Scuola Internazionale Superiore di Studi Avanzati - Trieste

Rewiring color categories:
The neural consequences of language contact

Dissertation by
Ana Laura Diez Martini

Supervisor
Giosuè Baggio

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in Cognitive Neuroscience

January 2015

SISSA - Via Bonomea 265 - 34136 TRIESTE - ITALY

INDEX

Abstract

In this thesis, through a combination of fieldwork, computational modeling, behavioral and neurophysiological (EEG) experimentation, we establish a neural precursor of the acquisition of lexical color categories. The thesis consists of 3 studies, each comprising a number of computational, behavioral and EEG experiments.

The first study (Chapter 1) reports our field research on the color systems of Galician and Spanish, two geographically contiguous and historically related languages. Our aim here was to explore similarities and differences in the way speakers of these languages set boundaries between color categories in the green-yellow-brown and blue ranges. We provide preliminary evidence that regional color meanings can co-exist in neighboring and connected populations.

In a series of computer simulations (Chapter 2) and laboratory experiments (Chapters 2-3), we recreated a minimal language contact scenario, in which speakers of different languages, with possibly different color systems, must coordinate and communicate by means of basic color terms used as signals.

In the second study (Chapter 2), participants learned during two consecutive days an artificial color system by playing as receivers in a signaling game with a computer. In a series of computer simulations, we show that the artificial color system is learnable by agents endowed with minimal cognition and limited memory. The stimuli consisted of an array of 5 Munsell colors that varied along the *hue* dimension from brown to green, with the most ambiguous color occupying the *middle* position in the array. At the end of day two, the EEG was recorded while participants were shown color-term (CT) and term-color (TC) stimulus sequences that were either learned, as part of the artificial color system, or incongruent. We found similar evoked responses to color terms in CT sequences and to colors in TC sequences, with larger late negative ERPs in incongruent than in learned trials. This EEG evidence for category-level color representations was supported by two independent color discrimination studies. ERP effects were largest for the more ambiguous colors, suggesting the strongest neural changes in the brain occur at the boundary of two color categories.

The second experiment was identical to the first, with the exception that color stimuli were different. Here, the five colors varied in *lightness* within the blue hue with an ambiguous blue/black color as the *terminal* color in the array. We obtained larger late negative ERPs, very similar to those observed in the first experiment, in incongruent than in learned trials in CT sequences for the most ambiguous color in the array. Our results indicate that these ERP effects may reflect the effort of rewiring the native color categorization of participants into the artificial color system they have learned.

Acknowledgement

The real authors of this thesis are the hundreds of excellent people I met during these 4 years: scientists, non-scientists and scientist manqué like me.

I would like to thank Giosuè Baggio who never stopped believing in me even when I came to him with the most absurd questions and useless ideas. As a supervisor he provided the perfect combination between freedom for creativity and rigorous structure for scientific thinking. As a mentor, he understood my limitations and showed me that everyone can do science, with his constant assistance, patience and encouragement on this most challenging times. But mostly, I am deeply thankful because he always put his students' happiness first.

I would like to express my great appreciation to Raffaella Rumiati who guided me during my first years and gave me the opportunity to be part of SISSA. This thesis would not have been possible without her. I thank Alessandro Treves for his time to discuss ideas that activated my curiosity.

I would like to offer my special thanks to Asifa Majid and Massimo Warglien for their exhaustive analysis of my thesis and their very valuable comments that let me improve it.

I am indebted to Paul Corballis and his lab at the University of Auckland for their great support and the knowledge shared. I thank them also for the financial support received for my experiments.

The first chapter of this thesis stems from the Evolution of Semantic Systems project and received financial support from the Max Planck Gesellschaft. I want to specially thank Daniel Areán Fraga and his family for the great help with my data collection in Galicia.

I would also like to extend my thanks to every person working at SISSA, who sometimes with a smile changed my day. In particular Alessio, Andrea, Federica and Riccardo for their great help with all the obstacles I encountered.

It was an honor for me to meet the most amazing and loving friends, without whom I would have never reached this moment. I thank them for the long days spent being there in times of happiness and hopelessness. Rtwik for the silence shared, Alessandro L. for his help with numbers and long talks with mate, Jenny, Laura, Alessandro, Giovanni and Vahid, with whom I started this path side to side. My lab friends Iga and Massimo, who I feel very lucky to have worked with. I want to specially thank people who brought music, adventure and happiness to my days: Alan, Alex, Andrea, Andreina, Angus, Arash, Athena, Cimi, Cinzia, Claudia, Duvan, Francesco, Georgette, Graciela, Indrajeet, Juan, Maria, Natalia, Milad, Paola, Pilar, Sahar, Shima, Silvia, Sina, Victor, Yamil, and many, many more. I never thought such a small white box could contain so many mind- boggling animals.

Finally, I am deeply grateful to my family. My family in Argentina who supported this journey with all the love, understanding and unconditional acceptance. My family in Norway for their advice and support. My guru and friend Jorge G.C. for helping me throughout my academic life, Fernanda and Julio, among others who were always there for me from the long distance. Lastly, I would like to thank Nan who promised to teach me how to fly after my defense.

General Introduction

1. Language contact and color categories

Language contact occurs in any situation in which speakers of different languages interact. In some cases, new dialects may arise, combining elements of the two original codes as well as adding elements that were not present in either language before interaction.

An example of this is the emergence of pidgin and creole languages. Pidgins are languages that arise when two speakers of different languages who have no common language try to converse. They often emerged in trade colonies which developed around trade forts or along trade routes (Mufwene 2002). They served as non-native lingua franca to speakers who used their native language for their day-to-day interactions. Pidgins are reduced in structures and specialized in functions (usually trade). The term pidgin apparently comes from a Chinese pronunciation of "business" (Peters 1845). When a pidgin language becomes the native language of the users in a community, it is considered a creole language. Examples of creole languages include Jamaican Patois, Bahamian Creole, Haitian Creole, Mauritian Creole, among others. Language contact also leave some of them on the verge of extinction (Lewis et al 2014) and some of them disappear completely (Crystal 2000). On the other hand, previously thought of as code-switching, adstraum or borrowing (Greenberg 1999), mixed languages started to be considered as a legitimate form of language contact (e.g. Thomason & Kaufman 1988). These are languages that maintain the complexity of two languages, using for example the grammar of one and basic vocabulary of another one (Bakker & Mous 1994). Speakers of mixed languages are often native speakers of the two languages that mixed. An example of a mixed language is Media Lengua or Quichuañol is a mixed language that consists of Spanish vocabulary and Quichua grammar, most conspicuously in its morphology (Muysken 1997).

The presence of bilingual speakers is a crucial factor of change in a language spoken in a monolingual community, since the languages spoken interact and influence each other. A number of studies have provided evidence that cultural evolution is a key factor shaping linguistic regularities (Dunn 2011).

The consequences of language contact are fascinating. One of them is the realization that the way of labeling and categorizing reality across different languages might not be equal and hence easy to translate word by word. This raises the question of whether the language we speak affects how we think of the world. Malt & Majid (2013) and Majid (2014) review more than a century of cross-linguistic research on the interaction between thought and language. They show many examples of how much the boundaries between one category and another one can vary across languages in many perceptual domains. Malt & Majid mention the case of body parts: even if joints create visually and kinesthetically salient a segmentation of the human body, humans around the world experience the body and its segmentation differently. For instance, some languages lack a word for the head, (Burenhult 2006) and some label arms and legs together with only a single word (Terrill 2006). Majid (2014) shows some examples of how sound qualities are also treated differently cross-linguistically (Eitan & Timmers, 2010). An English speaking music director would move her hand vertically to indicate a higher or lower pitch required but if she were to conduct an orchestra in Iran for example, she would probably be clearer using another strategy: speakers of languages like Farsi, Turkish and Zapotec describe pitches as thin or thick instead of low and high (Shayan 2011).

In the visual domain, color categorization is one of the most studied. Reseach such as De Valois et al (1966) used microelectrodes to monitor single neurons to study neural representation of color. From these neurophysiology studies we learned that there are cells in the lateral geniculate nucleus that respond to either green, blue, red, yellow, black or white, and that are inhibited by the opposite colour. Their hypothesis was that the outputs of these cells

correspond to the unique hues. However, the respective wavelengths chosen to correspond to the most representative colors of blue, yellow and green do not consistently match the predictions from neurophysiology.

