive feedback loop between the choice of which words to use and their success in communication is the main learning mechanism in the guessing game. Indeed, S&B claim that the feedback loop is a *necessary* requirement for cultural language development (sect. 5, condition 1), although in fact it is widely accepted that children receive little, if any, corrective feedback while learning words (Bloom 2000, but see Chouinard & Clark 2003, for an alternative account). In computational simulations of lexicon creation and learning, similar to those presented by S&B, we have shown that agents using a *cross-situational statistical learner* (a variant of Siskind's, 1996, cross-situational learner) can successfully develop a shared vocabulary of grounded word meanings *without* corrective feedback (Smith 2003; Vogt 2004). In our model, as in guessing games, hearers have to infer what speakers are referring to, but unlike in guessing games, the agents do *not* have any way of verifying the effectiveness of their attempts at communication. Instead, the agents use covariances to learn a mapping between words and categories based on the cooccurrence of words and potential referents across multiple situations.

Although young children do learn to relate colour terms to colours, it takes them a considerable length of time to find the appropriate mappings (e.g., Andrick & Tager-Flusberg 1986; Sandhofer & Smith 2001). For example, it has been estimated that, on average, children required over 1,000 trials to learn the three basic colour terms "red," "green," and "yellow" (Rice 1980, cited in Sandhofer & Smith 2001). Sandhofer and Smith suggest that children go through different stages in learning colour words: First they appear to learn that colour terms relate to the domain of colour, and only then can they actually learn the correct mapping. This has also been observed by Andrick and Tager-Flusberg (1986), who additionally suggest that children find it difficult to learn the boundaries of colour categories, thus slowing down the learning of colour words. Research into child lexical acquisition is, of course, dominated by the problem of referential indeterminacy, and many constraints have been suggested to explain how children reduce indeterminacy (see, e.g., Bloom 2000). Very few of these accounts, however, allow for the fact that children hear words in multiple different contexts, and can use this to determine the intended reference. Recent empirical research, indeed, shows that a cross-situational model of learning provides a robust account of lexical acquisition in general, and of the acquisition of adjectives, including colour categories, in particular. Houston-Price et al. (2003) suggest that the children in their study used cross-situational learning to disambiguate word reference, even though their experiments were designed with attentional cues. In addition, Mather and Schafer (2004) show that children can learn the reference of nouns by exploiting covariations across multiple contexts. Akhtar and Montague (1999) demonstrate that children use cross-situational learning to discover the meanings of novel adjectives. Klibanoff and Waxman (2000), furthermore, provide empirical support for their proposal that adjectival categories are learned cross-situationally, within the context of basic level categories.

A comparison of the guessing game and a cross-situational statistical learner, using computational simulations, has shown that, in the guessing game, coherence in production between agents is considerably higher and that learning is much faster (Vogt & Coumans 2003). This means that agents using cross-situational statistical learning have considerable difficulties in arriving at a shared lexicon, although in the end they manage to overcome them. Note, however, that cross-situational statistical learning improves when: agents' semantic categories are similar (Smith 2003); learners assume mutual exclusivity (Smith 2005); and the context size is relatively small (Smith & Vogt 2004). This slower rate of acquisition is thus consistent with the empirical evidence that children learn colour words relatively slowly. Importantly, as yet unpublished studies have shown that the category variance among agents in the cross-situational learner tends to be much higher than that seen from the guessing games. This suggests that negotiating category boundaries in the cross-situational learner is more

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

difficult, which could confirm Andrick and Tager-Flusberg's (1986) finding.

S&B have presented a model of learning colour words that is fast and based on corrective feedback. Research on child lexicon acquisition suggests, however, that colour categories are actually acquired slowly and through cross-situational learning. If cross-situational learning is, indeed, a more plausible model than the guessing game, then the results achieved by S&B may no longer hold for their account of cultural learning.

Interindividual variation in human color categories: Evidence against strong influence of language

Thomas Wachtler

Department of Physics/Neurophysics, Philipps-University, 35039 Marburg, Germany. **thomas.wachtler@physik.uni-marburg.de**
http://neuro.physik.uni-marburg.de/~wachtler

Abstract: With respect to human color categories, Steels & Belpaeme's (S&B's) simulations over-emphasize the possible influence of language. In humans, color processing is the result of a long evolutionary process in which categories developed without language. Common principles of color processing lead to similar color categories, but interindividual variation in color categories exists. Even color-deficiencies, causing large differences in color categories, remain inconspicuous in everyday life, thereby contradicting the hypothesis that language could play a role in color category formation.

The main focus of Steels & Belpaeme's (S&B) study is category formation in artificial agents and the role language could play in this process. Beyond that, they consider the possible relevance of language for color categories in humans. Neither issue seems to be adequately addressed by the simulations.

First, under the conditions assumed by S&B, it is almost trivial that language would make categories of the simulated artificial agents more similar. S&B specify their stimuli in a homogeneous color space. Because there are no constraints or dissipative mechanisms, *any* kind of coupling will increase the similarity between categories, and eventually lead to identical categories. S&B introduce such coupling in their simulations with color genes, language, or nonuniform stimulus distributions. Not surprisingly, in all of these cases, categories of different agents become similar. Even for the simple case of artificial agents, however, success of the "sharing by language" strategy requires that communication corresponding to "guessing games" would occur with fairly high frequency.

Second, for the case of human color categories, the scenarios considered by S&B are similarly inappropriate. They ignore, for example, the properties and constraints of neural processing and representation in the visual system. There are strong nonlinear mechanisms, such as the division in On- and Off-pathways, which effectively segregates color space into categorical half-spaces. Chromatic preferences of color-selective neurons tend to cluster, both at precortical stages (e.g., Derrington et al. 1984) and in the visual cortex (De Valois et al. 2000; Kiper et al. 1997; Komatsu et al. 1992; Lennie et al. 1990; Wachtler et al. 2003). In other words, not all chromaticities are equal. So far, the exact relation between coding at early stages of the visual system and perceptual categories is still unclear (see e.g., the comments on Saunders & van Brakel 1997; for a recent discussion see Valberg 2001). Nevertheless, nonuniform distribution of color preferences places constraints on category formation. Similarities between the properties of neurons in the visual system and efficient codes for natural colors (Caywood et al. 2004; Lee et al. 2002; Wachtler et al. 2001) further indicate that color vision is adapted to the statistics of natural chromatic signals, which implies shared categories.

The corresponding genetic coupling of color categories is not realized by "color genes," but rather by the genes that control the development and function of the visual system. These genes evolved over many millions of years, and evolutionary success was not determined by successful communication, but by efficient processing of visual information, probably including such important tasks as image segmentation and the finding of food (e.g., Mollon 1989). Experimental evidence for shared color categories has been found in other species, as well, such as chimpanzees (Matsuno et al. 2004) or even flies (Troje 1993).

Despite the common processing principles underlying human color vision, there are considerable interindividual differences in the prereceptoral, receptoral, and postreceptoral stages of visual processing. As a result, for example, the loci of unique hues are broadly distributed (e.g., Webster et al. 2000). How does this variability compare to the results of S&B's simulations? S&B fail to specify how similar categories have to be in order to be "sufficiently shared." In any case, with respect to human color vision, S&B's ideal of "complete" sharing is not realistic.

Variation in color vision is most striking in "color-blind" subjects. In dichromats, such as protanopes or deuteranopes, one type of cone photoreceptor is entirely missing. Interestingly, despite their receptoral color space of reduced dimensionality, dichromats use the same basic color terms as trichromats when asked to describe their color percepts (Boynton & Scheibner 1967). However, their category regions in color space differ considerably from those of trichromats (Wachtler 2004). This is not surprising, because certain colors belonging to different categories of trichromats, such as trichromats' reds and greens, are indistinguishable for dichromats. Nevertheless, dichromats seem to possess perceptual categories corresponding to those of trichromats, and they seem to achieve them by dividing their reduced color space using both spectral composition and luminance (Boynton & Scheibner 1967; Jameson & Hurvich 1978; Wachtler et al. 2004).

Several lines of evidence indicate that the color categories of dichromats revealed by color naming reflect perceptual categories. For example, dichromats claim that "red," "green," "blue," and "yellow" constitute unique and different percepts. Furthermore, dichromats consistently report a "red" contribution both in short-wavelength and in long-wavelength stimuli, asserting that the "red" is of the same perceptual quality in both cases (Wachtler et al. 2004).

The color naming behavior of color-deficient observers suggests that language plays a role in the acquisition of the lexicon of color names, but does not influence perceptual categories. Just like color-normals, dichromats have to learn the words to name their percepts. Given that the structure of their color space is different, they cannot achieve a perfect match, so they assign those names to their perceptual categories that constitute the best possible match to those of trichromats.

It is impossible for dichromats to have the same categories as color-normal trichromats. If communication about color would be as crucial as S&B suggest, dichromats would be lost in continuous frustration. No matter how long they would learn, it would be impossible for them to adjust their categories accordingly. Luckily, however, color naming plays only a marginal role in everyday life, situations that require accurate communications of color are extremely rare. Many color-deficient individuals are not even aware about their condition until their first color-vision test. Thus, language, although important in establishing a consistent lexicon for our color categories, is too weak a link to influence perceptual categories.

ACKNOWLEDGMENT

I thank Rainer Hertel for inspiring discussions and critical reading of the manuscript.

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

Categorization in artificial agents: Guidance on empirical research?

William S.-Y. Wang and Tao Gong

Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, China. **wsywang@ee.cuhk.edu.hk**
http://www.ee.cuhk.edu.hk/~wsywang/ **tgong@ee.cuhk.edu.hk**

Abstract: By comparing mechanisms in nativism, empiricism, and culturalism, the target article by Steels & Belpaeme (S&B) emphasizes the influence of communicational constraint on sharing color categories. Our commentary suggests deeper considerations of some of their claims, and discusses some modifications that may help in the study of communicational constraints in both humans and robots.

The article by Steels & Belpaeme (S&B)presents a multiagent model that adopts many prevalent mechanisms used in other similar cognitive or self-organizing models, such as neural networks (e.g., Munroe & Cangelosi 2002), associative networks (e.g., Smith et al. 2003), and strength-based competition (e.g., Steels et al. 2002). The authors refrain from any judgment on mechanisms that might be more realistic. However, further discussions are required to assess some of their conclusions.

Their article summarizes the categorical repertoire-sharing process by using four types of simulations: (a) acquisition of repertoires with the same learning mechanism (individual learning), (b) individual learning and adjustment of acquired repertoires during language communication (cultural transmission), (c) genetic transmission of repertoires with occasional mutation (genetic evolution), and (d) genetic evolution and cultural transmission. The comparisons of Category Variance (CV) of (a) and (b), as well as (a) and (c) lead to the compelling conclusion that "both a cultural learning hypothesis . . . and a genetic evolution hypothesis . . . could explain how agents in a population can reach a shared repertoire of categories. . . . The difference between the two models appears to be in terms of the time needed to adapt to the environment or reach coherence" (sect. 5). Then, the authors suggest "the collective choice of a shared repertoire must integrate multiple constraints, including constraints coming from communication" (Abstract). However, deeper discussions of these claims are necessary.

First, in their model, the rate of genetic evolution is controlled by adjusting the parameters in the neural network. The rate of cultural transmission is determined by a different set of parameters that associate categories and their symbols. Although there is a general consensus that cultural transmission operates at a much higher rate, it is not clear how the two sets of parameters can be made commensurate with each other and meaningfully compared.

Second, to support the authors' suggestion, is it necessary to show why we must integrate cultural transmission, because genetic evolution alone can already achieve category sharing? Can cultural transmission influence genetic evolution, and if so, what is the influence? The answers to these questions lie in the comparison of the CV difference between (c) and (d), or between (b) and (d). In fact, this topic is touched on by Munroe and Cangelosi (2002) in their mushroom-foraging model (M & C model). Based on the *Semiotic Square* (Steels 2002, see Fig. 1), in the M & C model, genetic evolution adjusts the sensorimotor tools (neural network's connection weights, Sensation aspect, aspect A), and cultural evolution introduces changes to the outputs of neural networks in the previous generation, the combination of input times being the connection weights (Representation aspect, aspect B). The M & C model shows that cultural transmission can assist genetic evolution; the learning time under cultural transmission and genetic evolution is much shorter than that under only one of these mechanisms. However, in the M & C model, both mechanisms work on the internal aspects (aspects A and B), and it neglects the Symbol aspect (aspect D). The framework of the target article covers all four aspects of the semiotic square. The neural network handles the color representation, and genetic evolution

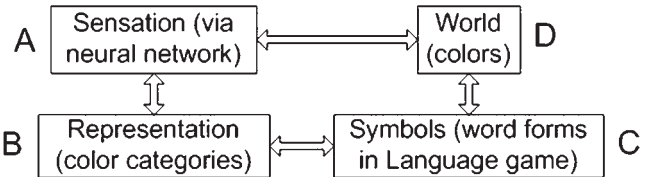


Figure 1 (Wang & Gong). Semiotic Square modified from Steels (2002).

adjusts these representations; the associative network handles the mappings between semantics and symbols; and cultural transmission adjusts these mappings. Therefore, besides demonstrating that both cultural transmission and genetic evolution can achieve the sharing task, this model can also explore whether these two mechanisms, separately working on different aspects, can affect each other by comparing CV or the number of games necessary to acquire certain CV in different simulations.

Finally, S&B should state clearly what "sufficiently shared" means in their claim that "a perceptually grounded categorical repertoire can become sufficiently shared among the members of a population to allow successful communication" (sect. 6). In their case study, because of identical learning mechanisms and limited features considered in creating categories, the categories created by different agents for the same color may be very similar, and successful communication may require the sharing of identical categories. It implies that only by sharing identical categories can successful communication be possible.

However, in general, this is not the case. Considering heterogeneous sensorimotor systems or learning mechanisms adopted by agents and the multiple features contained in world items, it is possible for agents, through different learning mechanisms, to create different categories for the same world item based on its different features. Besides, if both the categories partitioned in semantics inside one agent and the word forms partitioned in symbols can distinguish world items, it is possible that each agent will develop its own associative network between word forms and its categories, and successful communication is still possible even though there are no shared categories. This is more obvious when humans perceive abstract concepts like "friend," "loyalty," "game," and so forth. Different criteria are developed to represent these concepts, and communication is still available on a certain level.

In addition, some of their methods need modification if the authors want their results to be "relevant to a much broader audience of cognitive scientists" (sect. 6). First, the language game (Steels 2001a) in this model adjusts the association between words and categories, and the association between symbols and world items. No matter whether or not it can represent the speaker's word form, the hearer always gets the association between the symbol and the world item that this symbol represents in the speaker's mind. However, this method, similar to mind-reading, is too strong to be realistic, because in actual conversations there are communications in which the hearer gets no hints or even gets wrong ones. This indicates that language, or other communicational constraints, is not always reliable. Besides, as Quine's (1960) question about *gavagai* shows, nonlinguistic feedback only provides limited confirmation. Therefore, even without noise, misunderstanding is inevitable, and mind-reading does not simulate the actual influence of communicational constraints. Whether or not communicational constraints still have similar effects on sharing categories when occasional misunderstanding is allowed is worth studying, and this is already discussed in some models (e.g., Gong et al. 2004).

Also, this model adopts a Genetic Algorithm (GA) (Holland 1995) without crossover, in which, mutations, "happen with a probability inversely proportional to discriminatory success" (sect. 3.4). This method will undoubtedly accelerate the acquisition of

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

common categories because categories that are not successfully used will undergo more mutations. Therefore, this GA introduces a selective force though the mutation itself has no intelligence about what is good change. Genetic operations, like mutation, should be independent of certain factors outside the genome. Besides, the main driving force for evolution is the reorganization of the available materials (crossover), instead of the occasional mutation (Holland 2005). However, in this model, asexual reproduction does not incorporate crossover, and the low mutation rate may not explicitly represent the speed of the genetic evolution.

ACKNOWLEDGMENTS

The authors of this commentary would like to thank James W. Minett and Wong Chun-Kit for their useful discussions and resourceful suggestions. Our work is supported in part by grants from the Research Grant Council of the Hong Kong SAR: CUHK-1224/02H and CUHK-1127/04H.

Variations in color naming within and across populations

Michael A. Webster[a] and Paul Kay[b]

[a]*Department of Psychology, University of Nevada – Reno, Reno, NV 89557;* [b]*International Computer Science Institute and University of California – Berkeley, Berkeley, CA 94704.* **mwebster@unr.nevada.edu kay@icsi.berkeley.edu http://www.icsi.berkeley.edu/~kay/**

Abstract: The simulations of Steels & Belpaeme (S&B) suggest that communication could lead to color categories that are closely shared within a language and potentially diverge across languages. We argue that this is opposite of the patterns that are actually observed in empirical studies of color naming. Focal color choices more often exhibit strong concordance across languages while also showing pronounced variability within any language.

