Understanding the origins of colour categories through computational modelling

Tony Belpaeme*
Artificial Intelligence Lab, Vrije Universiteit Brussel
Pleinlaan 2, 1050 Brussels, Belgium
tony@arti.vub.ac.be

October 31, 2002

Abstract

Human colour perception is continuous, but humans categorise the colour continuum and often label the resulting colour categories. The debate on whether colour categorisation is an individual process, or whether it is embedded in genetic constraints has not been settled yet. Furthermore, as colour categories have colour names, it is claimed that language could have an influence on the categorisation. This paper reports on agent-based simulations that test the validity of different theories, and uncovers the weak and strong points of each. We conclude, from experiments using AI techniques, that colour categorisation is most likely to be cultural process.

1 Introduction

The chromatic perception of our environment is continuous, and we can distinguish millions of different colours. Nevertheless, we attach more importance to some colours than to others by cutting up the colour continuum into *colour categories*. Perceived colours can have a degree of membership to colour categories, the colour of a tomato for example has a high degree of membership for the RED category, but not so much for the PURPLE category and even less for the BLUE category. It is agreed that humans categorise the colour continuum, and that the categories can be labelled (e.g. the RED category is labelled "red" in English and "rouge" in French). We also assume that colour categories have a prototypical nature (Rosch et al., 1976), meaning they have a degree of membership which is maximal for one or more prototypical colours, and a gracefully decreasing degree of membership for others colours.

Although scholars largely agree on the nature of colour categories, the debate on the origins of colour categories has been raging for almost a century

*http://arti.vub.ac.be/~tony

now. Three major positions can be distinguished, (1) genetic determinism, (2) empiricism and (3) culturalism. The first position, genetic determinism, claims that colour categories (and other perceptual categories, such as olfactory and auditory categories) are determined by constraints on human biology. The physiology of the retina and the neural pathways to the visual cortex are largely genetically determined, and –so it is claimed– colour categories reflect this structure. This view is particularly supported by the field work of Berlin and Kay (1969), who observed that different languages seem to have similar colour categories. The order in which names for these categories appear in a language is also fixed: cultures with only two colour terms in their language always have terms for "black" and "white", languages with three terms have an additional word for "red", and so on. According to Berlin and Kay, languages can in total have eleven basic colour terms, which will appear in the following order: black and white, followed by red, green or yellow, blue, brown, and finally, in no particular order: purple, pink, brown, grey. The fact that all languages have a term for red, and that there is intracultural agreement on what "red" is, lends support to the theory that colour categories are innate.

Empiricism, on the other hand, claims that no representations are innate. Instead, the way we process and learn from our perceptual data is innate, and any representations are the result of a learning process of the individual and are thus under the influence of environmental stimuli that the *individual* receives (Elman et al., 1996; Piaget, 1977). Precisely this ontogenetic aspect, the *individual* learning of representations, is unacceptable for scholars defending culturalism (Tomasello, 1999). Instead they claim that the learning of representations is a cultural process, embedded in social *and* linguistic interactions. In this way, colour categories are the result of a self-organising process driven by environmental stimuli and social interactions. Categories are not only learned, they are also sustained at a cultural level. This position, where language has an effect on thought is also known as the Sapir-Whorf thesis (Whorf, 1956; Kay and Kempton, 1984).

Although all three position have been zealously defended –using anthropological experiments and rhetoric building on circumstantial evidence from psychology, neurobiology and neurophysiology– still no consensus has been reached (Hardin and Maffi, 1997; Saunders and van Brakel, 1997). The research presented in this paper tries to shed new light on the discussion by simulating the three positions. Three different computational models have been designed, each modelling a stance in the discussion. The technical details of the models are explicated in section 2. The models all share components on which a consensus has been reached, but differ in specific ways to allow the verification of different hypotheses. The models should fulfill following requirements:

- The colour perception should modelled on human colour perception.
- The colour categories should have a focal point and fuzzy boundaries.
- If linguistic interactions are involved, the model should facilitate lexicalisation of categories.