Berlin and Kay (1969) argued that color naming was not linguistically constrained. Their work was one of the first cross-linguistic study of basic color terms and before that it was thought that languages split the color spectrum into categories without any constraint (Gleason 1961). They compared nearly 100 languages and analyzed what speakers thought was the best exemplar of each color category (focal colors). They demonstrated that in languages that had the same amount of color terms, focal colors tended to cluster reliably in relatively narrow regions of the array of all colors presented. However, boundaries between one and another category are drawn with low consistency and consensus for any language.

Color naming universals may arise from learning and perceptual biases of human learners, brought out by cultural transmission (Xu et al. 2010). Models and simulation studies have offered a valuable parallel contribution to the subject. Komarova et al. (2007) provide a mathematical model in which color naming systems rapidly evolve in populations of communicating agents. Loreto et al. (2012) describe a theoretical framework where the origin of the color hierarchy is explained as the result of purely cultural processes constrained by basic properties of human vision. Although language contact is a widespread cultural phenomenon, its consequences for brain physiology and organization are poorly understood: what changes occur in the brain of language users when they acquire lexical meanings, specifically, when we categorize a continuous space such as color?

In this work, we consider a limited artificial vocabulary of color terms as a test case. Fonteneau et al. (2007) first reported a neural correlate of color categorical perception. They found the peak latency in ERPs for a small color difference bridging a color-term boundary to be quicker than that for an equally sized color difference not bridging a color-term boundary. Other studies show that learning associations between artificial words and color

categories increases grey matter volume in V2/3 of the left visual cortex (Kwok et al. 2011). In an fMRI study, Bird et al (2014) presented participants with two colors in a successive fashion without requiring any kind of judgment of them. The stimuli were 4 squares, each of a different color, one of them expected to be named green and the three remaining blue in a naming task they conducted following the fMRI sessions. They had six conditions where stimuli in a block were either same- or different-category, and where the difference in hue between stimuli in a block was either absent (same condition) or "small," "medium," or "large". They found regions of the brain that independently code for differences in color category and the size of the differences in hue between colors. The middle frontal gyrus in both hemispheres showed stronger activation for same- vs. different-category color differences but was invariant to the size of the hue differences. They did not find any color categorical effects in the visual cortical regions. However, there was a region of the visual cortex that was sensitive to the size of the hue difference.

Roberson & Davidoff (2000) showed participants a color and then asked to read color terms, adding a verbal interference, or look at a multicolored dot pattern adding a visual interference. They were then shown 2 color chips - the original color and one that was 1 or 2 color chips away in an array. Finally they asked them which the original color was. They found that verbal interference only interfered in their accuracy in trials with colors of different categories (across-category identification). They concluded it is verbal encoding what causes judgments of greater perceptual distance between one and another color.

On the other hand, Tan et al (2008) compared, using fMRI, "easy-to-name colors" to "hard-to-name" (more ambiguous) colors in Chinese. By presenting participants with a color discrimination task they found that perceptual identification of easy-to-name and hard- to-name colors activated largely overlapping brain areas; easy-to-name colors, however, were more related to activation of the left posterior temporo-parietal circuits. These are the same regions that also contribute to word-finding processes engaged when a color

is named aloud. This shows that the language processing areas of the brain are involved in visual perceptual decision.

To sum up, cross-linguistic research on color categorization and brain research on color perception are constantly expanding subjects that need more meeting points. Here, we combine fieldwork and laboratory experimentation (using psychophysics and EEG) to investigate the neural changes produced by learning new color categories in a simplified model of language contact, in which speakers of different languages, with possibly different color categorizations, are forced to communicate using a common code, with a chance that one party will have to acquire the code of the other.

2. Fieldwork

We conducted field research to determine whether regional color meanings can co-exist in populations of language users. As part of a collaborative effort[1], we administered two color forced choice labeling tasks to Spanish (N=20) and Galician (N=20) speakers. Spanish and Galician are geographically contiguous languages that display a certain degree of semantic diversity. The analysis of our labeling data shows that differences exist in how the two languages set boundaries between some basic colors.

3. Laboratory experiments

3.1. The signaling paradigm

In the core part of the project, we study languages in contact via a variant of the signaling game (Lewis 1969; Skyrms 2010). Signals carry information of the state of world, which can be in different states. In a signaling game, there is a sender and a receiver and they both want the same outcome of the game, so they will both cooperative to arrive to a common code. The sender knows something about this state and selects a signal with which she will communicate this information. The receiver has to choose a response and he will get a reinforcement that will help him interpret the association between

[1]Evolution of Semantic Systems (EoSS). Consortium based at the Max Planck Institute for Psycholinguistics, Nijmegen.

the state (unknown to him) and the signal. Eventually, this signal becomes more reliable and sender and receiver start to coordinate, i.e. the game arrives to a stable solution.

In our studies, two players, a sender and a receiver, must converge on a common code to communicate efficiently. Semantic conventions, where signals are systematically mapped to meanings (a set of colors, i.e., a color category) by sender and received, emerge rapidly, typically under 20 plays (Blume et al. 1998). In our investigation, the sender role is played by a computer which sends signals to participants (receivers).

3.2 Simulations

We conducted a series of computer simulations of signaling games, aimed at determining the frequency with which each of the colors in the array should be presented to the sender in order for the receiver/participant to be able to learn color categories with the least effort.

3.3 Experiment 1. Rewiring color categories in the hue dimension

3.3.1 Pilot on color categorization

A group of volunteers of mixed nationalities performed a computerized analogue of the naming task that was used in fieldwork, this was a forced choice labelling task. The purpose of this pilot was to construct arrays of 5 colors with basic colors as extremes and ambiguous transition colors. Ambiguity was defined as a particular color being labeled with color term X and color term Y each with a frequency close to 0.5.

3.3.2 The main experiment

In Experiment 1, during two consecutive days, Italian speaking participants learned an artificial color system by playing as receivers in a signaling game with a computer. The stimuli consisted of an array of 5 Munsell colors that varied along the hue dimension from brown to green. The array was based on the results of the pilot study (see 3.3.1). At the end of day two, neural reorganization was assessed by means of EEG, showing participants color-

term (CT) and term-color (TC) sequences that were either learned, as part of the artificial color system, or incongruent. We found similar evoked responses to color terms in CT sequences and to colors in TC sequences, with larger late negative ERPs in incongruent than in learned trials, and smaller early effects following the same pattern. This EEG evidence for category-level color representations was supported by two independent color discrimination studies. Moreover, we found that ERP effects were largest, not for the best exemplars of each color category, but for the more ambiguous colors, suggesting that the strongest neural changes in the brain occur at the boundaries between categories.

3.4 Experiment 2. Rewiring color categories in the lightness dimension

3.4.1 Pilot on color categorization: the blues

A group of native English speakers performed the same computerized forced choice labeling task of Experiment 1, but this time with colors that varied in lightness within the blue hue. The purpose of this pilot was to construct arrays of 5 blue colors with a blue color in the center of the array (color 3) that was considered blue with a labeling frequency close to 1 by all English speakers, and an ambiguous blue-black color at the end of the array (color 5). Again here, ambiguity was defined as a particular color being labeled with color term X and color term Y each with a frequency close to 0.5.

3.4.2 The main experiment

In Experiment 2 English speaking participants took part of the study. It resembles Experiment 1 except the stimuli were 5 Munsell colors that varied along the lightness dimension within the blue hue and the most ambiguous color in the artificial categorization did not coincide with the most ambiguous color in their native categorization. We obtained larger late negative ERPs in incongruent than in learned trials only in CT sequences for the most ambiguous color in their native color categorization. These results suggest that ERP effects may reflect the effort of 'translating' the native categories into the new artificial categories, and this neural process of translation may require

more effort where the native categorization is more ambiguous. However, the fact that we did not see any effect in TC sequences, leaves open the question of how ambiguity affects the processing of colors and color terms.

Chapter 1

Color categories in the wild: A case study of Spanish and Galician

1. Introduction

Seized by philosophical bewilderment, we may ask ourselves what it would be like to be 'inside' someone else's body, that is, how another person experiences reality. In truth, philosophers went as far as asking what it is like to be a different form of life, e.g., a bat (Nagel 1974). Our mind may be especially overwhelmed by the attempt to imagine what it is subjectively for an organism to have echolocation. However, it is no less challenging to try to picture, for example, how a dichromate perceives the world beyond the usual traffic lights example, or how speakers of a language see different shades of blue depending on whether that language has one or two words for light and dark blue (more on this in Chapter 3). Light is sensed by all human beings with the same type of retina cells, perception is mediated by the same visual cortex, and everyone can discriminate differences in hues with about the same precision. Still, boundaries between color categories can vary across speakers of different languages.