Steels & Belpaeme (S&B) use theoretical simulations to explore the potential role of physiological, environmental, and cultural (linguistic) constraints on the acquisition of shared color categories. Although their stated aim is to identify principles that could guide the design of communication among artificial intelligence systems, they emphasize that the results are also relevant for understanding color categorization in human observers. Our commentary focuses on the extent to which the trends they observe are evident in actual studies of color naming.

In S&B's simulations, whether or not a factor provides a loose or tight constraint is evaluated by measuring the variance in color categories across observers. In all cases, they find the variance to be greater for agents drawn from separate populations than for those drawn from the same population, yet this difference becomes dramatic when the categories are learned through language, in which case, the within-group variance approaches zero. This, in theory, points to a strong potential for cultural relativity in color naming.

What are the patterns of variance in empirical measures of color naming? There are two striking patterns. First, there are strong universal tendencies across languages. These tendencies were originally suggested by Berlin and Kay (1969) and have been confirmed by Kay and Regier (2003) in a recent analysis of the World Color Survey (WCS), which provides color-naming responses from an average of 24 primarily monolingual speakers from each of 110 unwritten languages. Specifically, they showed that the centroids of color-naming responses for different languages exhibit much stronger clustering than would be predicted by chance. This is qualitatively consistent with S&B's analyses, showing that physiological and/or environmental constraints can support some degree of consistency among speakers. Whether it is quantitatively consistent could potentially be evaluated by applying the authors' variance metric to the WCS data (which is available on-line at http://www.icsi.berkeley.edu/wcs/data.html). This might allow one to assess whether different languages show more concordance in color categories than would be expected from their models of physiological and environmental factors. Without such comparisons, it is difficult to interpret the relevance for human behavior of the values they derive from simulations.

The second prominent property of actual color-naming data is the pronounced variation among speakers of the same language. Individual differences in unique hue and focal color choices have been widely documented, though their causes remain poorly understood (Webster et al. 2000). For example, the wavelengths that individuals select for unique green within a linguistically homogeneous group span a range of more than 80 nm; these variations are in fact so large that the same wavelength might be chosen as unique green by one observer and unique yellow or blue by another (Kuehni 2004). Individual differences in focal color choices remain large for more naturalistic spectra like the Munsell chips and represent another obvious feature of the WCS data (as well as for most other data sets on color naming). Moreover, comparable differences persist even when the samples are restricted to individuals who select colors with the highest reliability (Webster et al. 2000). In sum, in actual measures of color naming, as contrasted to simulations, within-group variance is very large.

This fact appears difficult to reconcile with the minimal variance predicted by S&B to arise from adding communication to the simulated agents. Actual agents do not show the close agreement that language could potentially support. As an illustration of this, Table 1 compares the average within-language variance to the variance in mean foci across languages for "red," "green," "blue," or "yellow" terms for the WCS respondents, based on an analysis by Webster and Kay (in press). (These are calculated from the raw distances in the Munsell palette for the Hue and Value dimensions separately.) For each language, terms corresponding to the English terms were determined by finding the focal choices for con-

Table 1 (Webster & Kay). *Average variance in individual foci within a WCS language compared to the variance of mean foci between languages, computed for the hue or lightness of "red," "green," "blue," or "yellow."* F*-tests compare the between-language variance to the variance predicted by randomly sampling speakers of different languages. The hue scale runs from 1 = Munsell 2.5R, in 40 steps, to 40 = Munsell 10RP. The lightness scale is Munsell Value*

		Focal Hue					Focal Lightness				
Term	#	Mean	Variance	Predicted variance	F	*p*	Mean	Variance	Predicted variance	F	*p*
red	103	1.77	.46	.25	1.81	<.002	4.25	.095	.040	2.41	< e-5
green	73	18.9	3.01	.96	3.16	< e-8	4.74	.41	.099	4.12	< e-10
blue	50	27.7	2.45	.93	2.56	< e-5	4.30	.46	.093	4.84	< e-10
yellow	86	9.46	.65	.31	2.13	<.0002	7.79	.13	.038	3.38	< e-8

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

sensus terms closest to the English foci. For "red" and "yellow" these correspondences are obvious. For example, the means for the "red" and "yellow" clusters are separated by approximately ten times the cluster standard deviations, with only one language exhibiting a consensus term nearer to the intermediate focal point for English "orange." This finding echoes the consistent clustering demonstrated by Kay and Regier (2003). For green and blue the clustering is less obvious because many of the WCS languages apply a common "grue" term to this region. The values shown are thus restricted to the subset of languages that have both terms. Mean foci across languages vary much less than individual foci within languages. This suggests that a common language imposes only a weak constraint, and a difference in language produces relatively little divergence.

Another finding that may argue against a strong constraint of language on human color categories is that individual differences in focal choices for binary hues (e.g., blue-green or yellow-green) are not obviously distinct from the differences measured for primary hues (e.g., blue or yellow or green). Malkoc et al. (2002) in fact found less variation in "focal" blue-green than in the unique hues blue and green. That is, English speakers were more consistent at selecting the boundary between blue and green than at choosing either primary category's best example, even though there is no basic term targeting this boundary, and, as noted, many other languages do not have separate words for these categories.

Admittedly, the loci of color categories do vary significantly across different language groups, and there are both extreme (Davidoff et al. 1999) as well as more subtle examples of these differences (Webster et al. 2002). For example, Webster and Kay analyzed whether differences between the average foci for the WCS languages were larger than predicted by random sampling across languages; and as Table 1 shows, differences were significant for all terms. Nevertheless, as indicated earlier, the within language variances are much higher than the variances in mean foci between languages. As S&B show, interlanguage differences can arise from many sources. The question remains as to exactly what degree the existence of different languages is an actual contributing factor to the total interpersonal variation in color naming.

ACKNOWLEDGMENT

This work was supported by grants EY-10834 and NSF-0130420.

In the tiniest house of time: Parametric constraints in evolutionary models of symbolization

Chris Westbury and Geoff Hollis

Department of Psychology, University of Alberta, Edmonton T6G 2E9, Alberta, Canada. **chrisw@ualberta.ca**
http://www.ualberta.ca/~chrisw/

Abstract: Steels & Belpaeme (S&B) describe the role of genetic evolution in linguistic category sharing among a population of agents. We consider their methodology and conclude that, although it is plausible that genetic evolution is sufficient for such tasks, there is a bias in the presented work for such a conclusion to be reached. We suggest ways to eliminate this bias and make the model more convincingly relevant to the cognitive sciences.

> "When you really look for me, you will see me instantly – you will find me in the tiniest house of time."
> —Kabir (1440–1518)

We are sympathetic to computational models of language and symbolization, and believe many questions about the structure of symbols will only be answered by such modeling. We also agree that cultural dynamics must be valuable for grounding shared categories within populations. Our commentary focuses on some of the theoretical implications of Steels & Belpaeme's (S&B's) discussion on language and category sharing.

From a linguistic point of view, S&B's model is of course highly simplified, in keeping with their stated pragmatic goals. We appreciate that an attempt has been made, as the authors note, to keep the model "grounded, embodied, situated, and cultural." However, such uncontroversial foundational qualities of language as representation of second-order relations, word-order constraints, combinatorality, and traditional transmission (to name only a few) are not possible in the present model. This makes it dubious as a model of language per se. Nevertheless, such a simplified model may shed light on how symbols are grounded through situated interaction, and thereby shed light on what one might call the "prelinguistic stage" of language evolution. We take this to be the goal of the present exercise. For this reason, we do not wish to dwell on the model's insufficiencies with respect to language. Instead, our criticisms focus on the model's evolutionary parameters.

S&B provide a broad discussion that captures key points on how individual learning, genetic evolution, and cultural constraints tie into language development within a population. However, with such breadth, it becomes difficult to bring out the implications of each individual topic. We would like to bring up some points in regard to the methodology in this study, which suggest that the conclusions presented in the current article are made prematurely, based on the data presented. We have three main criticisms. The first has to do with the way mutation is performed during reproduction. The second has to do with the small population size. The third has to do with the immortality of organisms.

Mutation rates. Reproduction in the evolutionary computation paradigm, genetic programming, is generally a destructive process; offspring are likely to degrade in fitness with respect to their parents (Nordin & Banzhaf 1995; Streeter 2003). We assume this, or some similar property, is the motivation for reducing a new population member's mutation chance as a function of the parent's fitness. Such a decision is useful from an engineering standpoint, as it affects the trajectory of the evolutionary process; fit population members will be more likely to have offspring with similar categories to themselves, than they would without such a feature. However, in the current context, this means there is a bias towards evolving populations where color categories are constant across population members. It is plausible that individual agents might evolve mechanisms that protect their children from the destructive forces of reproduction (e.g., Soule & Foster 1998); however, building it directly into the evolutionary process seems out of place, when the question is if evolution will do this on its own.

Population size. Although the authors mention that their technique "scales up," we believe that using a small population affects the way their problem is solved in important ways. When using a small population to solve a large problem, initial populations cannot include elements from many different optima in the solution space. If the chances of finding a population member in a good path to an optimum are low to begin with, the first path discovered will probably be the only path explored. In Figures 6, 7, and 15 of the S&B article, no population member has any discriminative success until around the fifth generation; this is also true for communicative success, as evidenced by Figure 15. This suggests that the first population member to have any success at a task probably directs the future evolution of the population. We find it plausible that evolution can ground categories in an entire population, but it would seem premature to conclude this from the present results because the small population is able to explore only one optimum in the fitness landscape. The ability of evolution to ground categories in a population of agents needs to be tested in an environment where population members from many different optima can compete against each other in the same population pool.

Immortality. It is a common practice for evolutionary computation paradigms to allow parents and children to coexist indefinitely (e.g., Fogel 2002; Koza 1992; Westbury et al. 2003). Keeping par-

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

ents in the population pool reduces the chances that valuable information will be lost from generation to generation. However, such a property seems out of place in the current context for two reasons. First, organisms naturally die over time. Second, because of the lack of change that comes along with keeping population members across generations, there is a bias towards populations that evolve (perhaps just faster) to a set of shared categories. In any domain where there is more than one distinct solution to a problem and there is a claim that evolution will home in on only one of those solutions, convergence rates must be treated with special care. The results seen may depend crucially on how fast a population converges on a solution. To increase both the organic validity and the generality of the experiments conducted concerning genetic evolution in this article, we suggest putting an age limit on population members. Such a constraint would paint a much clearer picture of genetic evolution's role in category sharing.

S&B have achieved their pragamtic goal, by showing that evolution is sufficient for grounding category sharing with at least one set of parameters. However, the combined effect of those parameters is a large bias towards converging on one solution, which brings into question how useful their model is for understanding natural symbol grounding. It would be interesting to know if the results presented by Steels & Belpaeme can be recreated with less favorable parameter settings.

The question of the assumed givenness of the singularity of the target

Edmond Wright

Cambridge CB4 1DU, United Kingdom. **elw33@hermes.cam.ac.uk**
http://www.cus.cam.ac.uk/~elw33

Abstract: Interesting as the experiments are, their relevance to the real-life situation is rendered questionable by the unthinking use of given singularities as target objects. The evolutionary process does not respect what one agent *takes to be* a singular referent. A "singling" from the continuum is rather a varying feature of the necessity to track what is rewarding in it.

Although they are remarkably even-handed in their comparisons of what they call the "nativist," "empiricist," and "culturalist" approaches, claiming that they wish to leave the debate for later resolution (sect. 1.2), Steels & Belpaeme (S&B) make a presupposition about objectivity that produces some distortion of their experimental stance. For example, in considering the application of a colour category in the actual world (e.g., to mushrooms edible and nonedible, sect. 2.2), they assume a binary perfection *that is exactly the same for all agents* to be in existence before the categorization. It is interesting that Stevan Harnad used the very example of mushrooms to show the impossibility of accepting the idealizations of language. He described this in his amusing account of how he happily trusted a Russian student, when in autumn he went hunting for mushrooms near Moscow, but was deeply suspicious when the same student showed enthusiasm for hunting for the same mushrooms in New Jersey (Harnad 1990, personal communication). Harnad has emphasized the provisionality of categories because of individual variation. S&B do acknowledge the failure of categorical equivalence across populations (sect. 3.2), and Harnad's example falls within that characterization, but on the question of singular reference they do not get beyond Quine's "gavagai/rabbit" (sect. 4.1), and Quine's own error is to assume that a logically singular entity is there in the Real before human selection. Even though that singularity can be differently characterized, for Quine it still retains a given singularity (Quine 1960, pp. 20–46).

My argument has always been that singularity is a pragmatic but not an ontological necessity for real-life communication (Wright 1992; 2005). It is pragmatically necessary for two agents in communication to treat a puzzling region of the real *as if* it is singular, but that is merely to get their *differing* perspectives into some kind of harness so that correction of one by the other can go through. S&B concede relativity for agent to agent on the sensory level, but that obviously implies that a perfectly logical coinciding of their percepts can never be wholly achieved. The obvious inference is that pragmatic success, in what S&B call "collective decisions" (sect. 1.2), even over considerable time, does not guarantee that a perfectly singular objective referent preexists in the Real. Instead, a number of roughly coinciding referents exist at the focus of social interest, but they are to be *treated as* one, so that communication about that doubtful region can be accomplished. Indeed, evolution, which forms a key element in their thesis, can be viewed as a means whereby *continuous changes in the Real that provide reward* can be *tracked* by alteration of categorization across and within individuals, and not a tracking of preexisting singularities common to all. To believe in the givenness of singularities is an evolutionary recipe for disaster. Communication can often be concerned with a correction about singularity itself. To quote a sixth-century Indian Buddhist philosopher, "even 'this' can be a case of mistaken identity" (Dinnaga, see Matilal 1986, p. 332). So the investigation of the question of what S&B repeatedly call "adequacy" of response in real-life situations cannot be pursued if one begins with given countable distinctions, be they Munsell chips or anything else. A species that determinedly stuck to an achieved singularity in the face of changes in the Real would not survive, which is certainly an "inadequacy" of response.

One cannot neglect the "iconic," that is, the sensory registration, which is different for each individual (Harnad 1987, p. 550), nor the differences in learning histories resulting in differing criteria from person to person for what they *take as* the "same" object (Rommetveit 1978, p. 31), because it is these that allow for correction of an agent by another agent in a new circumstance. It also reveals why the chicken-and-egg problem does not arise for human beings (sect. 1.2) because the fictive *assumption* of perfect coreference needed for a roughly common focus to be arrived at seamlessly allows for its apparently paradoxical correction by another.

Where S&B consider real-life situations, they certainly make it clear that the "shared" convergence of categorical repertoires does not achieve perfect superimposition of discrimination (sect. 4.3). They make the comment that "different meanings may coexist until a situation arises that disentangles them" (sect. 4.2, Fig. 11), but, although the new situation can produce new discriminations, it cannot finally produce logical perfection across the population in what is being individually "singled."

Of course, what is left out of the artificial experiments is the *motivation* of the agents. It is that which prompts the agent to do the "singling" in the first place, for there is no knowledge or action without intention. So, regarding artificial experiments, not until the design of robots or computers has reached the inclusion of motivation (i.e., the effects of pleasure and pain modules) can the effects of "singling" in categorization be satisfactorily reproduced or its changes evolve across a population. S&B do conclude that "[i]t would be risky to rely on embodiment constraints and statistical clustering for forming the repertoire of perceptually grounded categories for use in communication" (sect. 5.1). They also insist in their conclusion that "there are important degrees of freedom left" for the individual in category formation, and it is those degrees of freedom that allow for the updating of one agent by another when the external Real produces its unexpected challenges.

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

What is culture made of?

Chen Yu and Linda Smith

Department of Psychology and Cognitive Science Program, Indiana University, Bloomington, IN 47405. **chenyu@indiana.edu smith4@indiana.edu http://www.indiana.edu/~dll/ http://www.indiana.edu/~cogdev/**

Abstract: Culture is surely important in human learning. But the relation between culture and psychological mechanism needs clarification in three areas: (1) All learning takes place in real time and through real-time mechanisms; (2) Social correlations are just a kind of learnable correlations; and (3) The proper frame of reference for cognitive theories is the perspective of the learner.

The target article argues that culture – sharing meanings – plays a special role in the creation of concepts and that this special role operates in addition to "empiricist" contributions to concepts. In brief, the take-home message is that empiricist approaches are not enough and learning in a socially guided culture is special. We agree fervently with the latter point and reject the former. Culture is just a broad – and deeply relevant – set of multimodal correlations. The moment-to-moment interactions of a learner in the world include things in the environment, the labels attached to those things, and interactions with other social beings. All these regularities are grist for the empiricist mill.

