Simulating the Formation of Color Categories

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

Abstract

This paper investigates the formation of color categories and color naming in a population of agents. The agents perceive and categorize color stimuli, and try to communicate about these perceived stimuli. While doing so they adapt their internal representations to be more successful at conveying color meaning in future interactions. The agents have no access to global information or to the representations of other agents; they only exchange word forms. The factors driving the population coherence are the shared environment and the interactions. The experiments show how agents can form a coherent lexicon of color terms and —particularly— how a coherent color categorization emerges through these linguistic interactions. The results are interpreted in the light of theories describing and explaining universal tendencies in human color categorization and color naming. At the same time, the experiments confirm aspects of the theories of Luc Steels [1997; 1998] who views language as a complex dynamic system, arising from self-organization and cultural interactions.

1 Introduction

Color has enthralled scientists for centuries. Many disciplines in science, among which physics, neurology, cognitive science, philosophy, psychology, linguistics and anthropology, have all contributed to a vast body of work on the aspects of human color vision, including color perception, color categorization and color naming. Cognitive processes concerned with color have often been considered as being ideal test ground for verifying theories proposed in the abovementioned disciplines. Moreover, empirical studies on color perception have always offered ample food for thought for quite a few different opinions in cognitive science; often the interpretation being changed to better fit this or that perspective.

The experiments described here study color categorization and color naming in artificial, well-controlled simulations; trying to provide justification or even new insights in theories on color categorization and color naming.

1.1 From color perception to color categories

Human color perception can be studied at several levels. At the neurological level, electro-magnetic energy is transformed in the photoreceptors of the retina into a neural signal, which is then conveyed to the brain. Humans have three different types of color sensitive photoreceptors: one sensitive for reddish light, one for greenish and one for bluish light. The cells are cone shaped and they are respectively called the L, M and S-cones; designating their sensitivity to long, middle or short wavelengths. Humans are thus trichromatic species. However, at the psychological level humans seem to react rather different than one would expect for a species having three types of photoreceptors. Hering's opponent color theory, later defined quantitatively by [Jameson and Hurvich, 1968] and observed experimentally during in vivo experiments on macaque monkeys by [DeValois *et al.*, 1966], argued for an antagonistic nature of color perception: color seems to occur in pairs, with black opposed to white, green opposed to red and blue opposed to yellow. This gave rise to the two stage color theory; in a first stage light is received by three types of photoreceptors, and in the second stage the outputs are interconnected to form red-green, yellow-blue and white-black channels. However, we are still left with a continuous color experience, handling color information would require cutting up that color continuum. This brings us to color categorization.

1.2 Color categories and color terms

Color appearance has a categorical nature; this is immediately suggested by the fact that every language has different color words to indicate different color sensations. The belief was long held that cultures divided the color spectrum into arbitrary categories. However, in 1969 Berlin and Kay published their influential monograph [1969], in which they provide empirical evidence for universal tendencies in color categorization. They conclude that humans have eleven basic perceptual color categories; *basic* meaning that the corresponding color term is a monolexemic, unique color term, salient and unambiguous to all language speakers. Human languages have at least two and at most eleven basic color terms referring to these perceptual color categories (English has all eleven of them: black, white, red, green, yellow, blue, brown, purple, pink, orange and gray). A second conclusion is that basic color terms appear in languages in a specific order. When a language has only two color terms it will be a term for BLACK and WHITE, when a third color term is added, it will be RED, next either GREEN or YELLOW is lexicalized and so on. At about the same time, new quantitative information on the opponent character of color vision seemed to support Berlin and Kay's theories very well [Kay and McDaniel, 1978]. Thus, the universalist stance quickly became widely accepted. However, recently critical views have been offered on universalist extremism, pleading for a more subtle attitude and for more carefully collected and interpreted quantitative data [Saunders and van Brakel, 1997; Lucy, 1997].

1.3 The relation to language

Investigating the formation of color categories and color terms can also help elucidate some aspects of human languages, such as concept formation, lexicon grounding and the propagation of lexicalizations through a population.

Language is unique to humans; although many animals are capable of communicating messages, they are not able of employing the full range of linguistic capabilities as we humans can. Concrete and abstract concept formation, extensive lexicalizations and syntax all seem to be exclusive to humans. The way humans handle abstract reasoning, hierarchical structures and arbitrary mapping is unsurpassed, and there is good reason to believe that language is crucial to all this. The nature of language and the origin of language might indeed help us understand human intelligence.