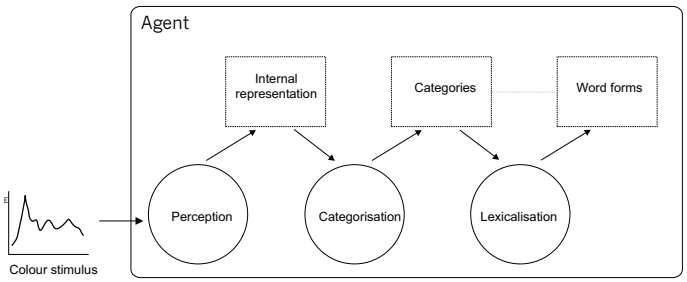

Figure 1: The conceptual structure of an agent.

We will use the models to investigate if they can account for the *sharing* of colour categories *in* and *across* communities. In the simulations the sharing of categories between individuals will be measured by computing the variability between categories. Low variability between categories is evidence for sharing, while a high variability demonstrates that the simulation fails to explain category sharing in a community. Results and conclusions are presented in section 3. The table below summarises the three positions, and how each explains the acquisition and sharing of colour categories in a culture.

Position	Acquisition	Sharing
Genetic determinism	Genetic expression	Gene propagation
Empiricism	Individual learning	Similar environment, ecology and biology
Culturalism	Social learning	Cultural self-organisation

Table 1: Summary of three positions on the origins of colour categories

2 The model and simulation

The simulations are agent-based, with each agent incorporating chromatic perception, categorisation, and where needed, lexicalisation and communication. Perception maps external stimuli onto an internal perceptual space. Categorisation defines categories on the internal space. Lexicalisation attaches labels to categories. And communication allows the agent to convey category labels to other agents. Figure 1 illustrates this.

2.1 The agent

2.1.1 Perception

Colour stimuli are offered to agents as spectral power distributions, which is a distribution of light energy at wavelengths in the visible spectrum. A red colour stimulus for example has high energy at long wavelengths, and no energy at short (or bluish) wavelengths. The human retina has only three photosensitive cells for colour perception, making humans a trichromatic species. One cell type is sensitive to reddish light, one to greenish light and one to bluish light. However, psychologically humans react in an opponent fashion to colour, with red opposed to green, blue to yellow and black to white. This effect has been rigorously described by Jameson and Hurvich (Jameson and Hurvich, 1955) and has been observed in the visual pathways of macaque monkeys (De Valois et al., 1966). We model this conversion from spectral energy to opponent chromatic reactions by using the CIE $L^*a^*b^*$ colour appearance model (Wyszecki and Stiles, 1982). This model converts a spectral representation of a given colour into a three dimensional representation $\{L^*, a^*, b^*\}$, where L^* corresponds to brightness, a^* to the red-green channel and b^* to the blue-yellow channel. The CIE $L^*a^*b^*$ has some interesting properties which are essential for our model: firstly, it mimics human colour perception well, and secondly it has a uniform distance measure between colours, which is needed for computing the membership of colour categories.

2.1.2 Categorisation

The perception only leaves us with a three-dimensional internal representation space, in order to have categories we need to cut up this space. The implementation of the categorisation should respect the properties of human colour categorisation, i.e. the categories should have one or more focal points with a fuzzy extent. Additionally, we need a membership function and we would like the implementation to allow learning, evolution and adaptation of the categories. We have taken *radial basis function networks* (RBFN) (Broomhead and Lowe, 1988) to represent categories.

Figure 2 gives an idea of how a RBFN could represent a prototypical category. Colour representations are fed to the input of the network, and the output returns a membership value of the category for the input. Each RBFN represents exactly *one* colour category. The input unit is fed with three dimensional vector $\mathbf{x} = \{L^*, a^*, b^*\}$. The hidden layer consists of locally tuned functions, this means that a hidden unit responds to a limited region in the representation space. This is implemented with Gaussian response function $z_j(\mathbf{x})$. The network has one output unit, which returns the membership value $y(\mathbf{x})$ of the input by computing the weighted sum of the outputs of the hidden units.