All learning, all development takes place in real time. Babies learn language from scratch, through millisecond by millisecond, second by second, minute by minute changes that arise from their own sensorimotor interactions with the world. Any theory of learning has to start here in real time in the regularities that accrue over those repeated interactions. It has been suggested that these regularities include correlations in the audio stream sufficient to create word-like units (Saffran et al.1996). These repeated interactions will also yield other regularities – those resulting from the physical structure of the world, or the time-locked multimodal dependencies arising from looking, seeing, touching and feeling, as well as those that arise from the actions of others. Mothers interact with their infants using hand signals, touching, eye gaze, and intonation, and all these cooccur in real time, with words, with objects, and with the infants' own actions and internal states. These patterns of social interactions surely reflect culture in that the mother's interactions are a product of her own developmental history and the social world in which that history was embedded. But from the developing infant's point of view, from the perspective of the mechanisms internal to the infant, it does not matter whether those correlations are the products of physics, human biology, or a history of learning in a social world. They are all learnable correlations.

Social correlations are just correlations, but ones that critically amplify other correlations. Recent studies in human development and machine intelligence show that the world and social signals encoded in multiple modalities play a vital role in language learning. For example, young children are highly sensitive to correlations among words and the physical properties of the world (Colunga & Smith 2005; Smith et al. 1996). They are also sensitive to social cues and are able to use them in ways that suggest an understanding of speaker's intent (Baldwin 1991; Tomasello 1992). The main point of Steels & Belpaeme's (S&B's) practical perspective is that these social cues dramatically enhance learning about relations between the physical world and language. This is an important point and the main contribution. But, this need not reflect a special link between concepts and culture, at least in a mechanistic sense. Social information can only be made manifest in correlations that arise from the physical embodiment of the mature partner (the mother) and the immature partner (the baby) in real time. For example, the mother "jiggles" an object, the infant looks, and simultaneously the mother provides the name. These time-locked social correlations play two roles. First, they add multimodal correlations that enhance and select some physical correlations, thereby making them more salient and thus learnable. Second, these time-locked and coupled interactions are, from the perspective of an observer (though perhaps not from the perspective of the infant), shared meaning. The viability of these ideas has been shown in a program of simulation and empirical studies presented by Yu et al. (in press) and Yu and Ballard (2004), demonstrating that body movements play a crucial role in creating correlations between words and world, correlations that yield world-word mappings on the baby's part that match those intended by the speaker. The simulations show that the coupled world-word maps between the speaker and the baby – what some might call the baby's ability to infer the referential intent of the speaker – are made from simple associations in real time and the accrued results over time of learning those statistics. Critically, these statistics yield the coupled world-word maps only when they include body movements such as direction of eye gaze and pointing. The power of these correlations results from this coupling of social partners that enhances and selects the right correlations (see Thelen & Smith 1994). But notice the mechanism: It is correlational learning all the same.

Culture and the frame of reference problem. One might call a learning system built on such social cues "culture," but from the perspective of the learner, it is just correlations. Culture is what we, as observers, see as a system of correlations, a system that perpetuates itself. However, from the perspective of a baby learning language, culture is not a separate and special source of information, but is one grounded in a sea of correlations. Importantly, these correlations form a complex and dynamic system. Moreover, the correlations are not passive statistical regularities independent of the learner. Rather, the learner is an active creator of the correlations in three senses: (1) The bodily orientation of the child's sensory system (e.g., where one looks) determines what will be learned; (2) the very activity of the sensorimotor system adds correlations to the mix; and (3) the child's activity elicits and is coupled to active social partners, creating even more dynamically complex and multimodal correlations. There is no culture separate from all these correlations in the child's point of view.

The target article makes an important contribution by pointing out the role of culture. But the implemented simulation may be too simple. We suggest that when an embodied agent has real-time experiences in the physical environment, rich social information in the environment can be acquired from multisensory correlations. These lead to coupled world-word maps between social partners. These coupled maps, in our view, embody what is meant by the inference of the internal state of another and also are absolutely crucial to language learning and human cognition more generally.

Authors' Response

The semiotic dynamics of colour

Luc Steels[a,b] and Tony Belpaeme[c]

[a]*Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Pleinlaan 2 – 1050 Brussels;* [b]*SONY Computer Science Laboratory, 75005 Paris, France;* [c]*School of Computing, Communications and Electronics, University of Plymouth, Plymouth PL4 8AA, United Kingdom.* **steels@arti.vub.ac.be tony.belpaeme@plymouth.ac.uk http://www.tech.plym.ac.uk/SOCCE/staff/TonyBelpaeme**

Abstract: The interesting and deep commentaries on our target article reflect the continued high interest in the problem of colour categorisation and naming. Clearly, colour remains for many cognitive science related disciplines a fascinating microworld in which some of the most fundamental issues for cognition and culture can be studied. Although our target article took the stance of practically oriented engineers who are trying to find the best solution for orchestrating the self-organisation of communication

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

systems in artificial agents, most commentators focus on the implications for cognitive science and we will do the same in our reply.

R1. Semiotic dynamics

In our target article, we have adopted a relativistic and complex systems viewpoint towards the origin and nature of colour categories and their names. By a complex systems viewpoint, we mean that we do not seek an explanation through the nature or functioning of a single specific aspect of the overall process, such as the statistical structure of the world, the nature of the human sensory apparatus, the conceptual space generated by early visual processing (implemented in our simulations as the CIE *L*a*b** space), the presence of genetically predetermined categories, the nature of the neural networks engaged in categorisation or naming, or the social shaping through language. Instead, we investigate how the dynamical interaction between *all* these various aspects, operating within individuals and, most importantly, across individuals in a population, can collectively give rise to a set of colour categories and names that are sufficiently shared to make successful communication possible.

Several commentators requested a clarification of what we mean by "sufficiently shared." We definitely do not mean that the categories have to be absolutely identical. On the contrary, in view of the individual variation in the data sets available for learning, the individual variation of the perceptual apparatus (e.g., dichromats vs. tetrachromats), individual variation in the history of interactions with the world, and cultural variation as reflected in language or cultural habits, we cannot expect absolute similarity. Many commentators provide further evidence that this variation is even larger than we originally thought. We have adopted therefore a relativistic view on colour, in the sense that individuals each generate their own ways of perceiving, categorising, and naming colours, but there is a mechanism, based on coupling the behaviours of different individuals, by which individual differences get resolved, even if only locally and temporarily for the purposes of a single conversation. The target article therefore advocates a shift from investigating the mechanism or mechanisms that establish the "optimal," "universally shared" set of colour categories (whether it is through biology, statistical learning, or genetic evolution) towards investigating the mechanisms through which a group coordinates their perception, categorisation, and naming conventions, despite individual differences and despite the absence of a central controller or telepathy. By "sufficiently shared" we therefore mean "sufficiently coordinated" to achieve successful communication and communal use, and this depends very much on the environment and ecology in which the group operates, that is, which distinctions are relevant in their ongoing interactions.

We have proposed a single key mechanism to achieve coordination, namely, that there must be a mutual coupling between the various processes in the total chain. For example, the low-level signal processing algorithms must get feedback on whether successful discrimination for language was possible so that they can be enhanced or reshaped if needed; the category formation process must get feedback on whether a category was successful in the language game so that this influence can be used to reshape the category; and the lexicon formation process must get feedback on whether the game succeeded so that the strength of associations between meanings and labels can be adjusted. In other words, there is not just an upstream flow of information from perception to naming, but also a downstream flow so that all processes can get coordinated for each individual and across individuals in the population. The target article has specifically focused on the coupling between the category acquisition process and the naming subsystem, showing how the activity of naming is able to relatively quickly coordinate the categorical repertoire of different agents so as to allow successful communication. It also showed that the coupling can either go through genetic evolution implementing a "memetic drive" (as pointed out in **Blackmore**'s commentary) or through cultural evolution by a direct coupling between individual category formation and naming. The latter is argued to be the fastest and most efficient way to reach a coordinated categorical repertoire.

It is, in principle, possible to study this semiotic dynamics in human populations, and some psycholinguistic studies have indeed tried to track the rapid cognitive and linguistic alignment that appears to take place when humans engage in communication (Garrod & Anderson 1987). The target article instead takes a theoretical stance and investigates what kind of dynamical relationships among all the aspects of the overall process (world, perception, categorisation, naming) are necessary and sufficient for a set of agents to arrive at a successful communication system. Our own motivation is pragmatic, because we try to find out how we can best build artificial agents that can self-organise their own communication systems. Moreover, it is impossible to integrate with total realism all relevant aspects of the real world, human ecology, human physiology, brain science, or human culture into theoretical models. So we are not trying to explain through these simulations why a particular human population might have adopted a particular categorisation or lexicon, or why there are specific universal tendencies in human colour categorisation, rather, we try to put just enough complexity into our simulations so that we can study the overall semiotic dynamics underlying social cognition.

Based on our concrete proposals of the dynamical interaction between categorisation and naming processes in individuals and among individuals, and on computer simulations testing their effectiveness, we derive two types of conclusions. The first type shows that the individual and collective dynamics we have introduced are indeed capable of leading to a coordinated set of colour categories and names adequate for successful communication in a particular environment. For example, Figure 15 shows that even a simple model of genetic evolution of colour categories coupled with a lexicon formation process based on a bidirectional associative memory and reinforcement learning allows a population to derive a successful communication system. Figure 12 shows a similar result for an entirely cultural evolution of colour categories.

The second type of conclusion argues that the semiotic dynamics we advocate is not only sufficient but also necessary, in other words, that relying exclusively on one source of constraints is insufficient to explain how a group of situated grounded agents arrives at an effective communication system that is adapted to their environment and ecology. For example, Figure 20 shows that relying only on the statistical structure of the world does not allow agents to build up a sufficiently shared categorical repertoire to allow suc-

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

cessful communication. Figure 7 shows that a population is able to reach a shared set of categories based on gene propagation, but that this process is slow to adapt to changes in ecology or environment, hence relying only on genetically innate categories makes it more difficult for a population to adapt to change.

Given this brief summary, we can now survey the different commentaries to this argument. They fall into three groups: (1) those that provide additional support for these two types of conclusions, (2) those that argue that they are not justified, and (3) those that provide suggestions for different experiments or enhancements of various kinds.

R2. Supportive commentaries

Commentaries that provide further justification for a relativistic, complex systems approach introduce either more data showing actual variation in colour categorisation and naming among human groups, or additional evidence why a particular source of constraints cannot be the sole determinant of a coordinated categorical repertoire in a population, and hence why all constraints have to operate together.

Thus **Davidoff & Luzzatti** argue in favour of the central thesis of the target article based on psychological and anthropological investigations of colour categorisation and naming. They take an even stronger stance than we do with respect to the sufficiency argument, namely, they argue that colour categorisation cannot even properly form without a labelling process stimulating it. They provide two sorts of empirical evidence, the first, from cross-cultural studies that show beyond doubt how the use of language influences categorisation, not only for language, but also for other cognitive tasks like memory or sorting tasks. Even more fascinating is the evidence reported on anomic aphasics, which suggests that without the activity of labelling, colour categorisation cannot function or get off the ground. It is true that in our simulations the discrimination game is used to generate colour distinctions and so produces a repertoire of categories even if language is not involved, contradicting this evidence. However, in a more realistic setting the discrimination game would only arise and be stimulated if it is part of a larger task, such as the guessing game. We wonder nevertheless whether other kinds of tasks (such as food selection) could also not be a stimulus for the discrimination game and in what way those patients would perform in such a task.

In the same line, **Roberson & O'Hanlon** point to additional empirical studies that demonstrate how the communicative needs of a community help to shape the specific categorical repertoires of its members, and that these differences not only show up in communication but also in other cognitive tasks. They paint a much more complex picture of natural colour categorisation, particularly in nonindustrial cultures, with very few hue-related colour terms and many concrete exemplars, as opposed to a few focal points. They emphasise that the similarity of stimuli plays an important role in shaping repertoires, for example, forcing categories to be continuous in the colour domain as opposed to disjoint. This similarity is implicitly embedded in the radial basis function network used in the target article, because the network carries out a nearest neighbour comparison and therefore groups colour experiences by similarity around a prototype.

The commentary by **Bimler** points out that "not all observers experience identical color distributions" and that there are significant "variations in color distribution among human habitats." Both of these factors make it obviously even more difficult to assume that the statistical structure of the environment is enough to make all agents converge on shared colour categories. Bimler also points out that the sensory apparatus and low-level visual processing in humans is highly varied, particularly if colour deficiencies are taken into account, consequently the conceptual space would show variation as well. Again, this is further evidence that this source of constraints cannot be the only force that pushes agents towards a coordinated categorical repertoire.

A similar point is made by **Jameson**. She criticises our simulations because we have assumed that all agents use the same CIE standard model as conceptual space and the same perceptual process for deriving it. She argues that this already puts heavy constraints on the kinds of categories that will evolve, which is of course true. We agree with Jameson that to examine our stance more thoroughly, we should also introduce variation on this aspect of the overall process. The reason why no variation was used at the perceptual level is because (1) we believe that the mechanisms put in place would already allow agents to arrive at a set of coordinated categories, even if their perception or conceptual space is as different as shown in the empirical evidence pointed to by Jameson. The categorical networks employed by each agent in the population would of course look quite different and it would be very hard to compare them using the measures employed, but the success rate and speed of convergence would be comparable, unless of course, there are basic incompatibilities so that successful communication is only possible by stepping outside the colour domain. (2) We wanted to focus particularly on the interaction between categorisation and naming, because that is the most controversial issue; with some researchers explicitly denying that language can have influence on category formation. By keeping the embodiment and colour space constant, the impact of different choices with respect to category formation and the influence of language can be clearly brought out.

The commentaries by **Bimler** and **Jameson** suggest another experiment in which we introduce variation in the perceptual apparatus and low-level processing. Such an experiment was recently carried out by Joris Bleys, who compared the performance of learning colour categories and their associated terms for a homogeneous population (as used in the target article), as well as a heterogeneous population. In the heterogeneous population, each individual had a random variation on its colour perception, implemented as a normal variation (with a standard deviation of 10) on each of the L^*, a^*, and b^* dimensions. Figure R1 shows the average communicative success for five homogeneous populations and five heterogeneous populations, each population having 20 agents. The communicative success of both kinds of agents evolves in the same way, showing that perceptual variations have very little influence on the communication of colour. The category variance is different, as expected: 3.18 ± 0.41 for the homogeneous populations, 6.45 ± 0.51 for the heterogeneous populations, showing that the categories of the heterogeneous population are less similar. These experiments confirm that conceptual coordination does not depend on a unique source of constraints, but on the interaction between constraints

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

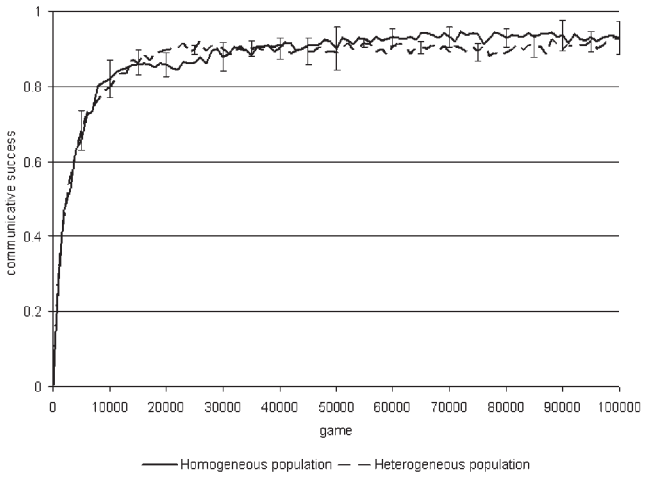


Figure R1. The average communicative success of five populations consisting of identical agents versus the average success of five populations consisting of agents having variations in their chromatic perception.

and the semiotic dynamics generated in the overall system. They confirm that agents can build a successful communication system even if their internal components are not all identical, they just need to be coordinated. These additional simulations also illustrate that, using the methodology adopted in the target article, it is straightforward to investigate additional questions or add additional constraints to bring in more realism.

Chella also discusses the issue of conceptual spaces. He correctly points out that our framework is quite general, in the sense that, given any kind of conceptual space, the mechanisms presented in the article show how it can become cut up into different regions. The conceptual space could be auditory, spatial, based on body position sensors, and so forth, as has been shown by Chella, **Ikegami**, and other roboticists referenced in their commentaries. Moreover, the dimensions need not be the direct input from the sensory channels, but could be processed in various ways, and thus they could also include relational information, that is, the dimensions of the space could reflect how adjacent or opposite two samples are with respect to some directly observed dimension (a question raised by **Bimler**).