On the origin of language, two extreme stances exist. Some assume that human language capacity is innate and at large genetically defined [Chomsky, 1980; Pinker and Bloom, 1990; Bickerton, 1998], while others believe that language emerges from the combined play of the human capacity of abstracting and learning and cultural interactions [Deacon, 1997; Steels, 1999].

Steels [1997; 1998] considers language to be the product of cultural evolution. According to Steels language can be seen as a distributed, dynamical and adaptive system. Language is not controlled by one central intelligence; instead, the knowledge of the language is distributed over its users. None of the users has perfect knowledge or control of the language. Language is also robust to changes in the population; users may leave or join the community without significantly affecting the language spoken in the community. In addition language can be seen as a complex dynamic system: categories, concepts, word forms, grammar, ... constantly emerge and change according to population dynamics which can be described using ideas from the field of dynamic systems. These theories have been successfully used, for example to explain the self-organization of universal tendencies in vowel systems [de Boer, 2001]. The simulation described in this paper use the same concepts.

The paper is structured as follows. Section 2 describes the internal organization of the individual agents (the representation of color perception, the categorization and the connection between color word and color categories). Section 3 provides details on the dynamics on the individual level and on the population level. Section 4 provides results illustrating some typical outcomes of simulations, while section 5 and 6 conclude.

2 The agents

The simulation use a population of agents. On an individual level, the agents all have the ability to perceive color, to categorize their perception, to lexicalize their color categories and to adapt to other agents in order to be more successful at communicating color meaning. On the population level, the agents communicate with each other using a simple protocol called the guessing game. In a guessing game, two agents communicate about a visual context. Through playing several thousands of these games a common lexicon is built up.

2.1 Color perception and representation

When perceiving the physical world, a mapping is made from the physical space to a representation in the psychophysical space. Upon this representation, further cognitive actions such as categorization or recognition are taken.

The color stimuli are presented to the agents as spectral power distributions, expressed as energy at wavelengths in the visible spectrum (ranging from 380 nm to 800 nm). No information on spatial or temporal properties are given, the colors are presented in "aperture" mode, void of any contextual information. The sensation $S \in \mathbf{S}$ is a physical stimulus and has to be mapped to a psychophysical representation $R \in \mathbf{R}$, for this a mapping $\mathbf{S} \rightarrow \mathbf{R}$ is needed. The representation $\mathbf{R}$ should fulfill three requirements. First and for all it should be a good model for how humans perceive color. Second, it should make discrimination possible: two stimuli are discriminable if and only if they map onto different points in the representation space. Third, one should be able to define a similarity measure over the representation space.

The CIE $L^*a^*b^*$ color space satisfies these requirements and has proven its merit at categorizing real-world color images (see [Lammens, 1994]). It is a three-dimensional color space designed to be perceptually equidistant and can be represented in Cartesian space. L^* represents lightness, a^* corresponds approximately to redness-greenness and b^* to yellowness-blueness. For a definition and for a conversion from spectral power distribution to CIE $L^*a^*b^*$ see for example [Wyszecki and Stiles, 2000])

2.2 Color categorization

When an agent is to communicate about the world, a symbolic representation of the perception is needed. This symbolic representation arises by cutting up and structuring the representation space.

The color space (which is the psychophysical representation space in these experiments) is used to define categories. A category has a number of features, in our case L^*, a^* and b^*, and for each feature a fuzzy membership function is defined. If an unknown stimulus is perceived, a measure is needed of how well a category matches the unknown representation. Several solutions are possible to represent a category, one could take a radially symmetric function such as a Gaussian; or a multilayer perceptron could be used. Here an adaptive network is chosen (an adaptive network resembles a radial basis function —see eg. [Broomhead and Lowe, 1988]; except that there is only one output unit, which is divided by the number of hidden units). Adaptive networks are our preferred choice for representing categories since they can divide the input space into regions while not being restricted in any way: the regions can be convex or not, symmetrical or not, connected or disparate, even overlap is possible. A second advantage is that adaptive networks are easily analyzed. This is valuable for monitoring the performance of the simulations.

