Reaching coherent color categories through communication

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

Abstract

The paper examines the formation of color categories and color terms in a population of autonomous individuals, i.e. simulated agents. Each agent is modeled to perceive color stimuli, to categorize the stimuli and to lexicalize the categories in order to communicate with other agents in the population. During these interactions the agents adapt their internal representations to be more successful in future interactions. The categorization of the color perception is individualistic and influenced only by the nature of the agents' perception and its environment. The color categories can be associated with word forms with which the agents communicate color meanings. The pressure to successfully convey color meaning gives rise not only to coherent color lexicons, but also to a coherent categorization of color perception. The results add to the view that certain aspects of language behave as complex dynamic systems, benefiting from self-organization and cultural interactions.

1 Introduction

Human color perception and related cognitive processes have been extensively studied and have been the topic of many discussions in philosophy, psychology, anthropology and linguistics. Although the focus has always been on descriptive analysis of color cognition, this paper investigates aspects of color using experiments done in artificial and well-controlled experiments to simulate color categorization and color naming in a population of agents. Results are given, and some speculation is offered on how the results can shed new light on old discussions.

Color perception can be studied at different levels. At the physical level, electromagnetic energy is converted in the photoreceptors of the retina into neural signals that are then conveyed to the brain. Humans have three types of color sensitive photoreceptors: one type being sensitive to long (reddish) wavelengths, one to middle (greenish) wavelengths and one to short (bluish) wavelengths. These chromatic photoreceptors are

*The author is a Research Assistant of the Fund for Scientific Research - Flanders (Belgium) (FWO).

cone shaped and are therefore called respectively the L-, M-, and S-cones. Humans are thus a trichromatic species. Psychologically however, humans react to color stimuli in an opponent fashion: human color perception has an antagonistic nature, with red opposed to green and blue opposed to yellow. Combining the trichromatic physiological and opponent psychological nature of color perception gave rise to the two-stage color perception theory of humans: in the first stage light energy is observed by three different kinds of receptors, after subtractively combining the resulting signals in a second stage, the perception runs along opponent channels. However, color perception is still continuous and can be used for little more than reactive behavior; handling color information as symbols requires a division of the spectrum in meaningful categories.

Now color perception is indeed divided into categories; this is immediately suggested by the fact that human languages all have a number of color terms to name some of those categories. The naming of color categories and the categories' referents have been investigated thoroughly over the last decennia. Of all research, the most well known is that of Berlin and Kay [2]. They concluded that there are a minimum of two and a maximum of eleven basic color terms in every language. The fact that color categories agreed very well over different cultures, suggests that color categories are universal. Berlin and Kay's theories are widely accepted, and later research, most notably by Rosch [7], has reconfirmed their conclusions. The universal agreement on color categories is attributed to the nature of human color perception; cultural and environmental influences have played only a minor (if any) role at explaining Berlin and Kay's results.

Recently some disparaging opinions have been published [9, 11]. The critique is mainly focused on the methodology of the experiments (the omission of context in the color samples, the prejudiced interpretation of the results) and the conclusions (the desire to have every language conform to the evolutionary order, the maximum boundary of eleven BCTs, the use of the opponent color space to fit the results on)

Investigating color categorization and color naming can indeed shed light on topics in linguistics such as concept formation, grounding of meaning and the influence of population dynamics on shared concepts. For example, the relation between language and thought has often been the subject of debate. Whorf [14] claimed that language acts as a mould to which thought adapts itself. The strong version of the thesis is not widely supported, but the weak version, claiming that language influences the way humans observe the world has won general acclaim. Color naming has been used as a test field for the Sapir-Whorf thesis [5], there indications have been found that language indeed influences color perception.

It is generally agreed that color perception is genetically defined and identical for all humans; apart from some variation in the sensitivity of color pigments in the photoreceptors and not considering humans with genetic defects (e.g. dichromatism). However, the universality of the categorization of color has only been identified through psychological experiments, and most explanations of these universal tendencies focus on the identical build of the visual pathways, completely ignoring cultural and environmental influences.

The experiments described here draw on theories by Luc Steels [12, 13] explaining language through cultural evolution. Steels considers language to be a distributed, complex dynamic system; in which self-organization in the representation of the individual and in the dynamics of the language community is responsible for stable states. The theory has been used to successfully study certain aspect of language and to offer viable

alternative explanations for linguistic phenomena, see for example [6] on the formation of vowel systems. The research presented in this paper follows the same tradition. In the following section the construction and results of an experiment in which artificial agents discriminate and communicate color are presented.