But, in contrast with **Chella**, we do not believe that we have actually shown how the conceptual spaces themselves are learned. The novelty of our approach (with respect to other work on conceptual spaces) is to show how verbal interaction can help to shape the way the conceptual space is divided up into different categorical regions. As pointed out by **Bódog, Háden, Jakab & Palatinus** (**Bódog et al.**), we assume the *L*a*b** space as given and fixed in all the experiments (because we specifically wanted to study the impact of categorisation and naming). It is perfectly possible however to make this aspect undergo either genetic or cultural evolution (as suggested by these commentators). In a recent piece of work, Nicolas Neubauer focused precisely on this topic, showing in computer simulations how the brightness dimension could become a separate conceptual space (leading to categories like "light" and "dark") in addition to the hue-based colour space, so that multi-word sentence like "dark blue" could be formed. Neubauer again showed convincingly how language can play a crucial role in fixing the use of a new conceptual space in a population, thus providing further computational support to the complex systems view advocated in the target article (Neubauer 2003).

Another commentator who argues that we should make even more aspects of the overall process variable, is **Wright**. He rightfully claims that we introduce a strong bias in the very beginning of the perceptual process by assuming the notion of different individual samples, whereas in realistic circumstances even the notion of what counts as a single object should be the subject of negotiated pragmatic convention. We agree with Wright's position, and in other – more realistic – experiments we carried out on real world robots, this was indeed a major issue (Steels & Kaplan 2002b). Even while using the same segmentation algorithm, two robots quite often diverge on what the boundaries of objects are, simply because slightly different light conditions or angles of view give different segmentation results. We believe that it is entirely possible to create the same sort of coupled mutual feedback that we have now set up between the category formation process and the naming process, between the object identification process and other aspects of the total cognitive chain, so that verbal interaction could play a role in deciding what counts as a singular object.

Various authors make comparisons to animal categorisation and signaling, particularly with respect to birds. **Christensen & Tommasi** point to empirical research showing that birds are not only capable of colour categorisation in referential communication but also of associating spontaneous vocalisations with new colour distinctions. This research is fascinating and it shows that the semiotic dynamics discussed in our article might already be operating in certain animal species. Given the simplicity of the mechanisms we have used, this is not surprising, but we would welcome further investigations for other species. The fact that animals might be able to self-organise a communication system and coordinate their categorisation of reality, strikes us as a much more significant step than the ability of some animals to mimic an existing human-invented semiotic system.

Harter & Lu summarise very accurately our position that multiple constraints act on the shaping of communication systems, and they argue for a balanced view in which the different dynamical systems at the genetic, individual learning, and cultural level interact. A similar position is taken by **Lehky**, who suggests that the cultural influence on colour categories through language can be stronger for some categories and weaker for others. Lehky points out that there might be basic categories that are so crucial to survival that there is no time to learn them, and there are indeed many categories relevant in human life that are not or only sparingly lexicalised (the domain of olfaction is a good example). In these cases, coordination would have to be achieved through other means than verbal interaction. It is difficult to disagree with this stance. As mentioned, the target article develops rather extreme positions so that the conclusions of the model are clear, but generally speaking, we argue that multiple constraints are at work, and we show through our models that learning without language or genetic evolution without language can also lead to the formation of perceptually grounded categories.

Two supportive commentaries by **Ikegami** and by **Yu &**

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

Smith both emphasise the coupled dynamical systems viewpoint, as advocated in the target article. Both commentaries emphasise also the need for a much more active role of the learner in shaping the environment available for learning, which implies that the input itself would also be influenced by the semiotic dynamics of interacting agents, a point we entirely agree with, although it has not been realised in the presented simulations.

More specifically, **Ikegami** reports experiments with mobile agents whereby sharing categories means sharing the sensory-motor coordination relevant for a joined activity (such as one agent tracking another one). In his view, the kind of verbal interactions we use in the target article to coordinate categorisation is only one example of the more general social process in which interacting agents coordinate their conceptualisations of the world: The role of verbal interaction is to stimulate categorical refinement.

Yu & Smith interpret the notion of culture as just one of the forces acting on the developing child, but it operates by generating correlations that would not be generated otherwise and is therefore similar to other forces, such as the physics of the real world, which also generate correlations. Yu & Smith believe that correlational learning (as convincingly demonstrated in their own work) can be the main source of category formation and category coordination (see also the commentary by **Vogt & Smith** on cross-situational learning), but they insist, as we do, on the framework of coupled dynamical systems. We agree entirely that it would be desirable to have much richer, active agents in much richer environments with developmental time-lines that are long enough to allow correlational learning to become effective, but there are, at this moment, practical limits to the simulation experiments that can be performed realistically. As robotics further develops in the direction of complex humanoid robots, such more complex experiments will perhaps become more feasible, but they would still require extraordinary effort.

Finally, **Blackmore** correctly remarks that our simulations (particularly those reported in Fig. 14) show how memetic processes can steer the genetic assimilation of categories, something she has called memetic drive. The "memes" in this case are words for naming certain colour categories and they evolve in a purely cultural fashion. The success of using a word depends on the nature of the colour category it employs and therefore, if there is a strong coupling between communicative success and fitness, that colour category will be genetically reinforced. However, in contrast with **Blackmore**, we believe that this type of genetic assimilation of perceptually grounded categories is rather exceptional, given that (1) it makes the population no longer able to adapt quickly to change, (2) it makes it more difficult to explain the observed variation across populations and individuals, and (3) it makes it hard to explain why somebody born from parents foreign to a particular culture can nevertheless perfectly well pick up the colour distinctions of another culture. We believe that our work is nevertheless very relevant to memetic theory, because it shows how certain behaviours, in this case ways of categorising reality, can replicate without genetic evolution, more specifically, through the coordinating force of verbal interactions. Often memetic theorists assume that a particular form of behaviour can be copied (supposedly by imitation) from one individual to another, but fail to be precise in how this copying is carried out. It can definitely not be based on telepathy, and imitation has turned out to be a very difficult task for which no operational models exist today.

R3. Opposing views

The commentaries that question the complex systems view advocated in the target article take two forms. Some argue that the empirical data of human colour categorisation contradicts the trends seen in the simulations. Others argue that a particular source of constraints might still be sufficient, if only we would have made the sensory input data, the perceptual process, the category formation process, or the language game itself much richer and more realistic compared to humans. In addition, there are some commentators who consider our simulations too simplistic to be relevant for human psychology.

Counterevidence based on empirical data is provided by **Webster & Kay**. They point out (as we also did in our target article) that there is strong evidence for universal tendencies in colour categorisation, and that these tendencies can be partly explained by the environmental, ecological, physiological, and cognitive constraints operating on colour categorisation and naming. We see similar tendencies in the simulations, and it would indeed be extremely interesting to follow up on Webster & Kay's suggestion to apply the variance metric used in the target article to the WCS data. We have also argued that many more constraints would have to be put in (e.g., more realistic ecologically valid input) to achieve tendencies closer to human languages.

Next, **Webster & Kay** point out that there are, in fact, important individual differences between speakers of the same language, whereas our simulations show that the group strongly converges to a similar set of categories. We accept the findings reported by Webster & Kay. One might think that these derive from the nature of the psychological testing that was used during the World Color Survey (Kay et al. 2003), which does not involve a communicative setting in which speakers have to use colour terms to achieve a communicative act (differences pointed out also by **Vogt & Smith**) or that the individual differences are an artefact of trying to shoehorn WCS colour terms into English colour terms. However, we have done some empirical testing of our own by asking humans from the same language family to play guessing games of the same sort used in the target article, and also found evidence for a similar sort of individual variation within the population (Belpaeme 2002a).

This individual intralanguage variation is very puzzling. If the focus for unique green for one observer may be the focus of unique yellow or unique blue for another, then communication becomes highly fallible. Our response to this is that we should not just take a snapshot from a randomly sampled population, but look at repeated interactions in a group of individuals, for example, the members of a family, the children in a class, a team of architects, designers, and builders. We predict that then a kind of converging semiotic dynamics will be seen, similar to the one discussed in the target article. In other words, we have to narrow down the interpretation of the simulations reported in the article. Rather than talk about languages, we should talk about groups of speakers who coordinate locally their colour categorisations and names. Of course, if one is already a speaker of a given language, the focal points would normally be already much closer and convergence would be much faster. At this moment, empirical data is almost en-

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

tirely lacking on this short-term coordination of colour categories.

Hampton argues that the conclusions of the target article may not hold up anymore if a more realistic source of sensory data and a more varied set of conceptual domains are used. We agree that focusing on colour as an exclusive dimension, and not taking into account many of the other dimensions that make up a normal sensory experience, is not realistic. We also agree that in other domains, such as biological classes, the constraints imposed by the structure of the world may be more constraining than the colour domain. As mentioned earlier, our position is not that one source of constraint is not sufficient, but rather that the combination of constraints is needed.

Grossberg (rightfully) argues that our simulations take a number of important shortcuts with respect to the source of sensory data and the human perceptual process. The samples used in the experiments do not take realistic lighting conditions or surface context into account. The neural network models could be made more realistic, for example, by using Grossberg's Adaptive Resonance Theory (even though we dispute that our networks are incapable of incremental learning, see Fig. 4). However, it is unclear whether making all these changes are crucial for the issues raised in the article. None of the experiments reported by Grossberg address the issue of sharing *for* communication, because they are all based on a single individual network, which is presented a series of samples from which it has to generalise. They do not investigate how a population of networks can become coordinated given different sets of samples, a changing environment and ecology, variation in the perceptual apparatus, and so forth. We think it is perfectly possible to redo all the experiments reported in the article using ART as a categorisation engine instead of the RBF networks, but that the semiotic dynamics reported in our simulations would be quite similar if the same mutual feedback relations between lexicon and category formation are put in place.

Another neural-perceptual argument as to why culture, ecology, or language need not influence the coordination of colour categories comes from **Wachtler**. He argues that neurons basically adapt to achieve optimal information coding of environmental stimuli. This has been demonstrated for visual pathways, and Wachtler believes that this would also be the case for categorisation targeting communication. As Wachtler states: "color vision is adapted to the statistics of natural chromatic signals, which implies shared categories."

Although we agree that the statistics of natural chromatic signals and its information-theoretic optimal coding could play an important role in low-level vision, there is a difference between the conceptual spaces generated by low-level vision and the way this space is cut up for the categorisation process. We are not aware of evidence that the colour terms found in natural languages can be explained on the basis of information-theoretic optimality. Furthermore, as **Wachtler** mentions, there is quite some interindividual variation in the perception of colour (as explored in Fig. R1). There are extreme variations, such as colour deficiency, but also there are subtle variations in colour perception in colour normals. Wachtler describes the idiosyncrasies in colour naming by colour deficient subjects, which, just as colour-normal subjects, have to learn lexical labels for their perceptual categories, but nevertheless, Wachtler believes that lexicon acquisition "does not influence perceptual categories." The subjects just acquire a lexicon with which they try to conform to the linguistic norm. But then, what are the perceptual categories used in this lexicon like? They cannot be identical to the categories of colour normals, but must be framed through the filter of colour deficiency. As Wachtler mentions: "Many color-deficient individuals are not even aware about their condition until their first color-vision test." Does this not illustrate that colour deficient individuals develop some sort of categorisation that is sufficiently coordinated to function in a collective setting?

Vogt & Smith question both the category formation process we have used and the essential mechanism by which categories get aligned through mutual feedback. These choices are argued to be unrealistic with respect to human psychology, and hence our claims "may no longer hold" if a more realistic model is used. Vogt & Smith argue that this more realistic model is cross-situational learning without feedback from success in communication on category formation or lexicon acquisition. Against this comment, we argue as follows.

First of all, the feedback used in the game is not linguistic, but pragmatic. The agents obtain evidence of whether or not the communicative act succeeded and have the opportunity to repair it by additional pointing when it failed. Neither the speaker nor the hearer ever explicitly corrects the words used. Evidence in the literature about the relative absence of corrective feedback concerns the latter not the former. No child psychologist has claimed that children or adults never have a clue whether their communication failed or that they never try to repair failed communication. If a mother asks her child for a cookie and gets a doll instead, she does not simply ignore this situation, but will try to repair it. If we never experience success or failure in communication, then why do we communicate in the first place?

Second, we insist that mutual feedback from language to category formation and from category formation to language are essential for agents to align their categorical repertoires.

Vogt & Smith seem to take an empiricist position, assuming that cross-situational learning is able to develop a repertoire of categories that can be shared without feedback on communicative success. The slow rate of learning they obtained in their experiments is taken to be a virtue, whereas, in fact, the self-organisation of a communication system is already so difficult in realistic circumstances (with many sources of stochasticity and error intervening) that a slow learning rate may mean that the shared communication system cannot get off the ground in more realistic circumstances, particularly if the categorical dimensions are no longer simple and given. Given the observations by **Webster & Kay** of individual variance within the same language group and our argument that humans must be able to quickly align their categorical repertoires within the context and time frame of a single conversation, we cannot afford a slow learning rate. Chromatic perception samples from a continuous space and the infinite number of values that a percept can take might hamper cross-situational learning, especially in large populations; this closely relates to the Sorites paradox characterised by **Davidoff & Luzzatti** (see also, Davidoff 2001). New experiments on cross-situational learning in continuous feature spaces will be needed to examine this argument.

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

Moreover, in another study, we used real robots roaming freely in the environment to compare cross-situational learning (called observational learning) and discrimination-based learning with pragmatic feedback (Steels & Kaplan 2002b). Our conclusion was that even though categories form in cross-situational learning, they are not sufficiently shared to be the basis of a successful communication system. This does not mean that cross-situational learning does not play any role at all in human learning. We believe that any method of learning is welcome and should be used. We do however believe that the coupling of language to category formation based on communicative success is essential for achieving effective communication systems.

Kotchoubey rejects our simulations, and particularly the guessing game, as being irrelevant for psychology and that is, in principle, fine with us, because we do not claim this relevance. However, we feel that he has not taken into account that "verbal behavior" is only possible when there is a categorisation of reality (a "segmentation of the face of nature," quoting Whorf). The studies undertaken in our target article could be (and have been) applied to domains like time or predicate-argument structure. One of our students, Joachim de Beule (2004), has recently carried out experiments in which agents have sufficient data about the temporal sequencing of events derived through visual processing and event recognition algorithms that they can generate a temporal conceptual space and start to cut it up into tense distinctions like present–past–future, or aspect distinctions like perfective–imperfective. Just as in the colour case, these distinctions must be coordinated among the members of the population, which in turn impact how these individuals structure their experiences for communication. These experiments are further illustrations of how the creation of a shared communication system can impact concept formation and thus shape the way that individuals structure their experience. Kotchoubey confuses, in our view, the ability to make a perceptual judgment (e.g., whether two hues or two sounds are similar or not) and a categorical judgment, which assumes a division of the continuous space. To take an example mentioned by **Davidoff**, an individual may be able to clearly see the difference between navy blue and royal blue but may not find it to be significant (and thus confuse samples with these colours), because these categories are not so commonly lexicalised in his or her language.

Maes questions the use of the guessing game, stating that our models "stay trapped in an 'Augustinian' referential theory of meaning" and do not adequately reflect a culturalist view, so that no conclusions can be drawn with respect to human psychology. It is true that we take shared attention for granted in our simulations. Agents are able to play a language game, but this critically involves a shared protocol of interaction and some form of attention influence independent of language, for example, through pointing or eye-gaze following. There is quite a lot of work in developmental robotics at the moment on how joined attention might be achieved, but the problems involved are enormous (see e.g., the review in Kaplan & Hafner 2004). In realistic circumstances, they involve the ability to guess the intentions of other agents and the ability to interpret reactions to verbal interaction. In the target article, all of these issues are shortcut in order to focus on the semiotic dynamics of the total game. But it is not true that the guessing games concern merely sticking labels on existing categories. The experiments show, on the contrary, how the formation of categories is stimulated by playing the language game, and how category prototypes and boundaries shift to become more aligned as a side effect of the game. Moreover, the agents are situated in a shared context in which there is a common task (namely, identify one of the samples in this context), so that discrimination in the specific context becomes a key source of constraints over and above the pointing. The category chosen must not simply be true for the topic, but it must distinguish the topic with respect to the other objects in the context.