Benjamin Whorf suggested that speakers are not led by the same physical evidence to the same way of perceiving the world, unless their linguistic experiences are similar or comparable (Whorf 1956). He used the Hopi language to illustrate this hypothesis. This language is a case in point as the Hopi people have a particular way of expressing the concepts of space and time: events that are far from the speaker are represented as having occurred in the distant past, whereas events that are closer in space are conceived as though they occurred closer in time. Hopi verbs have no grammatical tense but instead they are distinguished by their aspect (e.g., the extent of the event), the validity the speaker intends the statement to have (whether it is a report of the event or it expresses expectation regarding the event) and clause-linkage (the temporal relationship of two or more verbs) (Whorf 1956). In this

way, Whorf reckons, the Hopi language is framing the glasses through which the Hopi see and talk about their universe.

Returning to color terms, languages vary in the number of basic color words they have. Basic color terms are monolexemic (they consist of a single lexeme: 'blue' is a basic color, but 'baby blue' is not), they are used with relatively high-frequency in the language, and they are consistently agreed upon by speakers of the language in naming tasks. Also the application of these terms must not be restricted to a narrow class of objects (e.g. blond: human hair) and they must be psychologically salient for the speaker (this last one being the basis of the methodology of our study). Berlin & Kay (1969) showed that basic color terms form the basis of all color systems in the world. The number of basic color terms in a language can influence how easy it is for a speaker of that language to discriminate differences between colors (Kay & Kempton 1984). Roberson et al. (2000, 2005) studied speakers of three languages: English, Berinmo, a language of Papua New Guinea, and Himba. They found that the color stimuli that were in different color categories, depending on where each of the languages put the boundaries between one color category and the other, were remembered better, judged as being more different and word-color associations were learned better. Moreover, stimuli that straddle a category boundary are perceived as more distinct than equivalently spaced stimuli within a category (Kay 2006). Thus, despite universal tendencies in color naming across languages (Kay & Regier 2003), as color term boundaries vary, we may apprehend color differently depending on the language we speak.

If Whorf is right - or only 'half right' (Regier & Kay 2009) - then the language we speak influences our perception of color locally, i.e., at the boundary between two color categories. Here, the research agenda can be led down parallel paths, which correspond to the two main questions of this thesis. Consider two geographically contiguous and historically related languages, such as Galician and Spanish. The cultural divide between speakers of Galician and Spanish is nearly as small as a divide can be between speakers

of different languages. In spite of that, would these two languages set color boundaries differently? In other words, are different color vocabularies enough to support different color categorizations, or are 'deeper' cultural differences necessary? This is the first question raised in this thesis, and the focus of the present chapter. If such a purely vocabulary-driven effect on color categorization were to be found 'in the field', our next step would be to reproduce it in the laboratory. The second key question of the thesis, and the focus on Chapters 2 and 3, is: can learning of new color categories be induced simply through the acquisition of a new color vocabulary?

2. Methods

As part of the collaborative project Evolution of Semantic Systems (EoSS) that studies 50 Indo-European languages, we conducted field research to study how color categories vary in two geographically contiguous and historically related languages: Spanish and Galician. The aim of the EoSS project is to investigate how meanings vary over space and change over time, using a phylogenetic approach. It focuses on extensional semantics, i.e., how similar, or different are the referential ranges of words across languages.

The methods used here come from "Evolution of semantic systems procedures manual" (Majid et al 2011). There was a total of 6 tasks: the Color Blindness (Waggoner 2002), a test that evaluates the participants' color vision; Color Naming, in which participants were presented with a series of colors and were asked to name the colors they saw (Majid & Levinson 2007); A Focal Color task for which they were asked to point to the best example of colors in a list of color terms elicited in a pre-test (Majid 2008); a Spatial Relations Naming task, where the participant had to describe the relation between the figure and ground of a series of pictures, each one containing an orange figure object against a black ground object (adapted from Bowerman & Pederson 1992); a Body Part Naming task, in which the participant had to name a body part marked by a red dot (Jordan, Dunn & Majid 2009) and a Container Naming task, for which the participant had to name a container in a series of photographs of household containers (Ameel, Storms, Malt & Sloman 2005).

For the purpose of this study, here we focus on the first 3 tests: Color blindness, Color naming and focal colors.

2.1. Participants

Twenty native speakers of Spanish (14 females, mean age 24) and twenty native speakers of Galician (10 females, mean age 17) took part in the experiment after giving written informed consent. All participants' color vision was tested with a color blindness test (Waggoner 2002). All participants were trichromats. A biographical questionnaire was administered at the end of the session, developed by Dunn, Majid, & Jordan (2010) for the EoSS project. This questionnaire contained questions about their country of origin, the languages they spoke, their education level, and handedness (Oldfield, 1971).

2.2. Materials

2.2.1 Color blindness test

The test consisted of a booklet with 9 plates. There was one demonstration plate in order to explain the task. Six plates tested for the red-green color vision deficiencies. One of the plate tested for the type and degree of red-green defect and an additional plate tested for blue-yellow color deficiency. The materials were developed by Waggoner (2002).

2.2.2 Naming Task

The task was constructed using the Munsell (1912) color system. In this three-dimensional model, each color contains three attributes: hue, value and chroma. It is set up as a numerical scale with visually uniform steps for each of the three color attributes and each color has a logical and visual relationship to all other colors.

Hue defines pure color in terms of green, red, etc. and mixtures of two pure colors like red-yellow (orange), or yellow-green. It is dependent on its dominant wavelength. Chroma is the perceived strength of a surface color, the degree of visual difference from a neutral grey of the same lightness. It

has also been defined as "colorfulness" of an object relative to the brightness of a white object with the same illumination: colors of low chroma can be called “weak” and those of high chroma are highly saturated, strong or vivid. Value indicates the lightness of a color. The scale of value ranges from 0 for pure black to 10 for pure white. Black, white and the greys between them are called “neutral colors”. These colors have no hue. The stimuli consisted of 84 standardized Munsell colors, sampled with 20 equally spaced hues, 4 degrees of brightness, all maximum saturation, plus 4 achromatic colors. The colors were organized in a fixed random order. Saturation varied such that colors were generally at the maximal possible chroma for that point in the color space. This task was based on materials developed by Majid & Levinson (2007).

2.2.3 Focal colors task

The stimuli were the same 84 Munsell colors used in the naming task but here they were organized in a two dimensional array according to hue and lightness. Each color was identified by a letter according to the four levels of lightness (A-D) and a number that referred to the hue (1–20). The colors were shown as circles against a grey background within a rectangular plate. Participants were asked to verbalize the coordinates of the best example of each of the 10 basic colors terms the experimenter read out loud. To this end, we preliminarily constructed a list of basic color terms, using one of the attributes of the definition of “basic colors”, the psychological saliency to the informant, as a criterion. It is important to point out that if we had chosen a different set of criteria, we could have obtained different results. This list was taken from the responses to a free-listing task of 20 subjects (10 Spanish speakers and 10 Galician speakers) that did not take part in the main study. Participants were asked to list all of the colors they could think of, in the order they came to their mind, as quickly as possible, in a total of two minutes of time. Color terms were then ranked by how frequently they appear and their position in each list across participants. This task was based on materials developed by Majid (2008).

2.3 Procedure

We administered the two color naming tasks to speakers of Spanish at the Universitat Politècnica de Catalunya in Barcelona, Spain, and Galician at the IES nº1 de Ordes secondary school. Both testing rooms were quiet and there were good daylight lighting conditions, which were consistent across participants and between the two places. The entire session was recorded on an audio-recorder and then transcribed (see below).

The instructions for the tasks and questionnaire were translated into both languages from English. These translations were then back translated into English by a person who was not familiar with the study and corrected if necessary, to allow for a better cross-linguistic comparability. Before the tasks started, participants were given an overview of the project, and were given the opportunity to ask questions regarding the study.

2.3.1 Naming task

Participants were presented with each color individually on individual numbered plates and they were asked to name each color with the first term that occurred to them. In each trial, the experimenter asked in the participant's native language (Galician or Spanish): “What color is this?” Participants’ responses were audiotaped and later transcribed. If the participant gave a long description (e.g. “egg yolk or mustard yellow”, “strawberry red with more quantity of magenta”) they were asked to say it in another, possibly shorter, way. If at this point they did not offer a more concise answer, the experimenter proceeded to the next color.