2 The simulation

An agent in the simulation is able to perceive and categorize its perception, and can associate categories with word forms. The word forms are then used to communicate color meaning to other agents. The agents only share the common environment and the word forms uttered; they have no access to the internal representations of other agents. They can adopt their internal representations to the outcome of linguistic interactions, interactions which come in the shape of guessing games. Through playing several of these games the agents can arrive at a shared color lexicon.

2.1 The individual agent

Color stimuli are offered to agents as spectral power distributions (relative light energy for wavelengths in the visual spectrum from 380nm to 800nm). The colors are offered in *aperture* mode: no contextual information, such as shape or texture, is included. Since humans filter color stimuli through two stage color perception, a mapping is needed from the physical stimulus to a psychophysical representation.

The mapping has to fulfill three requirements; first, two dissimilar stimuli should map on different representations. Only then will discrimination of sensory perception be possible. Second, it should be possible to define a similarity relationship on it; a simple distance measure can be sufficient. And third, the mapping should be a good psychological model of human color perception. In the experiments, the spectral power distributions are mapped to the CIE $L^*a^*b^*$ color space[1]. It is a three-dimensional color space, intended to be perceptually equidistant. It is also an uncomplicated color space, which has proven to work well for categorization [8]. It has three dimensions, L^*, a^* and b^*; L^* represents lightness, a^* corresponds approximately to redness-greenness, and b^* corresponds approximately to yellowness-blueness. The CIE $L^*a^*b^*$ space can represent all human perceivable colors, it is however not able to represent self-luminous colors.

The agents need to be able to split up the color space in meaningful categories, without categories it is impossible to communicate about the perception using word forms. Instead of dividing the color space in discrete classes, a category is represented using an adaptive network, inspired on radial basis functions –see e.g. [4]. An adaptive network has a layer of hidden units, acting as a locally tuned receptors. Figure 1 gives an illustration.

The input unit $\mathbf{x}$ is a three-dimensional vector containing a L^*, a^* and b^*-value. A hidden unit i acts as a Gaussian receptor function with center $\mathbf{m}_i$ and width σ_i. The output $y(\mathbf{x})$ is composed of the weighted sum of the outputs of the hidden units, divided by the

[1]The reader might be familiar with the RGB color space. RGB serves well for technical purposes, but does a poor job at describing human color *perception*; mainly because it is extremely hard to define a perceptual similarity measure

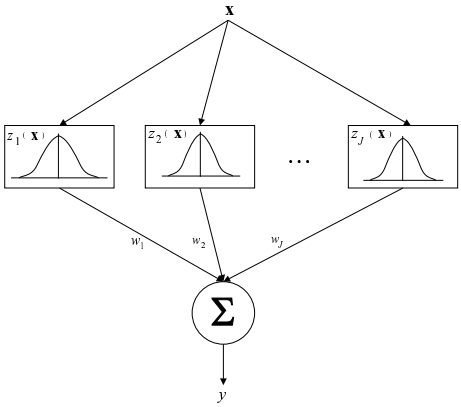

Figure 1: adaptive network for representing a color category, it consists of one hidden layer of locally tuned receptors fully connected to a linear output unit.

number of hidden units J, this to make the output independent of the number of hidden units.

$$y(\mathbf{x}) = \sum_{j=1}^{J} w_j \exp\left(-\frac{(\mathbf{x}-\mathbf{m}_j)^2}{2\sigma_j^2}\right)$$

The network can be adapted by adding or removing hidden units, by shifting the center and changing the width of the hidden units, or by changing the weights of the hidden units. In the simulations the width is set by default to $\sigma_i = 1$ (with $i = 1, ..., J$), and the center $\mathbf{m}_i$ once set is not changed anymore. The network in this way can be used for instance-based learning: when an exemplar with position $\mathbf{p}$ belonging to category c is learned, a hidden unit j with center $\mathbf{m}_j = \mathbf{p}$ and $\sigma_j = 1$ is added to the network representing category c.

The weights w_i of the hidden units are adapted according to the outcome of the inter-agent interactions, and are bound between $[0, 1]$.

The categories can be lexicalized, this is done by attaching word forms to categories; a word form is a string chosen from a finite alphabet. Each category can be associated with one or more word forms, allowing for synonymy; it also possible for the same word form to be associated with more than one category, allowing for polysemy. In this way an agent $\mathcal{A}$ contains a set of associations $\left\{\left\langle c_1^{\mathcal{A}}, F_1^{\mathcal{A}}\right\rangle, ...\right\}$, consisting of pairs containing a category c and a set of word forms $F = \{f_{1,...}\}$. Note that not all categories are lexicalized: when there is no need to communicate a particular category, no word form is assigned to it.

2.2 The dynamics

The agents play two kinds of games. One game, the discrimination game, is played at an individual level and serves to create categories to be able to successfully discriminate the environment. The other game, the guessing game, is a simple interaction played between two agents; a word form is uttered by one agent and interpreted by the other, when both

agents agree on the referent of the word form, the game succeeds. Both agents adapt their internal representations to be more successful at future guessing games. For details on both games, see [13].