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

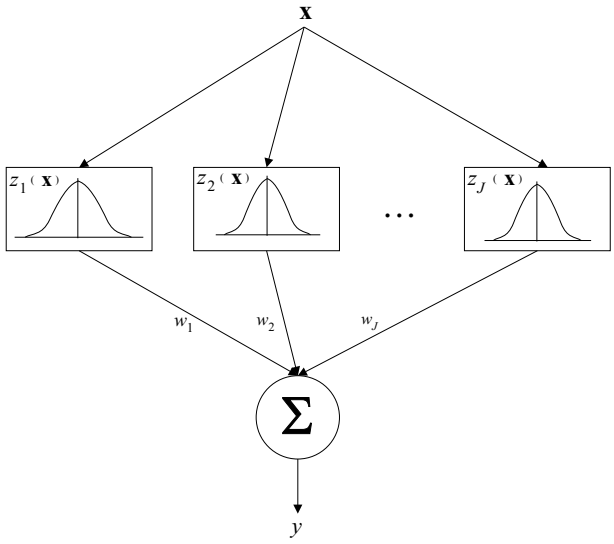

Figure 2: A radial basis function network having one input, one output and a hidden layer containing locally tuned units. The network represents one colour category.

$$y(\mathbf{x}) = \sum_{j=1}^{J} w_j z_j(\mathbf{x}) \tag{2}$$

Equation 1 computes the output of a hidden unit, $\mathbf{m}_{ji}$ is center of the Gaussian and σ_i the width. The width is set to a default value of 1. The output (equation 2) is the weighted sum of the hidden outputs.

2.1.3 Lexicalisation and communication

In some of the simulations the categories can be lexicalised, i.e. associated with word forms. A category c_i can be associated with no, one or more word forms f_j, allowing synonymy. The same word form can be associated with more than one category, allowing for polysemy. The strength with which a word form is associated is given by $\text{score}_{\text{ij}} = [0, 1]$. Initially the score of a word form is set to 0.8 (this number is arbitrary). During the simulation the score is determined by the frequency with which the word form is successfully used in interactions with other agents. Basically, this setup implements an associative memory:

$$\{\langle c_i, \{\langle f_1, \text{score}_{\text{i1}}\rangle, \ldots, \langle f_N, \text{score}_{\text{iN}}\rangle\}\rangle, \ldots\}$$

In the simulations involving linguistic interactions, the agents use these word forms to convey *colour meaning* to each other.

The mapping from colour stimuli to a colour representation and the representation of categories specify the internal structure of an agent. The behaviour and interactions between agents are specific to the position which we would like to simulate and are specified in the following section.

2.2 The simulations

In the simulations two *games* have been implemented. A game constitutes an interaction between the agent and its environment or between an agent and other agents (Steels, 1997; Belpaeme, 2002). There is a *discrimination game*, in which an agent has to distinguish colour stimuli and which serves to acquire and evaluate the repertoire of categories of the agent. And there is a *guessing game*, which is a simple linguistic interaction in which one agent names a colour and the other has to guess which colour was meant.

Building on these games, three different simulations have been constructed where

1. Agents learn colour categories individually (empiricism), this happens through playing discrimination games.

2. Agents learn colour categories through linguistic interactions (culturalism), this happens through playing guessing games between agents.

3. Agents evolve colour categories (genetic determinism), this happens through simulating genetic evolution of categories.

2.2.1 Discrimination game

The discrimination game serves to evaluate the repertoire of categories, and in the simulations where agents learn the colour categories it also serves to construct their repertoire of categories. The game follows a simple scenario (Belpaeme, 2001).

An agent has a (possibly empty) repertoire of categories C. A context O containing colour stimuli is offered to the agent. From the context, a topic o_t is chosen. The agent now finds the matching category for each colour stimulus (the best matching category is the one having the highest output y). The topic can be discriminated when there exists a category which has the highest output to the topic but not to any other colour in the context.