2.3.2 Focal colors

The goal was to elicit the best exemplars for the basic color terms in the language. The instructions were the following: “Here are the same colors you labeled in the previous task. They are all laid out for you. Now, I will give you a color word and I want you to point to the best example of that color”.

2.3.3 Color blindness test

Participants' color vision was evaluated by a color blindness test (Waggoner 2002) to exclude a possible color vision impairment. All participants were tested after the forced choice labeling task and the Focal Colors task.

Participants were asked to name the numbers on the plates that were held about 50 cm from the participant, directly facing them. The participant had 3 seconds to identify each plate. Hesitation was taken as an indication of color deficiency.

2.4 Data analysis

Our main objective with this analysis was to understand how much similarity there was between color representations across Galician and Spanish. In particular we wanted to detect possible category-boundary differences between the two languages. An example of a category-boundary difference would be a particular Munsell color (or a small set thereof) that is labeled by speakers of one language using color term X (say, 'green'), and by speakers of the other language using color term Y (e.g., 'yellow'). Other types of boundary differences include cases in which, in one language, a particular Munsell color lies across two categories, and is accordingly labeled with approximately equal frequency as an X or a Y, whereas in the other language it is reliably labeled as either an X or Y. As can be seen, our analysis focuses on single colors: the use of a limited sample of colors in a forced choice labeling study makes it difficult to trace boundaries across categories using multiple colors; moreover, as will become clear in Chapters 2 and 3, our methodology for laboratory research (whose aim is to investigate color-boundary effects in the brain) requires the identification of small sets of colors for which clear boundary effects are found.

2.4.1 Naming task

The data was transcribed verbatim from the audio recordings. The full responses of the participants (e.g. greenish blue, light brown) and the main response ('blue' for greenish blue, 'brown' for light brown) were written down in the answer sheets. To analyze our data we only considered the main response ('blue' in 'greenish blue') as the main response and we used only this term for statistics, and not the rest of the response ('greenish'). In order to analyze if Galician and Spanish speakers used the same color terms to name the same Munsell colors, for each of the color terms that were part of the basic colors list, we counted the number of times each Munsell color was given that name, in each of the two language groups separately (e.g. which Munsell colors were given the name 'azul' and how many Galician speakers used that term for each of the colors). Then we computed the difference in the amount of times a certain color name was given to each Munsell color between the two languages, and the amount of different color names each Munsell color was given in each language. Some color terms are not shared by both languages (e.g. the Galician term 'vermello' is translated as 'rojo' (red) in Spanish, 'amarelo' is 'amarillo' (yellow), etc.) and some are ('verde', 'azul', 'negro', etc.). In the cases where the term used was not one that is shared by both languages, we compared this term to the equal term in the other language (e.g. Munsell colors named with the term "vermello" by Galician speakers were compared to Munsell colors named "rojo" by Spanish speakers) Furthermore, some Galician speakers also used some color terms in Spanish instead of the Galician term (borrowing) in some trials (due to contamination of Spanish on the Galician language ("blanco" instead of "branco", "amarillo" instead of "amarelo", "naranja" instead of "laranxa"). In some cases, terms

To be included in our analysis, a Munsell color C had to meet three criteria: (1) C should be named using the same color term at least twice in at least one language (this shows a certain level of agreement among participants); (2) That color term should have at least a 30% difference in count between the two languages, i.e., at least 6 out of the 20 participants should name C using that specific color term in one of the languages and not

in the other one (we take this difference as a significant one between the languages); (3) No more than 4 names should be used for that Munsell color in either language (if a color is named with more terms, it suggests us that instead of showing a difference between languages, it shows this is a very ambiguous color for one of both languages). Our prediction was that if one color was found for which only 2 color terms are used, it is very likely that that color is at the boundary between 2 colors. We ran *t*-tests on all the Munsell colors that met the criteria to study whether and how the labeling patterns for each color differ from one language to another, i.e. if speakers put the boundaries between one color category and another one in a different point of the color spectrum. We also calculated for each language how frequently speakers used each color term for each Munsell color. If a Munsell color is named with the same term with a high frequency across speakers of the same language and with the cognate color word across languages, we can say it is a color that is prototypical for that color term across languages. If the term is used for a specific color with a high frequency only in one of the languages, we consider that Munsell color to be prototypical for that color term in that language. If instead a Munsell color was named with several different terms, we can say that color is ambiguous and we would expect speakers not to agree easily on the correct name for that color. This information allowed us to build the stimuli for the experiments we will explain in the next chapters.

We used Hamming distances, the proportion of positions at which two aligned sequences differ, and we estimated the variability between, within and across groups. We wanted to further measure the similarity of Galician and Spanish speakers' color space (Pinheiro, Seillier-Moiseiwitsch & Sen 1998) and to check if the colors which to the speakers represent a certain color term is the same across languages. We analyzed the color terms "azul", "amarillo/amarelo" and "verde". For each language we replaced all the subjects' responses "azul", to the number 1 and converted all other responses to 0. If we consider the list of Munsell colors with the responses 0 or 1 a vector space, a Hamming distance analysis allows us to see what is the number

of entries in the vector for one person or language group that would need to be changed in order to make it identical to the vector of the other person or language, or, the minimum number of substitutions required to change a color categorization of a speaker into another of another speaker. We ran *t*-tests to explore how consistent the responses used for each Munsell color were within the group (internal similarity) and how similar/different the Galician and Spanish speakers groups were in labeling the colors (cross similarity). We run the same procedure for terms "amarillo/amarelo" and "verde". In this way we had vectors of 0s and 1s as responses instead of the color terms.

2.4.2 Focal colors

The subjects were then asked to point to the best examples corresponding to 10 color terms. We then said the letter and number out loud and the responses were transcribed verbatim from the audio recordings. If the participant gave two different responses or hesitated, they were asked to confirm their response. If he/she said there was no match, we asked them to choose the closest matching color.

3. Results

When comparing the number of Spanish speakers and Galician speakers that named each Munsell color with each color term (or the cognate), two Munsell colors showed to be different in Spanish and Galician (Fig. 1).The Munsell color 5G 8/6 ($t(19)=-3.559$, $p=0.0021$) was one. It was named "azul" (blue) by 40 per cent of the Spanish speakers and by none of the Galician speakers, and it was named "verde" by 60% of the Spanish speakers and by 100% of the Galician speakers. The Munsell color 10Y 6/10 ($t(31.204)=-2.6261$, p-value=0.01327), which was named "verde" by 90% Spanish speakers and by 55% Galician speakers, and it was called "amarillo" (Spanish for yellow) or "amarelo" (Galician for yellow) by 5% Spanish speakers and by 35% Galician speakers ($t(26.603)=2.4936$, p-value=0.01917). It is important to point out that 6 Galician speaking participants also used the Spanish color term "amarillo" instead of "amarelo" to name the yellow Munsell colors.

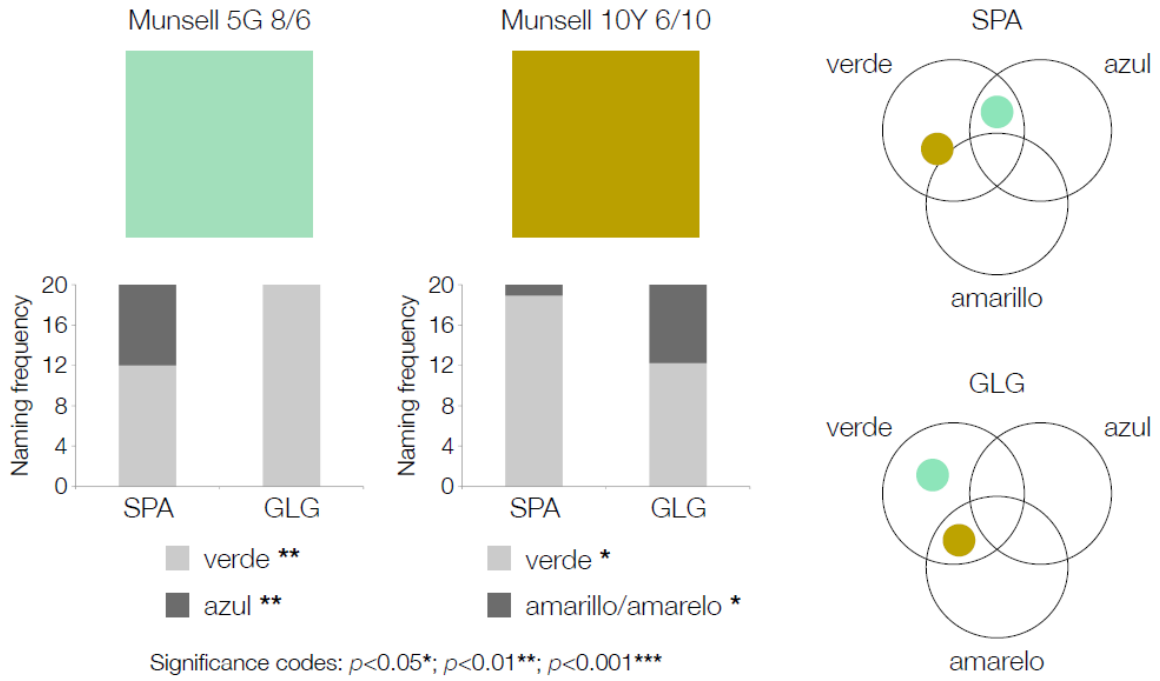


Figure 1. A) Munsell colors 5G 8/6 and 10Y 6/10. Below each color, the number of participants (labeling frequency) who used each color term in Spanish (SPA) and Galician (GLG). B) Diagrams of the use of the different color terms for the Munsell colors in each of the languages. Circles represent color categories, which are here assumed to have mutual intersection zones. The colors represented here may render inaccurately the actual Munsell colors.