The discrimination game The discrimination game serves to create sufficient categories to discriminate the environment. The environment consists of a set of color stimuli, this we call the "context". The game follows a simple scenario, and is completed by one agent. After a certain number of discrimination games, depending on the complexity of the environment, the agent has a set of categories that is sufficient to discriminate any context. For details on how agents create new color categories in the internal color space, the reader is referred to [1].

The guessing game The guessing game is played between two agents randomly selected from the population. One agent acts as the speaker, the other as the hearer. The goal of the game is for the hearer to guess what object the speaker is referring to. A context is presented to both agents, and a topic is selected from the context (only the speaker knows the topic). The speaker tries to discriminate the topic from the context by playing a discrimination game; when this already fails, then the guessing game fails. When the speaker finds a unique category for the topic, it picks one of the word forms associated with that category; if the category has no word forms associated with it yet, the agent creates a random word form and associates it to the category.

This word form f is then conveyed to the hearer. The hearer now looks if it knows the word form, if it doesn't the game fails. If the hearer does know the word form, the category c' associated with f is tried on all objects in the context: the hearer then "points" at the object which reacted best to c'. If this object is the topic, the game succeeds. If not, the topic is revealed to the hearer and it adapts its category c' to better match the topic in future games; this is done by adding a hidden units centered on the topic.

2.2.1 Adapting internal representations

Each time an agent is involved in a guessing game, the weights of all the hidden units are decreased; in this way, the contribution of the hidden units to the output of the category lessens. If the weights of all the hidden units have decreased to 0, the category is removed from the agents' category set. The same happens for unsuccessful word forms, if a word form has reached a certain age and it doesn't succeed enough in conveying meaning, the word form is removed from the agents repertoire. This takes care of the *forgetting* of inadequate categories and word forms.

On the other hand, when a category has been successfully used for discriminating the topic, the weights of its hidden units are increased. In addition, if a word form has been successful at communicating the topic, its success score is increased. In this way, the linguistic interaction has an influence on the internal representations of the agents (cfr. Whorf).

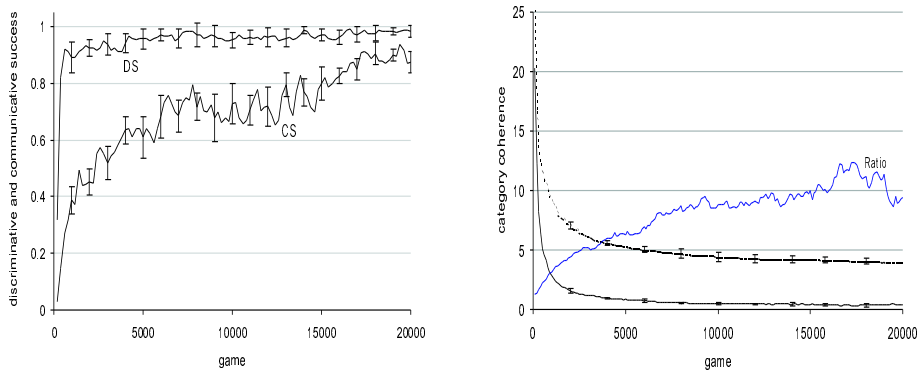

Figure 2: (a) Average discriminative and communicative success for a population of 10 agents. The context contains between two and four stimuli and is chosen from the full set of Munsell chips. (b) The coherence averaged over five runs with communication (full curve) and five runs without communication (dotted curve); for each curve the standard deviation is added. The third curve shows the ratio between the category coherence without communication and with communication.

3 Results

The data set used for the experiments consists of spectral measurements of over 1200 Munsell color chips [10]. A context is selected from this data set, with the constraint that all color stimuli in the context should have a certain distance from each other. By this we avoid that the context contains similar colors; the motivation for this constraint is that humans will not resort to color information for discriminating objects if the color differences are not apparent enough: colors should be distinct before discrimination makes sense. Note that the context complexity increases if there are many elements in it and if the distance between the context elements is small.

The results show that agent create categories with which they can discriminate any context offered, and that the agents' lexicons are coherent enough to enable successful communication. Figure 2a shows the population average of the discriminative success (the ratio of the number of games that an agent discriminated the topic correctly over the last 50 games) and the communicative success (representing its success in the guessing games). The discriminative success quickly rises to 100%, the communicative success rises and oscillates between 80 en 90%, this is due to the dynamics of the simulation. The a priori communicative success is 28.6%, this is the success rate if the hearer would take a blind guess for the topic.