This scenario can fail in two ways: (1) either the agent has no categories yet or (2) if it has categories, no discriminating category could be found.

In the simulation where categories are learned, this provides an opportunity to extend the repertoire or adapt the categories. Extending the repertoire happens through creating a new RBFN with one hidden unit centred on the topic.

Adapting a category is done by adding an additional hidden unit to the RBFN, this shifts the sensitivity of the RBFN in colour space. The decision to either extend the repertoire or to adapt a category depends on the discriminative success of the agent. The discriminative success is a the percentage of games in which the discrimination game succeeded. If it is lower than 95% the category set is extended, else the category set is adapted. By playing several of such games, an agent is able to construct a full repertoire of categories[1].

2.2.2 Guessing games

The guessing game implements a linguistic interaction between two agents: one agents is designated as the *speaker*, the other as the *hearer*. Again this game follows a simple scenario. First a context O is offered to both agents, of which only the speaker knows the topic o_t. The speaker then first plays a discrimination game to find the category that discriminates the topic. Then it looks up the word form having the highest score for the topic category and conveys this to the hearer. The hearer looks up the category associated with the word form, and tries to point out what the speaker meant.

This interaction can fail at several points:

1. The speaker fails at the discrimination game.
2. The speaker has no word form associated with the topic category.
3. The hearer does not know the word form.
4. The hearer knows the word form, but still points out the wrong colour.

However, failure provides opportunity to learn. Each type of failure has a corresponding reaction of the hearer and speaker:

1. The speaker extends and adapts its repertoire of categories as described in 2.2.1.
2. The speaker creates a new word form an associates it with the category matching the topic.
3. The hearer, after the speaker has pointed out the topic, learns to associate the word form with the category having the highest membership value for the topic.
4. The hearer, after the speaker has pointed out the topic, adapts the category to better match the topic.

[1] When learning categories, there is an additional mechanism which takes care of "forgetting" unused categories. This is done by decreasing the weights in the network. Only when a category has proven to be successful at discriminating, the weights are increased again. This is a "house hold" procedure, to prevent the repertoire from being filled with unresponsive and unused categories.

When the game succeeds however, the score of the word form is increased, strengthening the association between the word form and the category.

The interesting point in this process is that the colour categories are under influence of the linguistic performance of the agents: colour categories which are poorly communicated tend to disappear from the repertoire, even though they might be performing well at discrimination.

2.2.3 Genetic evolution of categories

While in the discrimination and guessing game there is opportunity to learn categories and lexical entries, we need a different mechanism to investigate the *evolution* of colour categories. Again we have a population of agents and initially the agents start off with an empty repertoire. Through a process of mutation and selection they eventually arrive at a full-blown repertoire of colour categories. The agents are selected on their successfulness at playing discrimination games; agents who can distinguish colours thrown at them are preferred over agents who cannot. Four mutation operations are implemented: adding a category at a random location in the internal colour space, removing a random category, extending a random category by adding a hidden unit, and removing a random hidden unit from a category. When moving to a next generation, 50% of the population (the fittest agents) is retained, while the agents performing poorly at discrimination are replaced by mutated copies of the fittest agents. The reproduction is asexual as we are not interested in faithfully modelling human reproduction, but rather we are interested in the possibility of categories being transmitted genetically.

3 Results

For the experiments two different colour stimuli sets are used: one containing spectral measurements of over 1200 colour plates of the Munsell colour set (Munsell, 1976), and one containing measurements of natural colours of plants, bark and flowers.

Results show that all three approaches (learning with and without linguistic interaction, and genetic evolution) arrive at category repertoires which perform well at the discrimination task. Figure 3 shows the success rate at discrimination of a typical run for all three approaches. The discriminative success –a measure telling how many times the agents have been successful at a discrimination game– quickly rises over 95%. Note that the time scale on which this happens does not allow proper comparison, as the processes involved for reaching the discriminating repertoire are all very different.