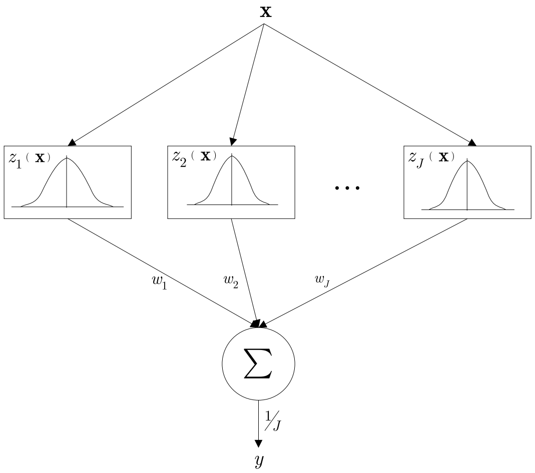

Figure 1: adaptive network for representing a color category, it consists of one hidden layer of locally tuned receptors fully connected to a linear output unit.

Figure 1 shows the adaptive network. It consists of a layer of an unspecified number of hidden units acting as tuned receptors and one output unit. The input $\mathbf{x}$ is a three-dimensional vector containing a L^*, a^* and b^*-value. The hidden units are Gaussian functions $z_j(\mathbf{x})$, with center $\mathbf{m}_j$ and width σ_j. The output of the network $y(\mathbf{x})$ is the weighted sum of the Gaussians, weighted by the number of hidden units.

$$z_j(\mathbf{x}) = \exp\left(-\frac{(\mathbf{x}-\mathbf{m}_j)^2}{2\sigma_j^2}\right)$$

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

For adapting the network, a combination of instance-based and reinforcement learning is used. There are four possible actions: adding or removing a hidden unit, and increasing or decreasing the weight w_j of a unit. The width σ_j of a unit —initialized to a default value— is never changed.

2.3 Lexicalizing categories

Finally, the agents need word forms in order to communicate about color categories: word forms are the only information exchanged by the agents. A category can be associated with one or more word forms, allowing for synonymy. It is also allowed for the same word form to be associated with more than one category, allowing for polysemy. Note that categories are only lexicalized when they need to be communicated; often agents have categories with no word form associated.

An agent has a set of meanings M, a meaning is a pair containing a category $c_i \in C$ and a set of word forms F_i (F_i can be empty).

$$M = \{\langle c_1, F_1\rangle, ...\}$$

Word forms are randomly selected from a finite alphabet, no other restrictions are applied to the creation of word forms (for example, it is not the case that more often used words tend to be shorter, as observed in human languages).

3 The simulation

During a simulation step two kinds of games[1] are played. The *discrimination game* is played at the individual level. The *guessing game* is played at the population level. More on both games can be found in [Steels, 1998]. More details on the implementation of the games can be found in (references to own work are not allowed by the anonymous review process).

3.1 The discrimination game

The goal of the discrimination game is to construct categories in order to successfully distinguish color stimuli. It follows a simple scenario, and is completed by one agent without the need for interactions with other agents.

An agent has a, possible empty, set of categories C. A random context $O = \{o_1, ..., o_N\}$ is created and presented to the agent. It contains N objects o_i (in this case color stimuli) of which one object o_t is the topic. The topic has to be discriminated from the rest of the context. The game proceeds as follows.

1. Context $O = \{o_1, ..., o_N\}$ and the topic $o_t \in O$ are presented to the agent.
2. The agent perceives each object o_i and returns a sensory representation for each object: $S_{o_i} = \{s_1^{o_i}, ..., s_M^{o_i}\}$.
3. For all N sensory representations, the closest matching category $c_{S_o} \in C$ is found.

   $$\forall c \in C : y_c(S_o) \leq y_{c_{S_o}}(S_o)$$

   y_c is the output of the adaptive network belonging to category c, and $y_{c_{S_o}}$ is the output of the adaptive network reacting best to S_o.
4. The topic o_t can be discriminated from the context when there exists a category matching the topic but not matching any other objects in the context.

   $$\left\{c_{S_{o_1}}, ..., c_{S_{o_N}}\right\} \cap c_{S_{o_t}} = \left\{c_{S_{o_t}}\right\}$$

This scenario can fail in two ways. First, the agent has no categories yet ($C = \emptyset$); in this case a category is created with its center on the topic. Second, no discriminating category can be found: the category found for the topic is also

[1] The use of games for studying basic linguistic interactions is largely inspired by Wittgenstein [1953].

found for other objects. When this category is far from the topic (according to some distance measure), a new category is created. When it is closer than a certain threshold distance, the category is adapted by adding a new hidden unit with its center on the topic.