The fact that the communicative success is high indicates that the agents each have a lexicon coherent enough to allow the transfer of meaning. The coherence (see [1]) is another measure for this; it shows how similar the categories are over the population. Figure 2b shows the interpretation coherence of two runs. In one run the agents do not communicate with each other, although the discriminative success will be high, the coherence

between the categories stays rather low since only the shared environment accounts for some rise in the coherence. In the second run, the agents are allowed to communicate (in the form of guessing games) and this significantly increases the coherence of the color categories. The agents all reach a coherent categorization of the color space, and for this the linguistic interactions are in a very significant way responsible.

4 Discussion and conclusion

When there is no communication and when no bias is introduced (for example to have a preference for categories at the opposing locations in the color space), the categorization of the color space is entirely opportunistic. Only the shared environment manages to introduce a slight coherence between the color categories. However, when the need is introduced to communicate color, this –together with the adaptive representations– first drives the agent to develop a shared lexicon, but particularly it also introduces coherence in the color categories of the agent.

However, the distribution of the categories in the color space does not have the regularity observed in human color categories. The agents will in most cases have a category for light-warm and for dark-cool colors, but this is as far as the analogy goes. When no bias is imposed on the environment or on the context, there is no pressure for the categories to resemble color categories as observed by Berlin and Kay.

It is often injudicious to extrapolate results from artificial simulations to real-world phenomena, but often insight can be gained and new approaches can be offered to explain why things are as we observe them. The experiments illustrate that linguistic interactions can drive coherence in a population's color perception, however it is problematic to explain the universality of human color categories solely by cultural factors. Infants for example seem to have a preference for color categories at an age where they do not yet possess any form of language [3]. There are two ends to the spectrum of explanations for the nature of human color categories: either the preference for certain categories emerges from the neurological build of our color perception, or there are near-universal environmental constraints which shape our color perception. The most widely accepted explanations are based on the former. Indeed, biology does account for a large amount of observations, but still some inconsistencies remain; see [11].

The experiment consist of a population of artificial agents –each able to perceive, categorize and lexicalize color stimuli– involved in simple linguistic interactions. It is shown how through adaptation and through the dynamics of the interactions, the agents arrive a common lexicon with which they can convey color meaning to each other. What is remarkable is that while the agents have no access to the internal representations of others, their color categories become coherent by interacting. Although not demonstrated in this paper, the color lexicons and categories stabilize under a large range of parameter settings. The resulting lexicons and the nature of the categories depend on external factors, such as the complexity of the context and the bias in the colors used in the context; and internal factors, such as the learning parameters.

There are still plenty of possibilities for further investigation. For example, the influence of environment bias on the character of the color categories can be studied further. In addition, the circumstances under which categories and word forms emerge that resem-

ble the ones observed in human languages should be further looked into, as well as the influence of contextual information on color categorization and the possible link to color constancy.

References

[1] Tony Belpaeme. Simulating the formation of color categories. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01), Seattle, WA*, 2001.

[2] Brent Berlin and Paul Kay. *Basic Color Terms. Their Universality and Evolution*. University of California Press, Berkeley, CA, 1969. Reprinted in 2000.

[3] Marc H. Bornstein. Color vision and color naming: a psychophysiological hypothesis of cultural difference. *Psychological Bulletin*, 80(4):257–285, 1973.

[4] D.S. Broomhead and D. Lowe. Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2:321–355, 1988.

[5] Roger W. Brown and Eric H. Lenneberg. A study in language and cognition. *Journal of Abnormal and Social Psychology*, 49:454–462, 1954.

[6] Bart de Boer. *The origins of vowel systems*. Oxford University Press, 2001. To appear.

[7] Eleanor (Rosch) Heider. Universals in color naming and memory. *Journal of Experimental Psychology*, 93:10–20, 1972.

[8] Johan M. Lammens. *A computational model of color perception and color naming*. PhD thesis, State University of New York, Buffalo, 1994.

[9] John A. Lucy. The linguistics of "color". In Clyde L. Hardin and Luisa Maffi, editors, *Color Categories in Thought and Language*. Cambridge University Press, 1997.

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

[11] Barbara Saunders and Jaap van Brakel. Are there nontrivial constraints on colour categorization? *Behavioral and Brain Sciences*, 20:167–228, 1997.

[12] Luc Steels. The synthetic modeling of language origins. *Evolution of Communication*, 1(1):1–35, 1997.

[13] Luc Steels. Synthesising the origins of language and meaning using co-evolution, self-organisation and level formation. In James R. Hurford, Michael Studdert-Kennedy, and Chris Knight, editors, *Approaches to the Evolution of Language*. Cambridge University Press, 1998.

[14] B.L. Whorf. *Language, Thought and Reality*. Cambridge, MA, The MIT Press, 1956.