Figure 4 shows the communicative success –measuring how good the agents are at conveying colour meaning to each other– of agents culturally learning colour categories, it shows how the agents attain a lexicon enabling them to communicate. The communicative success is never perfect but hovers around 80%, showing that in one out of five games the agents do not understand each

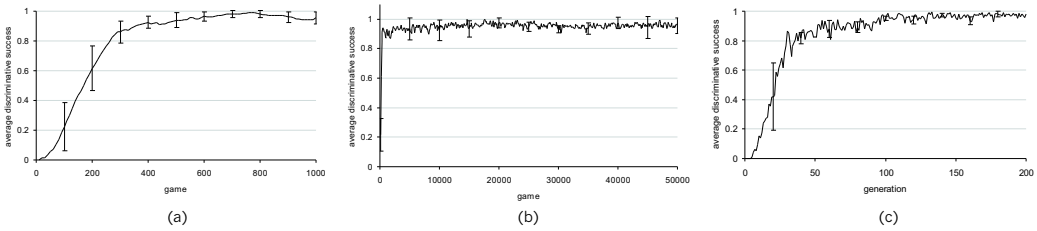

Figure 3: Discriminative success for (a) individual learning of categories, (b) cultural learning of categories and (c) genetic evolution of categories.

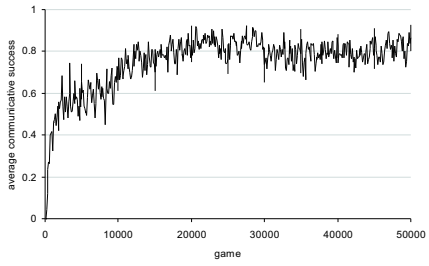

Figure 4: Communicative success for a population which culturally learns its colour categories.

other. This agrees with experiments on human subjects (Stefflre et al., 1966), where is has been noted that humans also exhibit less than perfect colour communication.

As cultures agree on colour categories, it is most interesting to see if the different approaches can arrive at a shared category repertoire within a population. If an approach fails to do so, this means that the approach is a poor candidate for explaining human sharing of colour categories or that some constraint is lacking in the simulation.

To compute the coherence of repertoires of different agents we use a measure called the *category variance* CV (eq. 3). It computes the variance between the categories of all agents of a population, $D_{\text{category set}}$ is a distance measure between categories of two agents A_i and A_j, and N is number of agents in a population (for more information see Belpaeme, 2002). If all categories would be identical, CV would be zero. The more the categories diverge, the higher CV will be.

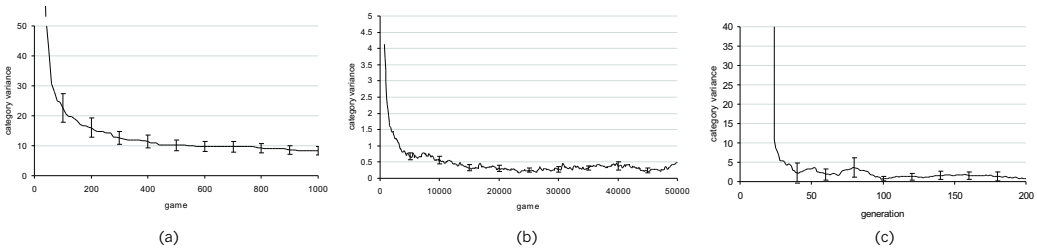

Figure 5: Category variance for (a) individual learning of categories, (b) cultural learning of categories and (c) genetic evolution of categories. Note the scale differences.

$$CV = \frac{1}{2N\left(N-1\right)} \sum_{i=2}^{N} \sum_{j=i-1}^{N} D_{\text{category set}}\left(A_i, A_j\right) \tag{3}$$

Figure 5 shows the category variance for the three approaches. Individual learning leaves the agents with relatively high category variance between the categories, meaning that the categories of agents are not similar. On the other hand, cultural learning and genetic evolution have very coherent categories between agents. With genetic evolution this is easy to understand: after a number of generations every agent will have descended from the same fit parents, and will therefore share the same "colour genes". The fact that cultural learning provides the agents with shared categories is more surprising: here language forms the glue between categories. If categories are not similar, communication would be impossible. We witness a self-organising effect between communication and categorisation, which facilitates the emergence of shared colour categories.