Some other commentators question our model of genetic evolution. For example, **Vilarroya**, who is generally supportive of our thesis, argues that we use a totally unrealistic and naive model of genetic evolution because phenotypical traits are represented directly in the genome. The same point is made by **Wang & Gong** who argue that using only mutations and no crossover slows down genetic evolution. **Westbury & Hollis** point out that because the population size is small, only small areas of the total search space can be explored, and because we let the mutation rate vary with fitness, there is a strong bias towards convergence, and hence towards the conclusion that genetic evolution is capable of generating a shared repertoire. We completely agree with these commentators that our model of genetic evolution is extremely simplistic (it is, in fact, the simplest model one might imagine). We did not make it more complex, simply because even this simplistic model serves its purpose, namely, it derives a set of categories in the population that is adequate for discrimination and undergoes genetic assimilation when coupled to language. Making this aspect of the overall process more complex would obscure the overall dynamics. We note though (against Westbury & Hollis) that there is evidence that mutation rates are influenced by fitness, as shown in the SOS gene response discussed by Radman (1975) and Matic et al. (1997), so that this assumption is not entirely devoid of biological plausibility. Moreover, we need to keep at least some parents and children interacting, because the lexicon is transmitted in a cultural way and not genetically. If the total population is replaced, the lexicon would be completely wiped out.

A similar comment against the simplicity of our models is made by **Satterfield**, who argues that we should take the developmental process much more seriously (this point was also made by **Vilarroya**). Indeed, it is well known that neonates and young children might not have the same colour perception as adults have. Children also take a considerable amount of time to learn the meaning of colour terms and to correctly map colour on colour terms (see the references in **Vogt & Smith**). This change could in principle be examined by having a mixed population of agents at different "age levels." It raises the issue of how children and adults are able to interact with each other, even when at different developmental stages. We are all in favour of refining our simulations to introduce this feature. But the question is again, whether making this aspect more realistic would disprove our central thesis. We believe it would not. On the contrary, we see variation caused by differences in developmental stage as an additional argument for why the formation of a repertoire of concepts for communication should be seen as a coordination problem involving constant alignment and adjustment, instead of the search for a single optimal solution. In our view, there is no final state of a language or its underlying ontology, as the group keeps

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

adapting both, whenever communicative needs or conventions change.

The commentary by **Harnad** is highly critical of our approach, perhaps because he assumes that we are trying to explain how language (in its full richness) evolved and how categories that are definable in terms of other categories can be learned, but this is not what we try to do here. We only investigate how far perceptually grounded categories can become sufficiently coordinated in a population to be the basis of an effective communication system. A lot of his criticism is valid if we had this other goal in mind.

We do think however that the guessing game is representative of one of the basic functions of language, namely, drawing attention to some aspect of the environment using signalling, a function that is already present in alarm calls. Suppose two individuals are walking in a forest hunting for mushrooms and they see a number of mushrooms. When one of them says "don't eat the yellow ones," they play a guessing game of the sort used in the article. The speaker has discriminated the bad mushrooms from the others based on a hue distinction and used the word "yellow" to name that distinction. Although the same could be achieved with pointing (indeed, bootstrapping a shared lexicon and its associated categorical repertoire critically depends in our model on nonverbal communication), there are significant advantages over pointing, one of them being that gesture recognition is itself error-prone, another one is that communication then becomes feasible even if the individuals involved are not situated in exactly the same shared context or see the situation from the same viewpoint.

We also disagree with **Harnad** about the nature of categorisation. In our view, categorisation is always relative to some frame of reference and task. There is no absolute correct or absolutely optimal way of categorisation, and the context always plays a role. If all mushrooms are yellow, saying "yellow" does not help very much, the speaker must use more refined colour distinctions or distinctions in another domain, such as shape or size. So we therefore disagree, as well, with **Matthen** who argues that colour perception is not categorical. He makes a comparison with phonemes, which is interesting and relevant (and our group has worked extensively on how a repertoire of speech sounds can self-organise in a population, see de Boer 2000). Phonemes are similar to colour categories in the sense that there is a lot of cross-linguistic variation, so that a particular distinction in one language (like between keen and kin or lap and rap) may be phonologically irrelevant in another (the /ee/ and /i/ sounds are nondistinctive for Spanish and the /l/ and /r/ sounds are nondistinctive for Japanese). Speakers of these other languages therefore have difficulty in accurately hearing or reproducing those distinctions when they are speaking English. Moreover, although some consonants (like bilabial /b/ and dental /d/) show a clear difference because they involve different articulators, vowels and many consonants form prototypes in a continuous space, and so we get exactly the same situation as studied in the target article, namely, agents have to coordinate their ways of categorising sounds by progressive alignment.

Matthen then argues for a more mixed approach, in which some aspects of colour categorisation are genetically determined (e.g., the primary hues based on the *L*a*b** space generated by the opponent channels) and then modulated by cultural processes. This approach is widely accepted and also advocated by **Bódog et al**. It could be investigated further with the tools we have provided in our simulation experiments.

A final critical commentary is provided by **Cowley**. He makes a philosophical argument, attributed to Wittgenstein, that communication relies not on shared categories but on relations and on integration with social activity. We believe that our simulations capture many aspects of the Wittgensteinian approach to language, particularly if compared to a logicist view (as advocated by the early Wittgenstein) in which meaning is supposed to be absolute and shared independently of the social history of interactions illustrated in language games.

R4. Suggestions for additional experiments

Many commentators have proposed fascinating additional experiments and extensions. Most of these could be done in a straightforward manner within the framework we have developed and they would form excellent topics for a Masters or in some cases a Ph.D. thesis. We have already reported on one experiment earlier in which the agents use different low-level processing so that their conceptual space (the *L*a*b** space) is different. We briefly summarise some of the other suggestions:

1. **Bódog et al.** propose an experiment in which the agents first genetically evolve colour categories so that there is an initial base of sharing, and then learning takes over, further refining and adaptively shaping the categories or a particular language. They also suggest introducing a better model for the ecological constraints based on some relevance measure. Both of these experiments could be done in a straightforward way.

2. Based on the observations by **Bimler** and **Jameson** concerning the variation in the perceptual apparatus and the conceptual colour space among humans, we could set up an experiment introducing such a variation. In fact, such experiments were done by our student Joris Bleys.

3. **Hampton** suggests giving agents a different social status so that some are more authoritative than others. Several researchers have done this kind of experiment for the naming game, and show that this has a conservative effect on cultural evolution. This line of investigation is also interesting for current research on the formation of social networks, because language can clearly be a factor in the formation and sustenance of a network.

4. **Yu & Smith** argue for an active learner. The learner is not passively taking in perceptual and linguistic stimuli, but actively explores the world through directing its attention, through experiencing the world through its sensorimotor system and through eliciting interaction from social partners. Their suggestion is welcomed, as indeed our agents do not actively expand their knowledge, but instead passively play language games. The interactions between the agents will certainly benefit from an active exploration of the environment. A concrete way to achieve this is to let agents preferentially communicate about colour stimuli for which they have no label yet. This would speed up the acquisition of a lexicon, and in the case of cultural learning would have repercussions on the acquisition of categories.

5. **Satterfield** argues that we should give different agents different developmental time lines and introduce profound developmental change, including to the conceptual spaces used by the agent depending on age. Such a fas-

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

cinating experiment would be highly valuable, particularly if it could be based on empirical data of child development.

6. **Wang & Gong** suggest adding stochasticity to the feedback received by speaker and hearer (something they have explored in their own work). Additional stochasticity could come from errors in transmission, from cognitive errors in lexicon lookup or categorisation, or from slips in perception. In some other studies, we already showed that if the stochasticity is too high, the communication system will not get off the ground, but once the system is in place, it is sufficiently robust to cope with these errors (see Steels & Kaplan 1998), which is a counterargument to **Harnad**, who claims that verbal communication is not relevant to achieve more robust forms of shared attention.

Some other suggestions have already been investigated by ourselves and other researchers. For example, **Huyck & Mitchell** wonder how hierarchical categories could arise. This was already shown in some of our earlier simulations (see discussion of the "Talking Heads Experiment" in Steels & Kaplan, 2002b). In these experiments, a population of (up to 3,000) agents played the same sort of guessing game as used in the target article. But instead of radial basis function networks, agents used discrimination trees to cut up their conceptual space, with more abstract categories being developed and used before more specific categories. The Talking Heads experiment showed the same key phenomena as discussed in the target article, namely coordination of hierarchical categorical repertoires through language games, because the same mutual feedback between categorisation and naming was implemented. We believe that many different kinds of concept formation algorithms can be employed as long as this mutual feedback is put into place.

The Talking Heads experiment also addressed an issue raised by **van Brakel**, namely, how far is colour a cultural dimension? As van Brakel correctly remarks, colour is already built into the perception, naming, and categorisation behaviour of the agents, but this was only done to focus the experimentation. When agents are given access to more complex sensory data and when the kind of the categories are not predefined, the categories themselves will be subject to cultural pressures. In the Talking Heads experiment, agents use more complex vision algorithms to generate several sensory dimensions, based on size, shape, brightness, spatial position, and so forth and hence multi-dimensional conceptual spaces can be constructed that combine any of these. In this case, there is no guarantee that the hue-based colour space comes out as the basis for the agents' categorical repertoire, let alone that it is the only one. Indeed, in the experiment, we saw that sometimes two agents would use quite different conceptual spaces. For example, for one agent "the right-most object in the scene" (a positional category) could be the meaning of a word "babodo," whereas for another agent the same word might mean "a specific area in the *L*a*b** colour space corresponding to bluish green." This pattern could be stable for a while, leading to successful communication, until a situation arose where the right-most object did not have this colour, at which point the agents had to disentangle those meanings (see Steels & Kaplan 2002b).

These observations are also relevant to the commentary of **Wang & Gong**, who argued that

> Considering heterogeneous sensorimotor systems or learning mechanisms adopted by agents and the multiple features contained in world items, it is possible for agents, through different learning mechanisms, to create different categories for the same world item based on its different features. Besides, if both the categories partitioned in semantics inside one agent and the word forms partitioned in symbols can distinguish world items, it is possible that each agent will develop its own associative network between word forms and its categories, and successful communication is still possible even though there are no shared categories.

Indeed, this is entirely possible and undoubtedly also occurs in human language. **Wang & Gong,** for example, suggest that lexical labels such as "friend" or "loyalty" will be associated with concepts that differ wildly from individual to individual, this without ever hindering communication. We agree, and also believe that abstract concepts that rely on a multitude of sensory information and hierarchical concepts will diverge between individuals.

This brings us back to the initial point made in the beginning of our reply: By sufficient sharing we do not mean that the categories are identical, but rather that they are sufficiently coordinated to allow successful communication.

R5. Conclusions

There is a lot more to say about each individual commentary, and the richness and multiple views they introduce attest to the enormous complexity of human colour categorisation and naming. An important subset of commentaries has provided additional evidence from various angles showing that the environment, physiology, low-level visual processing, developmental stage, social status, and cultural influence are highly varied among individuals, even in the same population. This reinforces our thesis that there is not a single source of constraints on shaping the categories used for communication, but that individuals combine multiple sources of constraints to coordinate their categories. It also reinforces the idea that there is not a single optimal, universal set of categories that simply need to be labelled, but rather that categorical repertoires are shaped and reshaped in a relativistic fashion, even temporarily within the context of a single conversation.

On the other hand, it cannot be excluded that certain sources of constraints are found that are more powerful than the ones found until now, and we therefore also see much value in the commentaries from opponents. For example, the information-theoretic optimality coding suggested by **Wachtler** needs to be explored further and compared to human colour categories, and the more sophisticated and realistic neural networks of **Grossberg** may indeed provide constraints that push the categorical repertoires more towards human categories.

The many suggestions for increased complexity in the models also show that our target article cannot be seen as an endpoint, but rather as the starting point for investigations into the dynamical interactions between the many constraints operating on cognition and culture. It would be fascinating to collect much more data on how humans coordinate their colour categories and to relate these data to the kind of theoretical models of conceptual and linguistic coordination discussed in this target article.

ACKNOWLEDGMENTS

We are extremely grateful to the many commentators for their highly insightful remarks and to the editors of *BBS* who have been

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

very supportive and effective in carrying out the editorial process. This research was sponsored by the Sony Computer Science Laboratory in Paris, the Flemish Fund for Scientific Research (FWO Vlaanderen), the OMLL initiative of the European Science Foundation, and the EU-FET ECAgents and Cogniron project.

References

Letters "a" and "r" appearing before authors' initials refer to target article and response, respectively.

Aisbett, J. & Gibbon, G. (2001a) A general formulation of conceptual spaces as a meso level representation. *Artificial Intelligence* 133:189–232. [AC]

(2001b) Conceptual spaces as voltage maps. In: *Connectionist models of neurons, learning processes, and artificial intelligence*, ed. J. Mira & A. Prieto, pp. 783–90. Lecture Notes in Computer Science 2084. Springer. [AC]

Akhtar, N. & Montague, L. (1999) Early lexical acquisition: The role of cross-situational learning. *First Language* 19:347–58. [PV]

Andrick, G. R. & Tager-Flusberg, H. (1986) The acquisition of colour terms. *Journal of Child Language* 13:119–34. [PV]

Au, T. K. & Laframboise, D. E. (1990) Acquiring color names via linguistic contrast: The influence of contrasting terms. *Child Development* 61:1808–23. [DR]

Aunger, R. (2000) *Darwinizing culture: The status of memetics as a science*. Oxford University Press. [SJC]

Baldwin, D. A. (1991) Infants' contribution to the achievement of joint reference. *Child Development* 62:875–90. [CY]

Balkenius, C. (1998) Are there dimensions in the brain? In: *Spinning ideas: Electronic essays dedicated to Peter Gärdenfors on his fiftieth birthday.* Available at: http://www.lucs.lu.se/spinning/categories/cognitive/Balkenius/index.html [AC]

Belpaeme, T. (2001) Simulating the formation of color categories. In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'01)* Seattle, pp. 393–98, ed. B. Nebel. Morgan Kaufmann. [aLS]

(2002a) Communicating colour embedded in a colour context. Report on experiments on communicating colour between Flemish informants. AI-memo 2002–10, Artificial Intelligence Lab, Vrije Universiteit Brussel. [rLS]

(2002b) *Factors influencing the origins of colour categories*. Doctoral dissertation, Vrije Universiteit Brussel, Artificial Intelligence Laboratory. [aLS]

Bennett, A. T. D., Cuthill, I. C., Partridge, J. C. & Lunau, K. (1997) Ultraviolet plumage colors predict mate preference in starlings. *Proceedings of the National Academy of Sciences of the USA* 94: 8618–21. [WDC]

Berlin, B. & Kay, P. (1969) *Basic color terms: Their universality and evolution*. University of California Press. [DB, MM, aLS, MAW]

Bernstein, B. (1981) Codes, modalities and the process of cultural reproduction. *Language in Society* 10:327–63. [JAH]

Biederman, I. (1985) Human image understanding: Recent research and a theory. *Computer Vision, Graphics and Image Processing* 32:29–73. [AC]

Bimler, D., Kirkland, J. & Jameson, K. (2004) Quantifying variations in personal color spaces: Are there sex differences in color perception? *COLOR Research and Application* 29(2):128–34. [aLS]

Blackmore, S. (1999) *The meme machine*. Oxford University Press. [SB, aLS]

Bloom, P. (2000) *How children learn the meanings of words*. MIT Press. [aLS, PV]

Bloomfield, L. (1935) *Language*. Allen & Unwin. [SJC]

Bornstein, M. (1985) On the development of color naming in young children: Data and theory. *Brain and Language* 26:72–93. [aLS, TS]

(1975) The influence of visual perception on culture. *American Anthropologist* 77:774–98. [aLS]

Bornstein, M. H., Kessen,W. & Weiskopf, S. (1976) Color vision and hue categorization in young human infants. *Journal of Experimental Psychology* 2:115–29. [aLS]

Bowerman, M. & Levinson, S. C., eds. (2001) *Language acquisition and conceptual development*. Cambridge University Press. [aLS]

Bowmaker, J. K. (1998) Evolution of colour vision in vertebrates. *Eye* 12: 541–47 Part 3B. [OV]

Boynton, R. & Scheibner, H. (1967) On the perception of red by "red-blind" observers. *Acta Chromatica* 1:205–20. [TW]

Briscoe, T., ed. (2002) *Linguistic evolution through language acquisition: Formal and computational models*. Cambridge University Press. [aLS]

Bruner, J. S. (2000) Tot thought. *New York Review of Books* 47(4):27–30. [JPMAM]

Bull, L., Holland, O. & Blackmore, S. (2000) On meme-gene coevolution. *Artificial Life* 6:227–35. [SB]

Burton, G. & Moorhead, I. R. (1987) Color and spatial structure in natural scenes. *Applied Optics* 26(1):157–70. [aLS]