To further explore the differences among these color terms, we analyzed the internal similarity of labeling patterns within each language group and the cross similarity between the two language groups using Hamming distances.

We calculated the internal similarity of Spanish speakers using the mean of the Hamming distances between all pairs of vectors X and Y (where X is the responses of one Spanish speaker, and Y is of another). That is, we measured the Hamming distances between the responses "azul" ("azul" coded as 1 and all other possible responses coded as 0) to a certain Munsell color of one participant and of another, and we repeated the same procedure with all pairs of participants. In the end computed the mean of these distances. A low mean

value suggests that the group of Spanish speaking participants was internally consistent in their responses. The same procedure was performed on the Galician speakers group. We also calculated the cross similarity between Spanish and Galician by comparing the means of the distances of pairs of Spanish speakers and Galician speakers. Finally, we compared the internal similarity of Spanish and Galician by means of a *t*-test. Afterwards we compared the internal similarity of Spanish to the cross similarity with Galician and the internal similarity of Galician to the cross similarity with Spanish also using *t*-tests.

Spanish and Galician speakers differed in their consistency in labeling colors “amarillo/amarelo” ($p<0.001$). On the other hand, they are similarly consistent in how they use the word "azul" ($p=0.527$) and “verde” ($p=0.018$). But then there are differences across the two languages: Spanish speakers were similar in their usage of "azul" ($p=0.0015$) and “amarillo/amarelo” ($p<0.001$) to other Spanish speakers than when we compared them to Galician speakers, but this was not true for “verde” ($p=0.341$). When we analyzed Galician speakers, we found they were more similar to other Galician speakers in the way they used “azul” ($p=0.0006$) and “verde” ($p<0.001$), than when we compared them to Spanish speakers, but not so much for “amarillo/amarelo” ($p=0.052$).

We found the color term “azul” to be especially interesting in the way it is used: the internal similarity of participant responses of Galician and Spanish is comparable (there is no statistical difference within each language group), but there are cross similarity differences in both directions (See Table 1). Thus, Spanish and Galician seem to use the same word "azul" for slightly different colors.

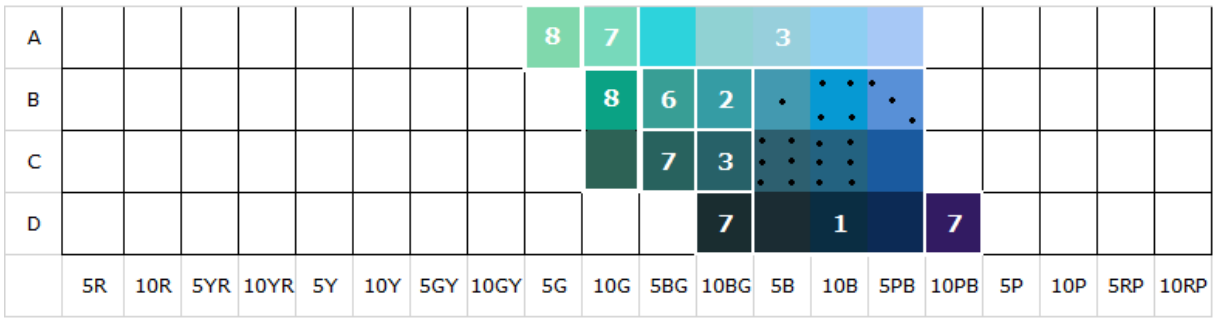
Figure 2 shows a general picture of the differences/similarities between Spanish and Galician for colors “amarillo”/”amarelo”, “azul” and “verde”.

	Amarillo/Amarelo	Azul	Verde
Internal similarity (IS) difference between SPA and GLG:	$t(19)=-9.0285$ $p < 0.001$	$t(19)=-0.6446$ $p=0.527$	$t(19)=2.5926$ $p=0.018$
Cross similarity (CS):			
IS SPA/CS GLG	$t(19)=-14.6865$ $p < 0.001$	$t(19)=-3.6989$ $p=0.0015$	$t(19)=-0.9759$ $p=0.341$
IS GLG/CS SPA	$t(19)=2.0673$ $p=0.052$	$t=-4.0936$ $p=0.0006$	$t(19)=-6.1436$ $p < 0.001$

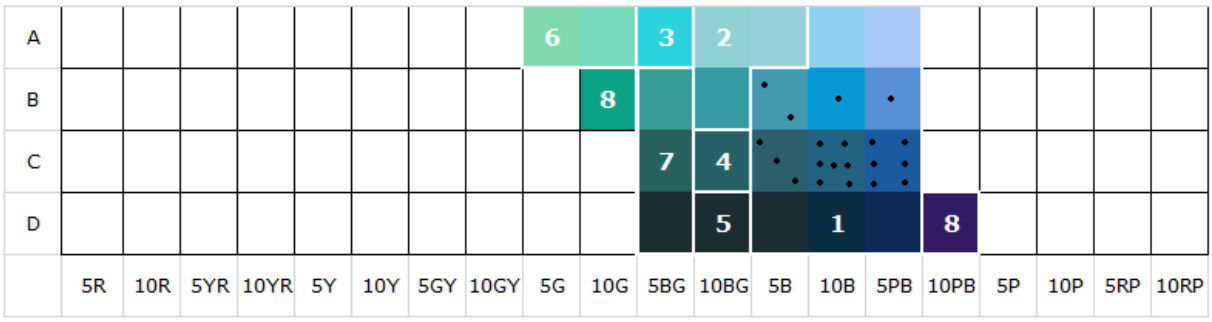
Table 1. Differences between Galician and Spanish. The internal similarity in using color "amarillo/amarelo" ($t(19)=-9.0285$, $p<0.001$) and "verde" ($t(19)=2.5926$, $p=0.018$) of the Galician speakers group was statistically different from the internal similarity of the Spanish speakers group. The opposite was true for "azul" ($t(19)= -0.6446$, $p=0.527$). Spanish speakers were more similar in their usage of "azul" ($t(19)=-3.6989$, $p=0.0015$) and "amarillo/amarelo" ($t(19)=-14.6865$, $p<0.001$) within their group than when compared to Galician speakers, but not of "verde" ($t(19)= -0.9759$, $p=0.341$). Galician speakers were more similar in their usage of "azul" ($t(19)=-4.0936$, $p=0.0006$) and "verde" ($t(19)= -6.1436$, $p<0.001$) within their group than when compared to Spanish speakers but not so much in their usage of "amarillo/amarelo" ($t(19)= 2.0673$, $p=0.052$).