Even though genetic evolution and cultural learning manage to attain sharing *within* a population, further experiments have shown that coherence *across* populations does not arise. The category repertoires of populations diverge as soon as the populations are not in contact with each other. With genetic evolution, undirected mutations are responsible for this, while with cultural evolution the dynamics of the simulation show that "dialects" emerge, each with a different conceptualisation of the colour space (Belpaeme, 2002).

Considering the fact that none of our simulations can explain the sharing across populations, one should wonder if the models or the simulations are not lacking realism. Our natural environment possibly contains structure which is reflected in natural categories (Shepard, 1987). And not only might there be a distribution on the stimuli presented to our senses, the stimuli might also have meaning: red could be strongly associated with fire, blood and danger for example. However, implementing this in a model is not straightforward. It would be a daunting task to measure the distribution of colour stimuli in the

world, and it is nearly impossible to model the semantics of colour stimuli[2]. Nevertheless, even though the model of the environmental stimuli is not faithful enough to explain cross-cultural agreement, the dynamics of the simulations show that cultural evolution is a good candidate for explaining the acquisition of natural categories.

4 Conclusion

The majority of scholars adheres to the nativist account of colour categorisation, but the results of the simulations convince us that cultural evolution is at least as powerful at explaining characteristics of perceptual categorisation. The idea that language has a causal influence on cognition is quite old (Whorf, 1956), but psychological data has never convinced opponents of its validity. We believe that the results reported here show that co-evolution of language and categorisation is a strong candidate for explaining the nature and origins of colour categories and possible other perceptual categories.

Acknowledgements

The author is a postdoctoral researcher of the Flemish Fund for Scientific Research (F.W.O. - Vlaanderen). He would like to thank the three anonymous reviewers who commented on an earlier short version of this paper.

References

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

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

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

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

De Valois, R., Abramov, I., and Jacobs, G. (1966). Analysis of response patterns of LGN cells. *Journal of the Optical Society of America*, 56(7):966–977.

[2]One could devise experiments to measure the salience of colours, but the results will be influenced by subject's language. For the purpose of the simulations presented here, one would like to measure the "pre-linguistic salience" of colours. As yet I have no idea on how to do that.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., and Plunkett, K. (1996). *Rethinking Innateness: A Connectionist Perspective on Development*. The MIT Press, Cambridge, MA.

Hardin, C. and Maffi, L., editors (1997). *Color categories in thought and language*. Cambridge University Press, Cambridge.

Jameson, D. and Hurvich, L. M. (1955). Some quantitative aspects of an opponent-colors theory. I. Chromatic responses and spectral saturation. *Journal of the Optical Society of America*, 45(7):546–552.

Kay, P. and Kempton, W. (1984). What is the Sapir-Whorf hypothesis? *American Anthropologist*, 86(1):65–79.

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

Piaget, J. (1977). *The essential Piaget*. Routledge and Kegan Paul, London. Edited by Gruber, H.E. and Vonéche, J.

Rosch, E., Mervis, C., Gray, W., Johnson, D., and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8:382–439.

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

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science*, 237:1317–1323.

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

Stefflre, V., Vales Castillo, V., and Morley, L. (1966). Language and cognition in Yucatan: a cross-cultural replication. *Journal of Personality and Social Psychology*, 4(1):112–115.

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

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

Wyszecki, G. and Stiles, W. (1982). *Color Science: Concepts and Methods, Quantitative Data and Formulae*. John Wiley and sons, New York, 2nd edition. Reprinted in 2000.