Cairns, P., Huyck, C., Mitchell, I. & Wu, W. (2001) A comparison of categorisation algorithms for predicting the cellular localization sites of proteins. *Knowledge and Information Systems: An International Journal* 12:296–300. [CH]

Camazine, S., Deneubourg, J., Franks, N., Sneyd, J., Theraulaz, G. & Bonabeau, E. (2001) *Self-organization in biological systems*. Princeton University Press. [aLS]

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

Cangelosi, A. & Harnad, S. (2001) The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in perceptual categories. *Evolution of Communication* 4(1):117–42. Available at: http://cogprints.org/2036/ [SH]

Cangelosi, A. & Parisi, D., eds. (2001) *Simulating the evolution of language*. Springer. [aLS]

Carpenter, G. A. (2001) Neural network models of learning and memory: Leading questions and an emerging framework. *Trends in Cognitive Sciences* 5:114–18. [SG]

Carpenter, G. A. & Grossberg, S. (1987) A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing* 37:54–115. [SG]

(1991) Pattern recognition by self-organizing neural networks. MIT Press. [SG]

Carpenter, G. A., Grossberg, S. & Reynolds, J. H. (1991) ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks* 4:565–88. [SG]

Carpenter, G. A., Martens, S., Mingolla, E., Ogas, O. J. & Sai, C. (2004a) Biologically inspired approaches to automated feature extraction and target recognition. AIPR 2004: 33rd Workshop on Applied Imagery Pattern Recognition, October 13–15, 2004, Washington, DC. [SG]

Carpenter, G. A., Martens, S. & Ogas, O. J. (2004b) Self-organizing hierarchical knowledge discovery by an ARTMAP image fusion system. *Proceedings of the 7th International Conference on Information Fusion (Fusion 2004)*, pp. 235–42. [SG]

Caywood, M. S., Willmore, B. & Tolhurst, D. J. (2004) Independent components of color natural scenes resemble V1 neurons in their spatial and color tuning. *Journal of Neurophysiology* 91:2859–73. [TW]

Chella, A., Coradeschi, S., Frixione, M. & Saffiotti, A. (2004) Perceptual anchoring via conceptual spaces. *Proceedings of the AAAI-04 Workshop on Anchoring Symbols to Sensor Data.* AAAI Press. [AC]

Chella, A., Frixione, M. & Gaglio, S. (1997) A cognitive architecture for artificial vision, *Artificial Intelligence* 89:73–111. [AC]

(2000) Understanding dynamic scenes. *Artificial Intelligence* 123:89–132. [AC]

Chomsky, N. (1980) *Rules and representations*. Columbia University Press. [TS]

(2002) *On nature and language*. Cambridge University Press. [SJC]

Chouinard, M. M. & Clark, E. V. (2003) Adult reformulation of child errors as negative evidence. *Journal of Child Language* 30:637–69. [PV]

Churchland, P. S. & Sejnowski, T. J. (1992) *The computational brain*. MIT Press. [aLS]

Colunga, E. & Smith, L. B. (2005) From the lexicon to expectations about kinds: A role for associative learning. *Psychological Review* 112(2). [CY]

Cowley, S. J., Moodley, S. & Fiori-Cowley, A. (2004) Grounding signs of culture: Primary intersubjectivity in social semiosis. *Mind, Culture and Activity* 11(2):109–32. [SJC]

Darwin, C. (1871) *The descent of man, and selection in relation to sex. vol II.* John Murray. [SJC]

Dasgupta, D. & McGregor, D. (1992) SGA: A structured genetic algorithm. Technical report, University of Strathclyde, Strathclyde, UK. [CH]

Davidoff, J. (2001) Language and perceptual categorisation. *Trends in Cognitive Sciences* 5(9):382–87. [arLS]

Davidoff, J. & Roberson, D. (2004) Preserved thematic and impaired taxonomic categorisation: A case study. *Language and Cognitive Processes* 19:137–74. [JD]

Davidoff, J., Davies, I. & Roberson, D. (1999) Colour categories in a stone-age tribe. *Nature* 398:203–204. [JD, DR, aLS, MAW]

Davies, I. R. (1998) A study of colour grouping in three languages: A test of the linguistic relativity hypothesis. *British Journal of Psychology* 98:433–52. [aLS]

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

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

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

Dawkins, R. (1976) *The selfish gene*. Oxford University Press (new edition with additional material, 1989). [SB, aLS]
de Beule, J. (2004) Creating temporal categories for an ontology of time. In: *Proceedings of the 16th Belgian–Netherlands Conference on Artificial Intelligence*, ed. R. Verbrugge, N. Taatgen & L. Schomaker, pp. 107–14. University of Groningen. [rLS]
de Boer, B. (2000) Self-organization in vowel systems. *Journal of Phonetics* 28(4):441–65. [rLS]
de Boysson-Bardies, B. (1999) *How language comes to children: From birth to two years*. MIT Press. [aLS]
de Valois, R. L. & de Valois, K. K. (1975) Neural coding of color. In: *Handbook of perception, Vol. V: Seeing*, ed. E. C. Carterette & M. P. Friedman, pp. 117–66. Academic Press. [aLS]
de Valois, R., Abramov, I. & Jacobs, G. (1966) Analysis of response patterns of LGN cells. *Journal of the Optical Society of America* 56(7):966–77. [aLS]
de Valois, R. L., Cottaris, N. P., Elfar, S. D., Mahon, L. E. & Wilson, J. A. (2000) Some transformations of color information from lateral geniculate nucleus to striate cortex. *Proceedings of the National Academy of Sciences USA* 97:4997–5002. [TW]
Dedrick, D. (1996) Color language universality and evolution: On the explanation of basic color terms. *Philosophical Psychology* 9:497–524. [DR]
(1998) *Naming the rainbow: Colour language, colour science, and culture*, vol. 274 of *Synthese Library*. Kluwer Academic. [aLS]
Dennett, D. (1995) *Darwin's dangerous idea: Evolution and the meaning of life.* Simon & Schuster. [SJC]
(1999) The evolution of culture. Charles Simonyi Lecture, Oxford University, February 17. Available at: http://www.edge.org/3rd_culture/dennett/dennett_p1.html[SB]
Derrington, A. M., Krauskopf, J. & Lennie, P. (1984) Chromatic mechanisms in lateral geniculate nucleus of macaque. *Journal of Physiology* 357:241–65. [TW]
Dummett, M. (1975) Wang's paradox. *Synthese* 30:301–24. [JD]
Durham, W. H. (1991) *Coevolution: Genes, culture and human diversity*. Stanford University Press. [aLS]
Edelman, S. (1999) *Representation and recognition in vision.* MIT Press. [AC]
Elman, J. (1993) Learning and development in neural networks: The importance of starting small. *Cognition* 48:71–99. [TS]
Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D. & Plunkett, K. (1996) *Rethinking innateness: A connectionist perspective on development*. MIT Press. [aLS, OV]
Epstein, J. & Axtell, R. (1996) *Artificial societies: Social science from the bottom up.* Brookings Institution Press. [TS]
Ersoy, B., Grossberg, S. & Carpenter, G. A. (2002) Top-down expectations during cortical learning of recognition categories. *Society for Neuroscience Abstracts* 872.1. [SG]
Estes, W .K. (1994) *Classification and cognition*. Oxford University Press. [SG]
Fabre-Thorpe, M. (2003) Visual categorization: Accessing abstraction in non-human primates. *Philosophical Transactions of the Royal Society of London B* 358:1215–23. [SRL]
Fairchild, M. (1998) *Color appearance models*. Addison-Wesley. [aLS]
Ferber, J. (1998) *Multi-agent systems: An introduction to distributed artificial intelligence*. Addison-Wesley. [aLS, TS]
Fodor, J. A. (1983) *The modularity of mind*. MIT Press. [aLS]
Fogel, D. B. (2002) Blondie24: Playing at the edge of AI. Morgan Kaufmann. [CW]
Fogel, L. J. (1999) *Intelligence through simulated evolution: Forty years of evolutionary programming*. Wiley. [aLS]
Fussell, S. R. & Krauss, R. M. (1992) Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. *Journal of Personality and Social Psychology* 62(3):378–91. [SH]
Gamberale-Stille, G. & Tullberg, B. S. (2001) Fruit or aposematic insect? Context-dependent colour preferences in domestic chicks. *Proceedings of the Royal Society of London* 268B:2525–29. [WDC]
Gärdenfors, P. (2000) *Conceptual spaces.* MIT Press. [AC]
Gärdenfors, P. & Williams, M. (2001) Reasoning about categories in conceptual spaces. In: *Proceedings of the Fourteenth International Joint Conference of Artificial Intelligence*, pp. 385–92. Morgan Kaufmann. [AC]
Garrod, S. & Anderson, A. (1987) Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition* 27(2):181–218. [rLS]
Gegenfurtner, K. R. & Sharpe, L. T., eds. (1999) *Color vision: From genes to perception*. Cambridge University Press. [aLS]
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. [aLS]
Gentner, D. & Goldin-Meadow, S., eds. (2003). *Language in mind*. MIT Press. [aLS]
Gerhardstein, P., Renner, P. & Rovee-Collier, C. (1999) The roles of perceptual and categorical similarity in colour pop-out in infants. *British Journal of Developmental Psychology* 17:403–20. [aLS]
Gibbons, R. (1992) *Game theory for applied economists*. Princeton University Press. [aLS]
Goethe, J. (1959) *Faust*, trans. P. Wayne. Penguin. [SJC]
Goldberg, D. E. (1989) *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley. [aLS]
Goldstein, K. (1948) *Language and language disturbances*. Grune and Stratton. [JD]
Gong, T., Ke, J., Minett, J. W. & Wang, W. S.-Y. (2004) A computational framework to simulate the coevolution of language and social structure. In: *Artificial Life IX: Proceedings of the 9th International Conference on the Simulation and Synthesis of Living Systems*, ed. J. Pollack, M. Bedau, P. Husbands, T. Ikegami & R. A. Watson, pp. 158–64. MIT Press. [WSYW]
Griffin, L. D. (2004) Optimality of the basic colours categories. *Journal of Vision* 4:309a. [DB]
Grossberg, S. (1976a) Adaptive pattern classification and universal recoding, I: Parallel development and coding of neural feature detectors. *Biological Cybernetics* 23:121–34. [SG]
(1976b) Adaptive pattern classification and universal recoding, II: Feedback, expectation, olfaction, and illusions. *Biological Cybernetics* 23:187–202. [SG]
(1980) How does a brain build a cognitive code? *Psychological Review* 1:1–51. [SG]
(1999) The link between brain learning, attention, and consciousness. *Consciousness and Cognition* 8:1–44. [SG]
Grossberg, S. & Todorovic, D. (1988) Neural dynamics of 1-D and 2-D brightness perception: A unified model of classical and recent phenomena. *Perception & Psychophysics* 43:241–77. [SG]
Guest, S. & Van Laar, D. L. (2002) The effect of name category and discriminability on the search characteristics of colour sets. *Perception* 31:445–61. [DR]
Gumperz, J. J. & Levinson, S. C. (1996) *Rethinking linguistic relativity*. *Studies in the Social and Cultural Foundations of Language 17*. Cambridge University Press. [aLS]
Güntürkün, O. (2000) Sensory physiology: Vision. In: *Sturkies avian physiology, 5th edition*, ed. G. Causey Whittow, pp. 1–19. Academic. [WDC]
Harnad, S. (1987) Category induction and representation. In: *Categorical perception: The groundwork of cognition*, ed. S. Harnad. Cambridge University Press. Available at: http://cogprints.org/1572/ [SH]
(1990) The symbol grounding problem. *Physica D* 42:335–46. [aLS]
(2000) From sensorimotor praxis and pantomine to symbolic representations. *The Evolution of Language. Proceedings of the 3rd International Conference*, pp. 118–25. Ecole Nationale Supérieure des Télécommunications, Paris – France. Available at: http://cogprints.org/1619/ [SH]
(2005) Cognition is categorization. In: *Handbook of categorisaton in cognitive science*, ed. C. Lefebvre & H. Cohen. Available at: http://cogprints.org/3027/ [SH]
Hassoun, M. (1995) *Fundamentals of artificial neural networks*. MIT Press. [aLS]
Hauser, M. D. (1996) *The evolution of communication*. MIT Press. [AB]
Hauser, M. D. & Fitch, T. (2003) What are the uniquely human components of the language faculty? In: *Language evolution: The states of the art*, ed. M. H. Christiansen & S. Kirby. Oxford University Press. [AB]
Hauser, M. D., Chomsky, N. & Fitch, T. (2002) The faculty of language: What is it, who has it, and how did it evolve? *Science* 298:1569–79. [AB]
Heider, E. & Olivier, D. C. (1972) The structure of the color space in naming and memory for two languages. *Cognitive Psychology* 3:337–54. [DR]
Hendley, C. D. & 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):870–73. [aLS]
Hockett, C. (1960) Logical considerations in the study of animal communication. In: *Animal sounds and communication*, ed. W. E. Lanyon & W. E. Tavolga, pp. 392–430. American Institutes of Biological Sciences. [AB]
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. [aLS]
(1995) *Hidden order*. Perseus. [WSYW]
(2005) Language acquisition as a complex adaptive system. In: *Language acquisition, change and emergence: Essays in evolutionary linguistics*, ed. J. W. Minett & W. S-Y. Wang, pp. 374–98. City University of Hong Kong Press. [WSYW]
Hong, S. & Grossberg, S. (2004) A neuromorphic model for achromatic and chromatic surface representation of natural images. *Neural Networks* 17(5–6):787–808. [SG]
Houston-Price, C., Plunkett, K., Harris, P. & Duffy, H. (2003) Developmental change in infants' use of cues to word meaning. Paper presented at the Eleventh European Conference on Developmental Psychology, Catholic University of Milan, Italy. [PV]