Usage of the word “azul”

Galician

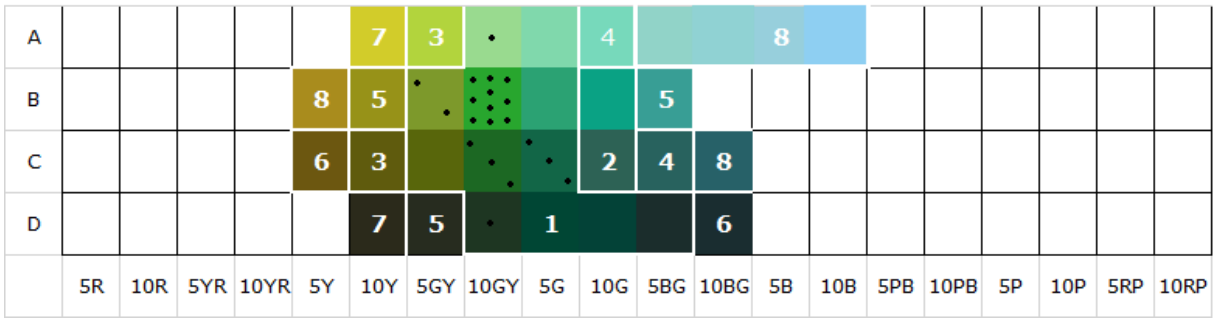


Spanish

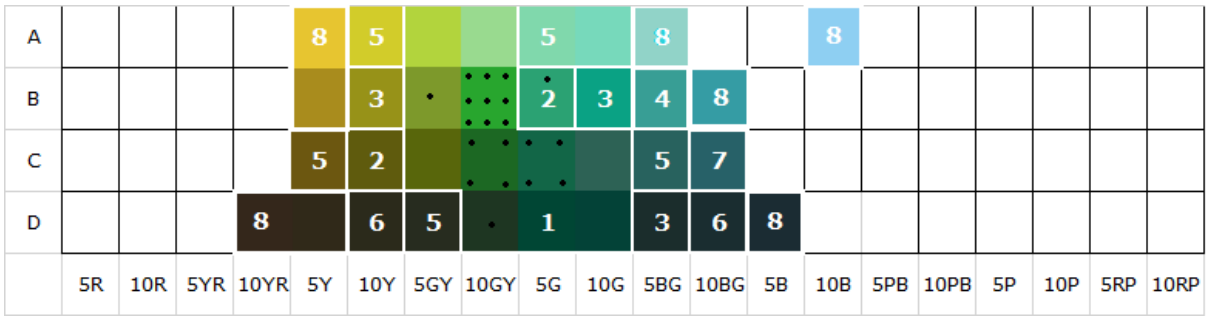


Usage of the word “verde”

Galician



Spanish



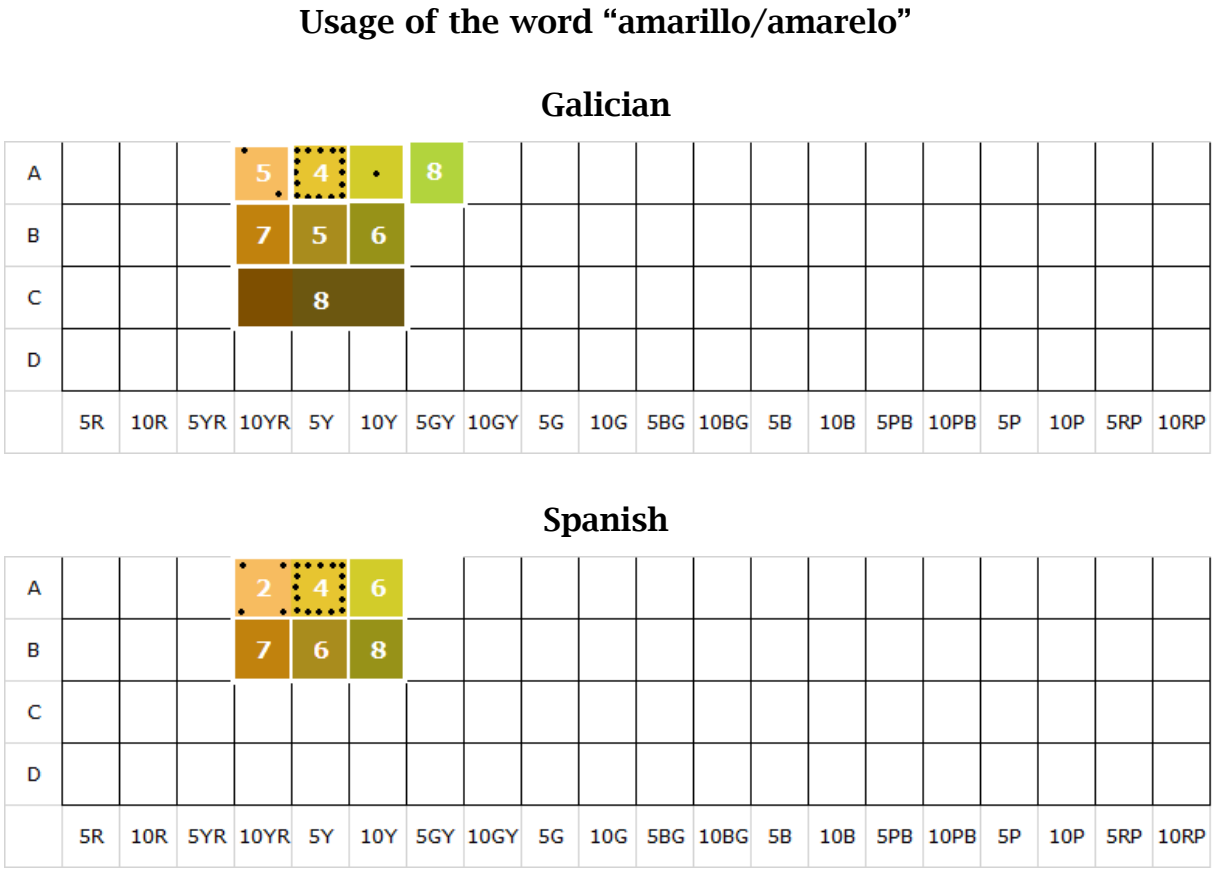


Figure 2. Array of colors used in the focal colors task. Only the Munsell colors for which the word "azul"/"verde"/"amarillo/amarelo" (both terms were merged as one in the case of Galician) was used in each of the languages are shown. Each color was identified by a letter according to the four levels of lightness (A-D) and a number that referred to the hue (1–20). Here we show in the horizontal axis the Munsell codes of the different hues. The numbers (1-8) inside the cells represent different clusters of colors. The clusters that contain the number 1 refer to the Munsell colors that were named with the specific color term by 100% of the speakers. The colors in cluster 2 were called with that color term by 95% of the speakers, cluster 3 by 80 - 90 % of the speakers, cluster 4 by 65-75%, cluster 5 by 50-60%, cluster 6 by 35-45%, cluster 7 by 20-30% and cluster 8 by 5-15% of the speakers. The dots on the grid represent the number of subjects who designed that color as best example of the category in the focal colors task. For Galician, in the focal color task the color term for which the participants had to point the best exemplar was "amarelo". R, red; Y, yellow; G, green; B, blue; P, pink. The RGB colors presented here to depict each Munsell color are only approximate conversions.

4. Discussion

Could a Spanish speaker consider the leaves of a tree in Autumn to be still green and a Galician speaker next to her say they are already yellow, about to fall, just because they speak different languages? Consider a group of tourists from different parts of the world, where different languages are spoken, on an island observing the ocean in a clear sky day. Would they all look at the horizon as the boundary between one color category and another?

In this study we analyzed how speakers of Galician and speakers of Spanish labelled 84 different Munsell colors in a forced choice labeling task. We computed the labeling frequency of each color and the quantity of names each color was given in each language. We also studied how participants of both languages responded to a task in which they were asked to point to the best examples of 10 color terms on a grid containing the same 84 colors. Finally, we studied the similarities within and across language groups in labeling the Munsell colors.

The analysis of our labeling data revealed differences in how the two languages set boundaries between the categories of yellow and green (hue), and within the blue category (lightness).

To sum up, contrary to Abrams & Strogatz's (2003) conclusion that societies in which two languages coexist are mostly split populations that lived without much interchange in separate monolingual groups, our results show that regional color meanings can co-exist despite interactions between populations (Abrams & Strogatz 2003). Galician is an Indo-European language spoken in the northwest of Spain, in Galicia Autonomous Region. It is closely related to Portuguese, with which it originally formed a unique language: Galician-Portuguese. Due to a political division of the territory, Galician and Portuguese became two different languages, although they have an 85% mutual intelligibility (Hall 1989). Galician and Spanish now coexist in the north of Spain and they are spoken in two connected populations. Even though spoken by the majority of the Galician population, Galician language is under threat from Spanish language contamination (relexification, lexical borrowing,

language convergence, etc.). However, the fact that regional meanings survive shows that there is still some degree of separation in how the two populations use language.

If the Spanish speaker and the Galician speaker who are looking at the tree had to reach to an agreement of what color the leaves are, or if they were exposed for the first time to each other's language, and the Galician had to explain what "azul" means for her using only signals, would there be any differences in the way they process these colors when they see them since they started categorizing them differently? Would there be any changes in their brains were they exposed to these colors after their learning interaction? The conclusions of this first study took us to further research into the effects of language contact in the brain.