When playing several of these discrimination games an agent is able to create categories that discriminate one object (i.e. color stimuli) from others. Next to basic mechanism, the weight of the hidden units are decreased with every game. Over time, this results in the "forgetting" of hidden units. Only when a category has proven to be useful in an interaction, the weights are increased again.

3.2 The guessing game

For the guessing game, two agents are randomly chosen. One acts as the *speaker*, the other as the *hearer*. A context O is presented to both agents, but only the speaker knows the topic. The game goes along the following scenario.

1. The speaker tries to discriminate the topic by playing a discrimination game, if it finds a discriminating category $c_{S_{o_t}}$ the game continues, otherwise the game fails.

2. The speaker looks if any word forms are yet associated with $c_{S_{o_t}}$. If not, a new word form f is randomly created and associated. If however one or more word forms are already associated with $c_{S_{o_t}}$, then one word form f is selected according to its success in previous guessing games. The word form f is then conveyed to the hearer.

3. The hearer looks if it has f in its associative memory, if not the game fails: the hearer is shown the topic o_t and it learns the proper word form for it by adding a category for the topic.

4. If the hearer does have the word form f in its lexicon it finds the associated category c' and tries to point out the topic. This will only work when the hearer can discriminate the topic from the context, otherwise the game fails.

5. If the hearer succeeds in pointing out the correct object as the topic, the game is successful. If the hearer points out the wrong object, the speaker identifies the topic and the hearer adapts its category c' by adding a hidden unit, so the category resembles the topic better in future games.

When a guessing game is successful the weights of the hidden units of the category $c_{S_{o_t}}$ are increased, this strengthens a category making the probability of it being used in future games higher. Categories which do not contribute to the game are through this less likely to be used in future interactions.

The interactions are responsible for achieving agent lexicons which are coherent over the population. And because the agents adapt their categories according to the outcome of the guessing game, the categories of the agents show coherence as well. This is an illustration of the Sapir-Whorf thesis, which claims that language influences the way its users experience the world [Whorf, 1950].

4 Results

For the simulations two data sets are used as input: one containing spectral measurements taken from over 1200 color chips of the Munsell color notation system and one with 300 measurements of colors of plants and flowers. A context is selected out of these databases, containing minimum 2 and up to 10 different color samples. As a reference, artificial created color samples are use to test the robustness of the system.

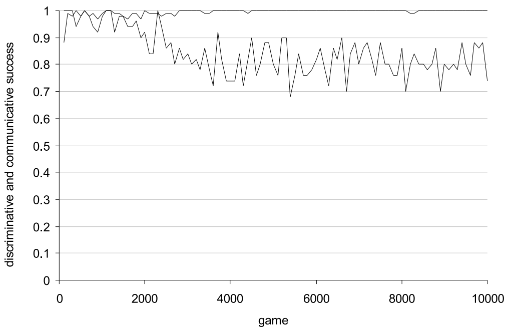

Figure 2: the average success rate plotted for 20 agents over 10000 games. The upper curve shows the discriminative success, the lower curve the communicative success.

The results show that the agents create a set of categories with which they can discriminate any color context offered (provided that the color stimuli in the context are dissimilar enough; for example, the colors can not be metameric). Figure 2 shows the success rate of a population of agents during a typical run[2]. The discriminative success —telling how good the agents are at discriminating the context— quickly rises to 100%. The communicative success —measuring how good the agents are at conveying meaning— rises quickly, then drops off as the games reach their full complexity and then gradually rises again as they agree on a common lexicon.

Another result demonstrates the benefit of communication. When no guessing games are played, so that there is no interaction between the agents, the agents do not manage to form coherent category sets. Clearly the environmental binding is not strong enough to obtain shared categorizations. When the interactive component is introduced, by letting the agents play guessing games, the coherence of the color categories rapidly increases. The coherence is computed by cross-summing the similarity for all categories of the entire population. Let $C' = \cup C$ contain all the categories of all the agents, the coherence is then computed as,

$$\text{coh} = \sum_{i=1}^{C'} \sum_{j=i+1}^{C'} \text{sim}\left(c'_i, c'_j\right)$$

The higher the coherence, the better the categories of the agents agree. Categories are matched according to a similarity function. The similarity between two categories c_a and c_b

[2] Note that similar phenomena are observed for a wide variety of parameter settings, the limited space however does not allow extensive reporting of results.