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

Howard, C. M. & Burnidge, J. A. (1994) Colors in natural landscapes. *Journal of the Society of Information Display* 2(1):47–55. [aLS]
Hurford, J. R. (1989) Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua* 77(2):187–222. [aLS]
Iizuka, H. & Ikegami, T. (2004) Simulating autonomous coupling in discrimination of light frequencies. *Connection Science* 16(4):283–99. [TI]
Jameson, K. A. (2005a) Culture & cognition: What is universal about the representation of color experience? *Journal of Cognition and Culture* 5(3/4). [KAJ]
(2005b) Why GRUE? An interpoint–distance model analysis of composite color categories. *Cross-Cultural Research: The Journal of Comparative Social Science* 39(2):159–204. [KAJ, MM]
Jameson, D. & Hurvich, L. M. (1978) Dichromat color language: "Reds" and "greens" don't look alike but their colors do. *Sensory Processes* 2:146–55. [DB, KAJ, TW]
Jameson, K. A. & Alvarado, N. (2003a) Differences in color naming and color salience in Vietnamese and English. *Color Research & Application* 28:113–38. [DR]
(2003b) The relational correspondence between category exemplars and names. *Philosophical Psychology* 10(1):25–49. [aLS]
Jameson, K. & D'Andrade, R. G (1997) It's not really red, green, yellow, blue: An inquiry into perceptual color space. In: *Color categories in thought and language*, ed. C. Hardin & L. Maffi, pp. 295–319. Cambridge University Press. [DB, aLS]
Jameson, K. A., Highnote, S. & Wasserman, L. (2001) Richer color experience for observers with multiple photopigment opsin genes. *Psychonomic Bulletin & Review* 8:244–61. [KAJ]
Jones, C. D., Osorio, D. & Baddeley, R. J. (2001) Colour categorization by domestic chicks. *Proceedings of the Royal Society of London* 268B:2077–84. [WDC]
Kaiser, P. & Boynton, R. (1996) *Human color vision*. Optical Society of America. [aLS]
Kalish, C. W. (1995) Essentialism and graded membership in animal and artifact categories. *Memory & Cognition* 23:335–53. [JAH]
Kaplan, F. & Hafner, V. (2004) The challenges of joint attention. In: *Proceedings of the 4th International Workshop on Epigenetic Robotics*, ed. L. Berthouze, H. Kozima, C. Prince, G. Sandini, G. Stojanov, G. Metta, & C. Balkenius, pp. 67–74. Lund University Cognitive Science Studies 117. [rLS]
Kay, P., Berlin, B., Maffi, L. & Merrifield, W. (1997) Color naming across languages. In: *Color categories in thought and language*, ed. C. Hardin & L. Maffi. Cambridge University Press. [DB, aLS]
(2003) *The world color survey*. Center for the Study of Language and Information, Stanford. [arLS]
Kay, P., Berlin, B. & Merrifield, W. (1991) Biocultural implications of systems in color naming. *Journal of Linguistic Anthropology* 1(1):12–25. [aLS]
Kay, P. & Kempton, W. (1984) What is the Sapir-Whorf hypothesis? *American Anthropologist* 86:65–78. [BK]
Kay, P. & Maffi, L. (1999) Color appearance and the emergence and evolution of basic color lexicons. *American Anthropologist* 101(4):743–60. [DB, aLS]
Kay, P. & McDaniel, C. (1978) The linguistic significance of the meanings of basic color terms. *Language* 54(3):610–46. [MM, aLS]
Kay, P. & Regier, T. (2003) Resolving the question of color naming universals. *Proceedings of the National Academy of Sciences* 100(15):9085–89. [aLS, MAW]
Kendal, J. R. & Laland, K. N. (2000) Mathematical models for memetics. *Journal of Memetics* 4(1). Available at: http://jom-mit.cfpm.org/2000/vol4/kendal_jr&laland_kn.html [SB]
Kiper, D. C., Fenstemaker, S. B. & Gegenfurtner, K. R. (1997) Chromatic properties of neurons in macaque area V2. *Visual Neuroscience* 14:1061–72. [TW]
Klibanoff, R. S. & Waxman, S. R. (2000) Basic level object categories support the acquisition of novel adjectives: Evidence from preschool-aged children. *Child Development* 7(3):649–59. [PV]
Kohonen, T. (1984) *Self-organization and associative memory*. Springer. [SG]
Komatsu, H., Ideura, Y., Kaji, S. & Yamane, S. (1992) Color selectivity of neurons in the inferior temporal cortex of the awake macaque monkey. *Journal of Neuroscience* 12(2):408–24. [aLS, TW]
Koza, J. R. (1992) *Genetic programming: On the programming of computers by means of natural selection*. MIT Press. [aLS, CW]
Krogh, A. & Hertz, J. A. (1995) A simple weight decay can improve generalization. In: *Advances in neural information processing systems 4*, ed. J. Moody, S. Hanson & R. Lippmann, pp. 950–57. Morgan Kauffmann. [aLS]
Kuehni, R. G. (2004) Variability in unique hue selection: A surprising phenomenon. *Color Research and Application* 29:158–62. [MM, MAW]
Kuschel, R. & Monberg, T. (1974) 'We don't talk much about colour here': A study of colour semantics on Bellona Island. *Man* 9:213–42. [DR]
Lakoff, G. (1987) *Woman, fire, and dangerous things.* The Universtiy of Chicago Press. [TI]
Lammens, J. M. (1994) *A computational model of color perception and color naming*. Doctoral dissertation, State University of New York. [aLS]
Langton, C. G., ed. (1995) *Artificial life: An overview*. MIT Press. [aLS]
Lantz, D. & Stefflre, V. (1964) Language and cognition revisited. *Journal of Abnormal and Social Psychology* 69(5):472–81. [aLS]
Leavens, D. A., Hopkins, W. D. & Bard, K. A. (1996) Indexical and referential pointing in chimpanzees (*Pan troglodytes*). *Journal of Comparative and Social Psychology* 110(4):346–53. [SH]
Lee, T.-W, Wachtler, T. & Sejnowski, T. J. (2002) Color opponency is an efficient representation of spectral properties in natural scenes. *Vision Research* 42:2095–2103. [TW]
Lehky, S. R. & Sejnowksi, T. J. (1999) Seeing white: Qualia in the context of decoding population codes. *Neural Computation* 11:1261–80. [aLS]
Lenneberg, E. (1967) *Biological foundations of language*. Wiley. [TS]
Lenneberg, E. H. & Roberts, J. M. (1956) The language of experience: A study in methodology. *International Journal of American Linguistics* memoir 13. [aLS]
Lennie, P., Krauskopf, J. & Sclar, G. (1990) Chromatic mechanisms in striate cortex of macaque. *Journal of Neuroscience* 10:649–69. [TW]
Levinson, S. C. (1997) Yélî dnye and the theory of basic color terms. Paper presented at a seminar at the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. [DR]
Lewin, B., Siliciano, P. & Klotz, M. (1997) *Genes VI, 6th edition*. Oxford University Press. [OV]
Lewontin, R. C., Rose, S. & Kamin, L. J. (1984) Not in our genes. Pantheon. [JAH]
Li, P. & Gleitman, L. (2002) Turning the tables: Language and spatial reasoning. *Cognition* 83:265–94. [BK]
Liberman, A. M. & Mattingly, I. G. (1985) The motor theory of speech perception revised. *Cognition* 21:1–36. [MM]
Liberman, A. M., Cooper, F. S., Shankweiler, D. P. & Studdert-Kennedy, M. (1967) Perception of the speech code. *Psychological Review* 74:431–61. [MM]
Lickliter, R. & Honeycutt, H. (2003) Developmental dynamics: Toward a biologically plausible evolutionary psychology. *Psychological Bulletin* 129(6):819–35. [OV]
Lopez, A., Atran, S., Coley, J. D., Medin, D. L. & Smith, E. E. (1997) The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive Psychology* 32:251–95. [JAH]
Love, N. (2003) Rethinking the fundamental assumption of linguistics. In: *Rethinking linguistics*, ed. H. D. Davis & T. J. Taylor, pp. 69–93. Routledge. [SJC]
Lucy, J. A. (1992) *Language diversity and thought: A reformulation of the linguistic relativity hypothesis*. Cambridge University Press. [DR]
(1997) The linguistics of "color." In: *Color categories in thought and language*, ed. C. L. Hardin & L. Maffi, pp. 320–46. Cambridge University Press. [aLS]
Lucy, J. A. & Shweder, R. A. (1979) Whorf and his critics: Linguistic and nonlinguistic influences on color memory. *American Anthropologist* 81:581–615. [aLS]
Lumsden, C. J. & Wilson, E. O. (1981) *Genes, mind and culture.* Harvard University Press. [SB]
Luzzatti. C. & Davidoff, J. (1994) Impaired retrieval of object-colour knowledge with preserved colour naming. *Neuropsychologia* 32:933–50. [JD]
MacKeigan, T. (2004) A network analysis of the emergence of a new colour term. Paper presented at PICS '04 Progress in Colour Studies, Glasgow, Scotland. [DR]
MacLaury, R. E. (1987) Color-category evolution and Shuswap yellow with green. *American Anthropologist* 89:107–24. [DR]
(1997) *Color and cognition in Mesoamerica*. University of Texas Press. [aLS]
Malkoc, G., Kay, P. & Webster, M. A. (2002) Individual differences in unique and binary hues. *Journal of Vision* 2:32a. [MAW]
Manabe, K., Kawashima, T. & Staddon, J. E. R. (1995) Differential vocalization in budgerigars: Towards an experimental analysis of naming. *Journal of the Experimental Analysis of Behavior* 63:111–26. [WDC]
Markman, A. B. & Makin, V. S. (1998) Referential communication and category acquisition. *Journal of Experimental Psychology: General* 127(4):331–54. [SH]
Marmor, G. S. (1972) Age at onset of blindness and the development of the semantics of color names. *Journal of Experimental Child Psychology* 25:267–78. [DB]
Marshall, J. C. (1980) On the biology of language acquisition. In: *Biological studies of mental processes*, ed. D. Caplan, pp. 106–48. MIT Press. [OV]
Mather, E. & Schafer, G. (2004) Object-label covariation: A cue for the acquisition of nouns? Poster presented at the meeting of the International Society of Infant Studies, Chicago. [PV]
Matic, I., Radman, M., Taddei, F., Picard, B., Bingen, E., Denamur, E. & Eion, J. (1997) Highly variable mutation rates in commensal and pathogenic *Escherichia coli*. *Science* 277:1833–34. [rLS]

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

Matilal, B. K. (1986) *Perception: An essay on classical Indian theories of knowledge*. Clarendon. [EW]
Matsuno, T., Kawai, N. & Matsuzawa, T. (2004) Color classification by chimpanzees (*Pan troglodytes*) in a matching-to-sample task. *Behavioural Brain Research* 148:157–65 [TW]
Matthen, M. (2005) *Seeing, doing, and knowing: A philosophical theory of sense-perception*. Clarendon. [MM]
Maturana, H. & Varela, F. (1998) *The tree of knowledge*, revised edition. Shambhala Press. [aLS]
May, R. (1986) When two and two do not make four: Nonlinear phenomena in ecology. *Proceedings of the Royal Society of London B* 228:241–66. [aLS]
Maynard Smith, J. (1982) *Evolution and the theory of games*. Cambridge University Press. [SH, aLS]
Medgassy, P. (1961) *Decomposition of superposition of distributed functions*. Hungarian Academy of Sciences. [aLS]
Medin, D. L. & Smith, E. E. (1981) Strategies and classification learning. *Journal of Experimental Psychology: Human Learning and Memory* 7:241–53. [SG]
Medin, D. L., Dewey, G. I. & Murphy, T. D. (1983) Relationships between item and category learning: Evidence that abstraction is not automatic. *Journal of Experimental Psychology: Learning and Memory* 9:607–25. [SG]
Miklósi, Á., Gonda, Zs., Osorio, D. & Farzin, A. (2002) The effects of the visual environment on responses to colour by domestic chicks. *Journal of Comparative Physiology* 188A:135–40. [WDC]
Miller, G. (2000) *The mating mind: How sexual choice shaped the evolution of human nature*. Heinemann. [SB]
Mingolla, E., Ross, W. & Grossberg, S. (1999) A neural network for enhancing boundaries and surfaces in synthetic aperture radar images. *Neural Networks* 12:499–511. [SG]
Minsky, M. & Papert, S. (1969) *Perceptrons*. MIT Press. [aLS]
Mitchell, T. (1997) *Machine learning*. McGraw-Hill. [aLS]
Mizokami, Y., Webster, S. M. & Webster, M. A. (2003) Seasonal variations in the color statistics of natural images. *Journal of Vision* 3:444a. [DB]
Mollon, J. D. (1989) "Tho' she kneel'd in that place where they grew . . . ." The uses and origins of primate colour vision. *Journal of Experimental Biology* 146:21–38. [TW]
(2000) Cherries among the Leaves: The evolutionary origins of color vision. In: *Color perception: Philosophical, psychological, artistic and computational perspectives*, ed. Steven Davis, pp. 10–30. Oxford University Press. [AB]
Mollon, J. D., Pokorny, J. & Knoblauch, K. (2003) *Normal and defective colour vision*. Oxford University Press. [aLS]
Munroe, S. & Cangelosi, A. (2002) Learning and the evolution of language: The role of culture variation and learning cost in the Baldwin effect. *Artificial Life* 8(4):311–40. [WSYW]
Munsell (1976) *Munsell book of color, matte finish collection*. Munsell Color Company. [aLS]
Neitz, J., Carroll, J., Yamauchi, Y., Neitz, M. & Williams, D. (2002) Color perception is mediated by a plastic neural mechanism that is adjustable in adults. *Neuron* 35(4):783–92. [aLS]
Neitz, J., Neitz, M. & Jacobs, G. H. (1993) More than three different cone pigments among people with normal colour vision. *Vision Research* 33(1):117–22. [aLS]
Neubauer, N. (2003) Emergence in a multiagent simulation of communicative behavior. *PICS*, Publication Series of the Institute of Cognitive Science, vol. 11, 2004, Department of Cognitive Science, University of Osnabrück. [rLS]
Newport, E. (1990) Maturational constraints on language learning. *Cognitive Science* 14(1):11–28. [TS]
(1991) Contrasting conceptions of the critical period for language. In: *The epigenesis of mind: Essays on biology and cognition*, ed. S. Cary & R. Gelman, pp. 111–30. Erlbaum. [TS]
Nicolis, G. & Prigogine, I. (1989) *Exploring complexity: An introduction*. Freeman. [aLS]
Nijhout, H. F. (1990) Metaphors and the role of genes in development. *BioEssays* 12(9):441–46. [OV]
Nordin, P. & Banzhaf, W. (1995) Complexity compression and evolution. In: Proceedings of the Sixth International Conference on Genetic Algorithms, *ed.* L. Eshelman, pp. 310–17. Morgan Kaufmann. [CW]
Nosofsky, R. M. (1984) Choice, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: Learning ,Memory, and Cognition* 10:104–14. [SG]
(1987) Attention and learning processes in the identification-categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 13:87–108. [SG]
Nosofsky, R. M., Kruschke, J. K. & McKinley, S. C. (1992) Combining exemplar-based category representation and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 18:211–33. [SG]
Nowak, M. A. & Krakauer, D. (1999) The evolution of language. *Proceedings of the National Academy of Science* 96(14):8028–33. [aLS]
O'Hanlon, C. & Roberson, D. (2004) Learning in context: The effects of linguistic contrast and functional salience on children's acquisition of novel colour terms. Poster presented at PICS '04 Progress in Colour Studies, Glasgow, Scotland. [DR]
Oliphant, M. (1996) The dilemma of Saussurean communication. *BioSystems* 37(1–2):31–38. [aLS]
Owings, D. & Morton, E. (1998) *The evolution of vocal communication: A new approach*. Cambridge University Press. [SJC]
Oyama, S. (1985) *The ontogeny of information: Developmental systems and evolution*. Cambridge University Press. [DH, OV]
Page, M. P. A. (2000) Connectionist modelling in psychology: A localist manifesto. *Behavioral and Brain Sciences* 23:443–67. [SG]
Palmeri, T. J., Nosofsky, R. M. & McKinley, S. K. (1994) Recognition memory for exceptions to the category rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21:548–68. [SG]
Parkkinen, J., Hallikainen, J. & Jaaskelainen, T. (1989) Characteristic spectra of Munsell colors. *Journal of the Optical Society of America* 6(2):318–22. [aLS]
Parsons, O. & Carpenter, G. A. (2003) ARTMAP neural networks for information fusion and data mining: Map production and target recognition methodologies. *Neural Networks* 16:1075–89. [SG]
Pepperberg, I. M. & Wilcox, S. E. (2000) Evidence for a form of mutual exclusivity during label acquisition by grey parrots (*Psittacus erithacus*). *Journal of Comparative Psychology* 114:219–31. [WDC]
Pessoa, L., Mingolla, E. & Neumann, H. (1995) A contrast- and luminance-driven multiscale network model of brightness perception. *Vision Research* 35:2201–23. [SG]
Pfeifer, R. & Scheier, C. (1999) *Understanding intelligence*. MIT Press. [TI]
Pinker, S. (1994) *The language instinct: How the mind creates language*. Morrow. [SB, aLS]
(1997) *How the mind works*. W. W. Norton & Co. Inc. [SB]
Pinker, S. & Bloom, P. (1990) Natural languages and natural selection. *Behavioral and Brain Sciences* 13:707–84. [aLS]
Pinker, S. & Jackendoff, R. (2004) The faculty of language. What is so special about it? *Cognition* 95:201–36. [AB]
Posner, M. I. & Keele, S. W. (1970) Retention of abstract ideas. *Journal of Experimental Psychology* 83:304–308. [SG]
Premack, D. & Woodruff, G. (1978) Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 1:515–26. [SH]
Putnam, H. (1975) The meaning of 'meaning'. In: *Mind, language, and reality: Philosophical papers, vol. 2*, ed. H. Putnam. Cambridge University Press. [JAH]
Quine, W. V. O. (1960) *Word and object*. MIT Press. [aLS, EW, WSYW]
(1969) Natural kinds. In: *Ontological relativity and other essays*. Columbia University Press. [MM]
Quinlan, J. (1993) *C4.5: Programs for machine learning*. Morgan Kaufmann. [aLS]
Radman, M. (1975) Endonuclease III: An endonuclease from *Escherichia coli* that introduces single polynucleotide chain scissions in ultraviolet-irradiated DNA. *Basic Life Sciences* 5A:197–200. [rLS]
Raizada, R. & Grossberg, S. (2003) Towards a theory of the laminar architecture of cerebral cortex: Computational clues from the visual system. *Cerebral Cortex* 13:100–13. [SG]
Rice, N. (1980) *Cognition to language*. University Park Press. [PV]
Roberson, D. & Agrillo, C. (under review) Color language and color cognition: Brown and Lenneberg revisited. [DR]
Roberson, D., Davidoff, J. & Braisby, N. (1999) Similarity and categorisation: Neuropsychological evidence for a dissociation in explicit categorisation tasks. *Cognition* 71:1–42. [JD, BK, DR]
Roberson, D., Davidoff, J. & Shapiro, L. (2002) Squaring the circle: The cultural relativity of 'good' shape. *Journal of Culture and Cognition* 2:29–53. [JD]
Roberson, D., Davidoff, J., Davies, I. R. L. & Shapiro, L. R. (2004) The development of color categories in two languages: A longitudinal study. *Journal of Experimental Psychology: General* 133:554–71. [JD, DR]
(2005) Color categories: Confirmation of the relativity hypothesis. *Cognitive Psychology* 50:378–411 . [DR]
Roberson, D., Davies, I. & Davidoff, J. (2000) Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General* 129(3):369–98. [JD, DR, aLS]
Rommetveit, R. (1978) On negative rationalism in scholarly studies of verbal communication and dynamic residuals in the construction of human subjectivity. In: *The social contexts of method*, ed. M. Brenner, P. Marsh & M. Brenner. pp. 16–32. Croom Helm. [EW]
Roper, T. J. & Marples, N. M. (1997) Colour preferences of domestic chicks in relation to food and water presentation. *Applied Animal Behaviour Science* 54:207–13. [WDC]
Rosch, E. (1978) Principles of categorization. In: *Principles of categorisation, in cognition and categorisation*, ed. E. Rosch & B. Lloyd, pp. 27–48. Erlbaum. [aLS, CH]