With the laboratory experiments we will explain in the next chapters, we wanted to force a change in the native color categorization of people to model the scenario in which two languages get in contact and thus make the language evolve, with the final aim of studying the brain processes underlying this phenomenon.

Chapter 2

Rewiring color categories in the hue dimension: An EEG study

1. Introduction

A botanist sails around among islands inhabited only by natives whose language and culture he is ignorant about. On one trip he suffers from poisoning by eating seafood. He knows he needs the green root of a certain kind of plant to get some relief and be able to resume his travelling. On the island where he happens to be now, that plant is rare, and he must therefore rely on the natives to locate a place where the plant is to be found, and where the soil is humid such that green roots can grow. The natives use what appears to be a color word: 'duwi', to inform the botanist about the roots in a faraway area of the island. How is he to know what 'duwi' means, and whether it is the particular shade of green he is after?

This story echoes the classic thought experiment by the philosopher Quine (1960), who argued that solutions to similar cases of 'radical translation' are 'near miracles'. The meaning of 'duwi' is for the botanist hopelessly indeterminate, and his attempt to find green roots is fraught with uncertainty. A different answer was given by Lewis (1969), who showed it is possible for speakers to arrive at a shared meaning without prior agreements. If the native utter 'duwi' and point to a given area of the island, the botanist should be able to determine whether 'duwi' means brown or green by simply examining the roots there. If the procedure is iterated a number of times, the botanist may eventually refine his understanding of 'duwi', and find the roots he needs. For Quine communication between the botanist and the natives is as indeterminate as the meaning of their signals, whereas for Lewis communication can be fruitful so long as the parties can coordinate their choices, i.e., the signals the natives use to name certain colors, and the actions the botanist undertakes in response to the signals he receives. These are the basic elements of 'signaling games', devised by Lewis (1969) to model coordination among agents with common interests (Skyrms 2010).

Communication games similar to signaling have been used as a basis for computational models of the evolution of color systems. Komarova et al. (2007) showed that a color category system emerges in a population of communicating agents endowed with minimal perceptual discrimination capabilities, subject to pragmatic constraints such as 'hotspots' (e.g., colors signaling ripe fruit, or fresh roots). This work was later refined by Jameson & Komarova (2009a, 2009b). In a more parsimonious model, embodying a universal color discrimination principle, but no pragmatic constraints, Baronchelli et al. (2010) found that a population of communicating agents is capable of evolving a color category system that exhibits universal properties similar to those of the world's languages (Kay & Regier, 2003). Contrary to Quine's 'indeterminacy of translation' thesis, not only is coordination among pairs of speakers possible (e.g., between the botanist and the natives on the meanings of 'duwi' and 'green'), but the very same communication processes may lead to the emergence of fully fledged color lexicons in populations.

The computational models discussed above scaled up the problem of the emergence of shared color categories from the dyad, as in Quine or Lewis, to the population. Here, we wish to contribute to the current understanding of the same problem by moving in the opposite direction: from the dyad to the individual. We aim to investigate neural changes following the acquisition of color categories and vocabularies via communication games. We consider a limited portion of the color space: an ordered array of 5 colors in the green-brown continuum (Fig. 1). We use 2 artificial color terms, i.e., the monosyllables 'nu' and 'to', to partition the array into 2 categories, each containing either 2 or 3 colors. In our experiments, we employ all 4 possible categorizations, i.e., 'nu' for colors 1-2, and 'to' for colors 3-5; 'nu' for 1-3, and 'to' for 4-5; etc. Each participant learns one of these categorizations by playing a signaling game (Lewis 1969; Skyrms 2010) as the receiver; the sender role is taken by the computer. Roles are fixed at the start of a game. In recent computational and experimental work (Moreno & Baggio, 2014), we found that, in signaling games with fixed roles, in which both players are human volunteers, the receiver adjusts his code to that of the sender: coordination is achieved by unilateral agreement on the sender's code. The present

experimental setup, in which the sender/computer sticks to its code throughout a game, and where it falls to the receiver/participant to learn that code over signaling rounds, is therefore not implausible as it captures a key property of signaling games with fixed roles: that a code is transmitted from the sender to the receiver.

In the present study, once training in the signaling game is completed, and the participant has learned one of the 4 color categorizations, we assess neural changes using EEG. Participants are presented with visual sequences of either a color followed by a color term (color-term, or CT sequences), or a color term followed by a color (TC sequences). Each color-term and term-color pair can be as in the learned categorization, or incongruent with the learned color system. Event related brain potentials (ERPs) are obtained for the second stimulus in each sequence, i.e., the color term in CT sequences, and the color in TC sequences.

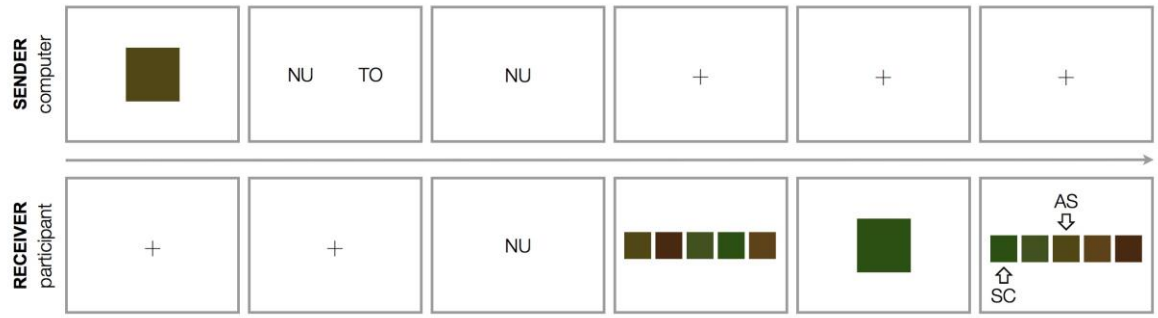


Figure 1. A trial of the signaling game used in the behavioral training sessions. The top and bottom rows show what the sender/computer and the receiver/participant see on successive screens, respectively. Roles are fixed at the start of the game. Time flows from left to right. The sender/computer plays according to a fixed categorization of the 5-color array, which must be learned by the receiver/participant. In each trial, the sender sees one color, and sends the signal associated to the category the color belongs to. The receiver sees the signal, and must choose from the 5-color array the one that may have been seen by the sender by pressing one of five keys on a standard full-size keyboard. The feedback shown to the receiver indicates, by means of two sets of arrows and labels, the color associated to the signal (AS) and privately seen by the sender, and the color chosen by the receiver as a response to the signal (SC). Over trials, the receiver will learn what colors are associated to each signal, i.e., the color categorization the sender is playing by.

We aim to test two hypotheses. The first concerns the perceptual/semantic nature of color category representations. Fonteneau & Davidoff (2007) reported an ERP correlate of the categorical perception of colors. They found the peak latency of ERPs in response to a small color difference bridging a color category boundary (195 ms) to be quicker than that for an equally-sized color difference not bridging a category boundary (214 ms). Moreover, differences for the between-category appeared at 160–200ms and data from within-category deviancy showed longer ERPs (160–280ms) which they think is related to a more difficult discrimination due to a psychologically compressed color space. These early ERP effects are suggestive of a perceptual basis for color categories. Using a visual oddball task, Holmes et al. (2008) reported earlier latencies of P1 and N1 components for between- relative to within-category deviant stimuli, and enhanced P2 and P3 waves, thought to reflect post perceptual processing. In an fMRI study, Kwok et al. (2011) showed that learning associations between artificial color terms and color categories increases grey matter volume in regions V2/3 of the left occipital cortex. Importantly, in none of these studies is the communicative dimension of color category learning very prominent. Therefore, a possibility is that learning color categories in a communication game like signaling may rewire semantic rather than perceptual systems. If that was the case, one might expect later ERP components, for instance the N400 (Kutas & Hillyard, 1980), to be modulated by processing colors or color terms inconsistent with the learned color system.

A second hypothesis concerns the locus of incongruity effects in ERPs. One possibility is that terminal colors in the array are represented as the best exemplars of each color category, so that the strongest incongruity effects are observed there. The weakest ERP effects would be seen for the middle color of the array, which is the most ambiguous. Another possibility is that ERPs indirectly reflect learning effort, which is assumed to be higher at the boundary between color categories. The strongest incongruity effects should then be observed for the 3 intermediate colors, and terminal colors should show weaker ERP effects, because it would be easier to assign them to either category during learning.