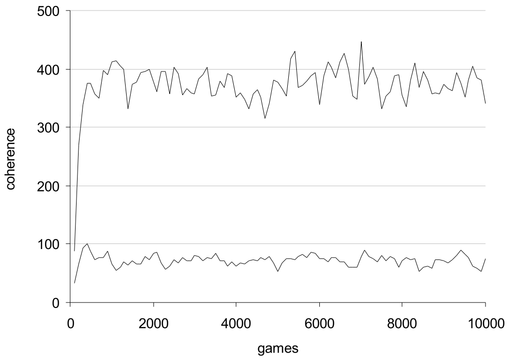

Figure 3: coherence of the color categories in a population. The bottom data series shows the coherence without interactions, the top series shows how interactions increase the coherence.

is computed as in eq. 1, with J_{c_a} and J_{c_b} being the number of hidden units for both categories. It is proportional to the inverse of the Euclidean distances between the centers of the hidden units of both categories.

$$\mathrm{sim}(c_a, c_b) = \frac{1}{J_{c_a} J_{c_b} \sum_{i=1}^{J_{c_a}} \sum_{j=i+1}^{J_{c_b}} \left\| \mathbf{m}_{c_a,i} - \mathbf{m}_{c_b,j} \right\|} \quad (1)$$

Figure 3 shows a typical run, with a population of 20 agents playing 20000 games, the context contains 5 color stimuli selected from the Munsell database. The coherence of a population only playing discrimination games is low compared to a population benefiting from interactions through guessing games; it demonstrates how linguistic interactions are responsible for coherence of categories. This might seem surprising because the agents never have access to the categories of other agents. However, through the linguistic interactions their internal representations adapt to allow for improved transfer of meaning. A side effect of this is that the categories become coherent over the population.

The number of color categories (and proportionally the number of associated word forms) created by the agents rises quickly and stabilizes after a while. Figure 4 shows a typical run: the number of categories stabilizes on an average of 9.4 color categories per agent. The number of color categories depends on several parameter settings. Parameters having a large influence on the number of color categories are (1) the number of color samples in the context, more color samples force the agents to create more color categories to be able to discriminate the colors and (2) the similarity of the color samples; if the samples are rather similar, fine grained categories are needed to discriminate them. In a nutshell, the context exerts pressure on the agents to create more or less categories[3].

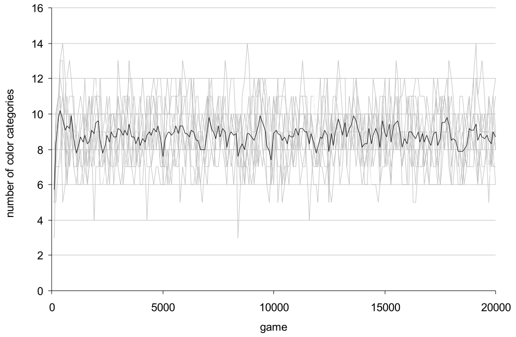

Figure 4: number of categories for 10 agents during 20000 games. The gray curves show the number of categories for each agent, the black curve shows the population average.

The environment does not only influence the quantity of categories, it also affects their quality. When the context contains only highly saturated colors, the agents will only create categories presenting high saturated colors. Likewise, when the context contains a significantly higher amount of red samples, the agents are likely to all have a category and word form for red.

5 Discussion

The evolutionary order of the emergence of named color categories, as observed by Berlin and Kay, does not show up in the experiments where there is no bias imposed on the color perception or on the environment. When in simulation the agents have only two color terms, they will have a term for warm-bright colors and one for dark-cool colors, which is in accordance with observations of human languages. However, when the agents have three or more color terms, there is no preference for creating a category for reddish colors: the creation of categories is entirely opportunistic. For humans the story is different, when human languages have three or more color terms, there will always be a term for reddish colors. Several explanations have been offered for this (see [Hardin, 1987] for an overview) that can be summarized in two ideas: the preference could be built into human biology or it could be rooted in near-universal environmental constraints. For years the focus has been on the former: by interpreting the neurophysiological structure of human color perception one is able to explain many observed phenomena of color lexicalizations. However, some discrepancies remain which can not be explained by the neurological makeup of color perception. For instance, why do some languages not follow the evolutionary order proposed by Berlin and Kay? Why do some languages have more than one word for blue? And why do languages spoken around the equator have less color terms? See [Saunders and van Brakel, 1997].