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

Rosch, E. R & Mervis, C. (1975) Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* 7:573–605. [CH]
Rosch, E. R., Mervis, C. B., Gray, W. D., Johnson, D. M. & Boyes-Braem, P. (1976) Basic objects in natural categories. *Cognitive Psychology* 8:382–439. [JAH]
Rosch-Heider, E. (1971) 'Focal' color areas and the development of names. *Developmental Psychology* 4:447–55. [aLS]
(1972) Universals in color naming and memory. *Journal of Experimental Psychology* 93:10–20. [aLS]
Rosch-Heider, E. & Olivier, D. (1972) The structure of the color space in naming and memory for two languages. *Cognitive Psychology* 3:337–54. [aLS]
Ross, D. & Dumouchel. P. (2004) Emotions as strategic signals. *Rationality & Society* 16(3):251–86. [SJC]
Rumelhart, D. & McClelland, J. (1986) *Parallel distributed processing: Exploration in the microstructure of cognition, vols. 1 and 2*. MIT Press. [aLS]
Rumelhart, D. E. & Zipser, D. (1986) Feature discovery by competitive learning. *Cognitive Science* 9:75–112. [SG]
Sacks, O. W. (1997) *The island of the colorblind and cycad island*. Knopf. [KAJ]
Saffran, J. R., Aslin, R. N. & Newport, E. L. (1996) Statistical learning by 8-month-old infants. *Science* 274:1926–28. [CY]
Sampson, G. (1997) *Educating Eve: The 'language instinct' debate*. Cassell. [aLS]
Sandhofer, C. M. & Smith, L. B. (2001) Why children learn color and size words so differently: Evidence from adults' learning of artificial terms. *Journal of Experimental Psychology: General* 130(4):600–20. [PV]
Sapir, E. (1921) *Language: An introduction to the study of speech*. Harcourt. [aLS, SB]
Satterfield, T. (2001) Toward a sociogenetic solution: Examining language formation processes through SWARM modeling. *Social Science Computer Review* 19(3):281–95. [TS]
(2005) The bilingual bioprogram: Evidence for child bilingualism in the formation of creoles. In: *Proceedings of the 4th International Symposium on Bilingualism,* ed. J. MacSwan. pp. 1070–90. Cascadilla Press. [TS]
Saunders, B. & van Brakel, J. (1997) Are there nontrivial constraints on colour categorization? *Behavioral and Brain Sciences* 20(2):167–228. [aLS, TW]
Schaffner, K. (1998) Genes, behavior, and developmental emergentism: One process, indivisible? *Philosophy of Science* 65:209–52. [OV]
Schlichting, C. D. & Pigliucci, M. (1998) *Phenotypic evolution: A reaction norm perspective*. Sinauer. [OV]
Schlimmer, J. S. (1987) Mushroom database. UCI Repository of Machine Learning Databases (aha@ics.uci.edu). [SG]
Sharpe, L. T., Stockman, A., Jägle, H. & Nathans, J. (1999) Opsin genes, cone photopigments, color vision, and color blindness. In: *Color vision: From genes to perception*, ed. K. R. Gegenfurtner & L. T. Sharpe. Cambridge University Press. [aLS]
Shepard, R. N. (1962a) The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika* 27:125–40. [AC]
(1962b) The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika* 27:219–46. [AC]
(1992) The perceptual organization of colors: An adaptation to regularities of the terrestrial world? In: *Adapted mind*, ed. J. Barkow, L. Cosmides & J. Tooby, pp. 495–532. Oxford University Press. [aLS]
(1994) Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin & Review* 1:2–28. Reprinted in: (2001) *Behavioral and Brain Sciences* 24(4):581–601. [KAJ, aLS]
(1997) The perceptual organization of colors: An adaptation to regularities of the terrestrial world? In: *Readings on color: Vol. 2, The science of color,* ed. A. Byrne & D. Hilbert, pp. 311–56. MIT Press. [KAJ]
Shepard, R. & Cooper, L. (1992) Representation of colors in the blind, color-blind, and normally sighted. *Psychological Science* 3:97–103. [DB, KAJ]
Simoncelli, E. P. & Olshausen, B. A. (2001) Natural image statistics and neural representation. *Annual Review of Neuroscience* 24:1193–1216. [SRL]
Siskind, J. M. (1996) A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition* 61:39–91. [PV]
Smith, A. D. M. (2003) Intelligent meaning creation in a clumpy world helps communication. *Artificial Life* 9(2):559–74. Available at: http://www.ling.ed.ac.uk/~andrew/publications/publications.html [PV]
(2005) Mutual exclusivity: Communicative success despite conceptual divergence. In: *Language origins: Perspectives on evolution*, ed. M. Tallerman, pp. 372–88. Oxford University Press. Available at: http://www.ling.ed.ac.uk/~andrew/publications/publications.html [PV]
Smith, A. D. M. & Vogt, P. (2004) Lexicon acquisition in an uncertain world. Paper presented at the Fifth Evolution of Language Conference, Leipzig. Available at: http://www.ling.ed.ac.uk/~paulv/publications.html [PV]
Smith, J. D. & Minda, J. P. (1998) Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 24:1411–30. [SG]
(2000) Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26:3–27. [SG]
Smith, J. D., Murray, M. J. & Minda, J. P. (1997) Straight talk about linear separability. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 23:659–80. [SG]
Smith, K., Kirby, S. & Brighton, S. (2003) Iterative learning: A framework for the emergence of language: *Artificial Life* 9:371–86. [WSYW]
Smith, L. B., Jones, S. & Landau, B. (1996) Naming in young children: A dumb attentional mechanism? *Cognition* 60:143–71. [CY]
Song, D. & Bruza, P. D. (2003) Towards context-sensitive information inference. *Journal of the American Society for Information Science and Technology (JASIST)* 54:321–34. [AC]
Sorace, A. (2003) Near-nativeness. In: *The Handbook of second language acquisition*, ed. C. Doughty & M. Long, pp. 130–51. Blackwell. [TS]
Soule, T. & Foster, J. A. (1998) Removal bias: A new cause of code growth in tree based evolutionary programming. In: 1998 IEEE International Conference on Evolutionary Computation, *ed.* D. Fogel, pp. 781–86. IEEE Computer Society Press. [CW]
Sperber, D. (1996) *Explaining culture: A naturalistic approach*. Blackwell. [aLS]
Spurrett, D. & Cowley, S. J. (2004) How to do things without words: Infants, utterance-activity and distributed cognition. *Language Sciences* 26(5):443–66. [SJC]
Steels, L. (1996a) Perceptually grounded meaning creation. In: *Proceedings of the International Conference on Multiagent Systems (ICMAS-96)*, ed. M. Tokoro, pp. 338–44. AAAI Press. [aLS]
(1996b) Self-organizing vocabularies. In: *Proceedings of the Conference on Artificial Life V (Alife V) (Nara, Japan)*, ed. C. Langton & T. Shimohara. MIT Press. [aLS]
(1997) The synthetic modeling of language origins. *Evolution of Communication* 1(1):1–34. [aLS]
(2001a) Language games for autonomous robots. *IEEE Intelligent Systems* 16:17–22. [aLS, WSYW]
(2001b) The methodology of the artificial. *Behavioral and Brain Sciences* 24(6):1077–78. A reply to Webb, B. (2001) Can robots make good models of biological behaviour? *Behavioral and Brain Sciences* 24(6):1033–50. [aLS]
(2002) Grounding symbols through evolutionary language games. In: *Simulating the evolution of language*, ed. A. Cangelosi & D. Parisi, pp. 211–26. Springer. [WSYW]
Steels, L. & Kaplan, F. (1998) Stochasticity as a source of innovation in language games. In: *Proceedings of the Conference on Artificial Life VI (Alife VI) (Los Angeles)*, ed. G. Adami, R. Belew, H. Kitano & C. Taylor, pp. 368–76. MIT Press. [arLS]
(1999) Situated grounded word semantics. In: *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI'99) (San Francisco)*, ed. T. Dean, pp. 862–67. Morgan Kauffman. [aLS]
(2002a) AIBO's first words: The social learning of language and meaning. *Evolution of Communication* 4(1):3–32. [aLS]
(2002b) Bootstrapping grounded word semantics. In: *Linguistic evolution through language acquisition: Formal and computational models*, ed. T. Briscoe, pp. 53–73. Cambridge University Press. [rLS]
Steels, L., Kaplan, F., McIntyre, A. & van Looveren, J. (2002) Crucial factors in the origins of word-meaning. In: *The transition to language*, ed. A. Wray, pp. 252–71. Oxford University Press. [aLS, WSYW]
Stengers, I. & Prigogine, I. (1986) *Order out of chaos*. Bantam. [aLS]
Streeter, M. J. (2003) The root causes of code growth in genetic programming. In: Genetic Programming, Proceedings of EuroGP '2003, *ed.* C. Ryan, T. Soule, M. Keijzer, E. Tsang, R. Poli & E. Costa, pp. 449–58. Springer. [CW]
Sturges, J. & Whitfield, T. A. (1995) Locating basic colours in the Munsell space. *COLOR Research and Application* 20(6):364–76. [aLS]
Suchman, L. (1987) *Plans and situated actions*. Cambridge University Press. [aLS]
Surridge, A. K., Osorio, D. & Mundy, N. I. (2003) Evolution and selection of trichromatic vision in primates. *Trends in Ecology & Evolution* 18(4):198–205. [OV]
Teller, D. (1998) Spatial and temporal aspects of infant color vision. *Vision Research* 38:3275–82. [aLS, TS]
Thelen, E. & Smith L. B. (1994) *A dynamic systems approach to the development of cognition and action*. MIT Press. [DH, CY]
Thelen, E., Schöner, G., Scheier, C. & Smith L. B. (2001) The dynamics of embodiment: A field theory of infant perseverative reaching. *Behavioral and Brain Sciences* 24(1):1–34. [DH]
Tinbergen, N. (1953) The herring gull's world: A study in the social behaviour of birds. Collins. [SJC]
Tomasello, M. (1992) The social bases of language acquisition. *Social Development* 1(1):67–87. [CY]
(1999) *The cultural origins of human cognition*. Harvard University Press. [SJC, SH, aLS]
Tommasi, L. & Vallortigara, G. (2004) Hemispheric processing of landmark and geometric information in male and female domestic chicks (*Gallus gallus*). *Behavioural Brain Research* 155:85–96. [WDC]

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press

Troje, N. (1993) Spectral categories in the learning behaviour of blowflies. *Zeitschrift für Naturforschung* 48:96–104. [TW]
Tversky, A. (1977) Features of similarity. *Psychological Review* 84:327–52. [MM]
Valberg, A. (2001) Unique hues: An old problem for a new generation. *Vision Research* 41:1645–57. [TW]
Vallortigara, G. (2004) Visual cognition and representation in birds and primates. In: *Vertebrate comparative cognition: Are primates superior to non-primates?* ed. L. J. Rogers & G. Kaplan, pp. 57–94. Kluwer Academic/Plenum. [WDC]
VanWijk, H. (1959) A cross-cultural theory of colour and brightness nomenclature. *Bijdragen tot de taal-, land- en volkenkunde* 115:113–37. [aLS]
Vogt, P. (2003) Anchoring of semiotic symbols. *Robotics and Autonomous Systems* 43(2):109–20. [aLS]
(2004) Minimum cost and the emergence of the Zipf-Mandelbrot law. In: *Artificial life IX: Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems*, ed. J. Pollack, M. Bedau, P. Husbands, T. Ikegami & R. A. Watson. MIT Press. Available at: http://www.ling.ed.ac.uk/~paulv/publications.html [PV]
Vogt, P. & Coumans, H. (2003) Investigating social interaction strategies for bootstrapping lexicon development. *Journal of Artificial Societies and Social Simulation* 6(1). Available at: http: http://jasss.soc.surrey.ac.uk/6/1/4.html [PV]
Vorobyev, M., Osorio, D., Bennett, A. T., Marshall, N. J. & Cuthill, I. C. (1998) Tetrachromacy, oil droplets and bird plumage colours. *Journal of Comparative Physiology* 185A:621–33. [WDC]
Vygotsky, L. V. (1934) *Thought and language*, trans. E. Hanfmann & G. Vakar. MIT Press. [JPMAM]
Wachtler, T. (2004) Dichromat hue scaling in color space. *Society for Neuroscience Annual Meeting, abstract* 174:4. [TW]
Wachtler, T., Dohrmann, U. & Hertel, R. (2004) Modeling color percepts of dichromats. *Vision Research* 44:2843–55. [TW]
Wachtler, T., Lee, T.-W. & Sejnowski, T. J. (2001) Chromatic structure of natural scenes. *Journal of the Optical Society of America* 18:65–77. [TW]
Wachtler, T., Sejnowski, T. J. & Albright, T. D. (2003) Representation of color stimuli in awake macaque primary visual cortex. *Neuron* 37:681–91. [TW]
Webster, M. A. & Kay, P. (in press) Individual and population differences in focal colors. In: *The anthropology of color*, ed. R. L. MacLaury, G. Paramei & D. Dedrick. Benjamins. [MAW]
Webster, M. A. & Mollon, J. D. (1997) Adaptation and the color statistics of natural objects. *Vision Research* 37:3283–98. [DB]
Webster, M. A., Miyahara, E., Malkoc, G. & Raker, V. E. (2000) Variations in normal color vision. II. Unique hues. *Journal of the Optical Society of America* 17:1545–55. [MAW, TW]
Webster, M. A., Webster, S. M., Bharadwadj, S., Verma, R., Jaikumar, J., Madan, G. & Vaithilingam, E. (2002) Variations in normal color vision. III. Unique hues in Indian and United States observers. *Journal of the Optical Society of America A* 19:1951–62. [MAW]
Werker, J. & Tees, R. (1984) Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behaviour and Development* 7:49–63. [MM]
Westbury, C., Buchanan, L., Sanderson, S., Rhemtulla, M. & Phillips, L. (2003) Using genetic programming to discover non-linear variable interactions. Behavior Research Methods, Instruments, and Computers 35(2):202–16. [CW]
Whorf, B. L. (1956) *Language, thought and reality: Selected writings of Benjamin Lee Whorf*, ed. J. B. Carroll. MIT Press. [aLS, SB, JD, SH]
Winderickx, J., Lindsey, D., Sanocki, E., Teller, D., Motulsky, A. & Deeb, S. (1992) Polymorphism in red photopigment underlies variation in color matching. *Nature* 356:431–33. [aLS]
Wittgenstein, L. W. (1953) *Philosophical investigations*. Macmillan. [SH, JPMAM, aLS]
(1958) *Philosophical investigations*, 2nd edition. Blackwell. [SJC]
(1969) *On certainty*. Blackwell. [SJC]
Worden, R. (1995) A speed limit for evolution. *Journal of Theoretical Biology* 176:137–52. [aLS, OV]
Wright, E. L. (1992) The entity problem in epistemology. *Philosophy* 67:259:33–50. [EW]
(2005) Perceiving socially and morally: A question of triangulation. *Philosophy* 80:311:53–75. [EW]
Wyszecki, G. & Stiles, W. (1982/2000) *Color science: Concepts and methods, quantitative data and formulae, 2nd edition*. Wiley. [aLS]
Yendrikhovskij, S. N. (2001a) A computational model of colour categorization. *Color Research and Application (supplement)* 26:235–38. [KAJ]
(2001b) Computing color categories from statistics of natural images. *Journal of Imaging Science and Technology* 45(5):409–17. [aLS]
Yu, C. & Ballard, D. H. (2004) A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perception* 1:57–80. [CY]
Yu, C., Ballard, D. H. & Aslin, R. N. (in press) The role of embodied intention in early lexical acquisition. *Cognitive Science.* Available at: http: http://www.indiana.edu/~dll/ [CY]

https://doi.org/10.1017/S0140525X05360080 Published online by Cambridge University Press