Although a strong caveat should be issued when generalizing from artificial simulations to real-world phenomena, simulations can sometimes offer new insights and might help us find new ways to tackle problems. The simulations show

[3] This bears resemblance to the folk theories on why equatorial languages have less color terms; it seems language communities there have experienced less pressure to extend their color vocabulary because color technology has evolved more slowly and has only been fairly recently brought up to Western standards.

how in simple artificial linguistic setting, a coherent color lexical and color categorizations can emerge in population of agents. This does not mean that shared human color categories emerge through cultural interactions; there is evidence that already infants have a categorical preference for certain strong hues, long before they engage in linguistic interactions [Bornstein, 1973]. However, it might very likely that it is the interplay between cultural dynamics and biological dispositions that is responsible for how we categorize color, and not color alone.

6 Conclusion

The experiments show how out of self-organization and linguistic interactions a coherent color lexicon can emerge. In addition, it shows how color categories can become shared among a group of language users solely by linguistic interactions. The color lexicons and category sets stabilize under a large range of parameter settings, showing the robustness of the system. The agents and their interaction dynamics form a dynamic system, with attractors in the form of stable lexical and category sets.

Many things still need to be investigated. Most important, the influence of bias on the perception and on the environment needs to be studied. The conditions under which more realistic color categories arise, including the evolutionary order, need to be studied. Already it is clear that the environment and contextual information will play an important role in this. As a bonus, it might be interesting to see if the system could learn color categories and names from a human instructor.

Acknowledgments

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

References

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

[Bickerton, 1998] Derek Bickerton. Catastrophic evolution: the case for a single step from protolanguage to full human language. In James R. Hurford, Michael Studdert-Kennedy, and Chris Knight, editors, *Approaches to the Evolution of Language*, pages 341–358. Cambridge University Press, 1998.

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

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

[Chomsky, 1980] Noam Chomsky. Rules and representations. *Behavioral and Brain Sciences*, 3:1–21, 1980.

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

[Deacon, 1997] Terrence W. Deacon. *The Symbolic Species: The Co-evolution of Language and the Brain*. Norton, 1997.

[DeValois *et al.*, 1966] Russell L. DeValois, I. Abramov, and G.H. Jacobs. Analysis of response patterns of lgn cells. *Journal of the Optical Society of America*, 56(7):966–977, 1966.

[Hardin, 1987] C.L. Hardin. *Color for philosophers*. Hacket Publishing Company, Indianapolis, 1987.

[Jameson and Hurvich, 1968] Dorothea Jameson and Leo M. Hurvich. Some quantitative aspects of an opponent-colors theory. i. chromatic responses and spectral saturation. *Journal of the Optical Society of America*, 49(9):890–898, 1968.

[Kay and McDaniel, 1978] Paul Kay and Chad K. McDaniel. The linguistic significance of the meaning of basic color terms. *Language*, 54:610–646, 1978.

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

[Lucy, 1997] John A. Lucy. The linguistics of “color”. In Clyde L. Hardin and Luisa Maffi, editors, *Color Categories in Thought and Language*, pages 320–346. Cambridge University Press, 1997.

[Pinker and Bloom, 1990] Steven Pinker and Paul Bloom. Natural languages and natural selection. *Behavioral and Brain Sciences*, 13:707–784, 1990.

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

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

[Steels, 1998] 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*, pages 384–404. Cambridge University Press, 1998.

[Steels, 1999] Luc Steels. The puzzle of language evolution. *Kognitionswissenschaft*, 8(4), 1999.

[Whorf, 1950] B.L. Whorf. *Four articles on metalinguistics*. Foreign Service Institute, Washington, 1950.

[Wittgenstein, 1953] Ludwig Wittgenstein. *Philosophical investigations*. Macmillan, New York, 1953.

[Wyszecki and Stiles, 2000] Gunther Wyszecki and W.S. Stiles. *Color science : concepts and methods, quantitative data and formulae*. John Wiley and Sons, 2nd edition, 2000.