Session	Task	Methods	Results	Tables	Figures
Day 1	Colorblindness test	2.7	-	-	-
	Labeling	2.1	3.2.1	-	-
	SG training	2.8	3.2.2	-	2
Day 2	SG training	2.8	3.2.2	-	2
	EEG session	2.9	3.3	3	3
Day 3	Colorblindness test	2.7	-	-	-
	Labeling	2.1	3.2.1	-	-

Table 1. Structure of the main study: organization of the tasks on three successive days. The table indicates the Methods section in which each task is described, and the Results section and the Tables and Figures in which results are reported.

The main study was carried out in 3 sessions over 3 consecutive days. The structure of the main study is outlined in Table 1. On the first day, we assessed participants' categorization of colors, including browns and greens, in their native language (Italian) by means of a forced choice labeling task. The aim was to determine whether the native and learned artificial color categories are compatible or otherwise. The first behavioral training session with signaling games also took place during the first day. On the second day, participants received further training with signaling games, on the same color categories they were trained on during day one, and they were then tested in a passive exposure task of CT and TC sequences while the EEG was being recorded. On the third day, the forced choice labelling task was repeated to determine whether learning artificial color categories affected native color categories. Besides, we conducted two sets of studies. The first was a series of computer simulations of signaling games, aimed at determining the frequency with which each of the colors in the array should be presented to the sender in order for the receiver/participant to be able to learn color categories with least effort. Furthermore, we administered two color discrimination tasks to an independent set of participants, who were also trained with signaling games

on one of the 4 categorizations, on two consecutive days, to determine whether via signaling games participants learn color categories, as opposed to mappings between colors and color terms. On each trial, participants saw a pair of colors, and had to judge whether the colors belonged to the same or a different category. Here the prediction is that responses should be faster for across-category trials compared to within-category trials (Winawer et al., 2007).

2. Methods

Below we first report on how the 5-color array used as a stimulus set in the main behavioral and EEG experiments was constructed based on results from a forced choice labeling study (2.1). Further, we describe a series of computer simulations aimed at setting stimulus presentation parameters so as to optimize learnability of color categories in the signaling game (henceforth SG; 2.2). Next, the structure on three successive days (2.3; Table 1), the participants (2.4), and the materials and methods (2.5-2.9) involved in the main experiments are described. Last, we present the method of two color discrimination experiments intended to demonstrate that SGs induce learning of color categories, as opposed to associations between colors and color terms lacking a set level representation (2.10).

2.1. Constructing color stimuli: forced choice labeling study

We conducted a forced choice labeling study to identify five colors that would form an ordered array with instances of two of the basic colors red, orange, brown, green or yellow as extremes, and transition colors in between. The resulting array would sample a uniform portion of color space varying along the hue dimension, at the boundary between two color categories. The array was intended to serve as a stimulus set in our behavioral and EEG experiments. Crucially, in our computer simulations and experiments, whether the middle color in the 5-color array belongs to one or to the other category is entirely determined by our training protocol with SGs. Therefore,

all 4 categorization possibilities, with each category comprising either 2 or 3 colors, involving the 5 colors in the array and 2 artificial color terms, could be included in the experimental design. The aim of this forced choice labeling study was to identify one middle color labeled with 2 basic color terms with similar frequencies (ideally ~0.5), and 2 terminal colors, each labeled with a different basic color term with high frequency (ideally ~1). Two more transition colors were found by interpolating the terminal colors with the middle color in color space. The results of the forced choice labeling study are reported in section 2.5.

2.1.1. Participants

Sixty native speakers of Italian (N=10), Russian (N=10), English (N=10), Chinese (N=10), Persian (N=10), and Argentinian speakers of Spanish (N=10) performed the task (mean age 27.32; 29 female). Participants had normal or corrected-to-normal vision. Their color vision was evaluated by the Ishihara Colorblindness Test (2.6).

2.1.2. Materials

The stimuli consisted of 65 Munsell (1912) colors varying in the chroma dimension in the red-yellow-green range. Colors were selected using a standard Munsell color chart. Translation to the RGB systems was done using a standard conversion table.

In all experiments we used the same hardware and software settings. The stimulus presentation monitor was a 22-inch LCD Samsung Syncmaster 2233RZ with 120 Hz refresh rate (frame duration 8.33 ms). The screen resolution was 1680×1050 pixels, and each pixel subtended ~1.7 arcmin. The minimum and maximum luminances of the screen were 0.19 and 134 cd/m2. Luminances were measured using a Minolta LS-100 photometer. A gamma-corrected lookup table (LUT) was used to ensure that luminance was a linear function of the digital representation of the image.

2.1.3. Procedure

Participants sat in front of the computer monitor and a standard full-size keyboard (viewing distance ~80 cm). Their heads were fixed by means of a chin rest. In each block, participants saw each of the 65 colors 10 times in random order. Each color was presented for 400 ms, followed by a mid-grey screen (RGB [128 128 128]) for 1750 ms. each color was shown as a square (size 120px × 120px) in the center of a mid-grey screen. The task consisted in pressing one of five keys to label the color just shown. The keys were labelled with either the first letter (in English, Spanish, Italian and Russian), the first two letters (in Persian) or the ideogram (in Chinese) of each basic color term in a language. The position of the labels on the keyboard was randomized and counterbalanced across participants. Participants had to use their index finger to answer and, after having pressed the key, to return with the finger to a marked point outside the keyboard equidistant to all five keys.

2.2. Learnability of color categories: computer simulations

We designed an algebraic computational model of the SG modified from Moreno & Baggio (2014). At the start of the game, a language is created, consisting of color categories, grouping together the 5 colors from the array in subsets of 2 or 3 adjacent elements (i.e., empty or singleton categories and categories containing 4 or 5 colors were excluded), and a color vocabulary, mapping each of 2 color terms to either set of colors. In each trial (Fig. 1), a color is drawn from the 5-color array with a certain probability (see 2.2.2), and is privately shown to the sender. The sender then selects the color term corresponding, in the language, to the category the color belongs to. The color term is sent as a signal to the receiver, who chooses one of the colors that he believes are associated with the color term in the sender's language. Feedback is provided as to whether the color as seen by the sender and the color chosen by the receiver in response to the color term match. The task for the receiver is to infer the color set (i.e., the category) assigned to each color term in the

language. SGs in the computational model have the same formal structure as in the experiments (2.8).

In the model, players are understood as 'mindless' agents with no capacity for piecewise remapping of colors to color terms if negative feedback is given, and no memory of past choices (for a computational study of 'memoryless' learners in color categorization games, see Komarova et al., 2007; for further motivation, see Moreno & Baggio, 2014). Remapping colors to color terms occurs as a random permutation and reassociation of elements from the 5-color set to the elements of the 2-term set. The receiver is endowed with the capacity to retain the current code if it matches with the sender's language. The aim of these simulations is threefold: (1) to determine whether color categories are learnable despite (a) the chosen form of feedback and (b) players' lack of intelligent or strategic behavior; (2) to set some performance benchmarks for accuracy and learning times, with which to compare human performance data; (3) to determine the optimal distribution of frequencies of colors as drawn from the 5-color set, so that learning times for the receiver are minimized; this frequency distribution will be used to set stimulus presentation parameters in the training protocol with SGs on human volunteers.

2.2.1. Model design

The model is written in MATLAB code and performs linear algebra operations on one set of two-column numerical matrices (mapping colors to terms) representing the player's code in different trials of the game. A similar 5-by-2 matrix represents the sender's language, which is fixed at the beginning of the game. The receiver's mapping is initialized by randomly drawing and pairing two elements: one from the color set, and one from the term set. Colors are eliminated from the color set as soon as they enter the mapping. Terms are removed from the term set if and only if they have been used k times, where k is the ratio between the number of colors and the number of terms in a game, rounded up to the next integer (i.e., here k=3). The use of k as defined here allows our model agent to attain optimal accuracy, i.e., the

average accuracy if the receiver, upon seeing the signal, chooses at random one of the 2 (or 3) colors given one (or the other) term. With equiprobable colors as drawn from the color set, optimal accuracy by random choices is 0.40, that is, 2/5*1/2 (for the color term associated to 2 colors the chance of success is 1/2) plus 3/5*1/3 (for the color term associated to 3 colors the chance of success is 1/3).

Once the language and the receiver's initial code are set, the game begins. A color is drawn with a certain probability (details in 2.2.2) from the color set, and the corresponding color term is sent to the receiver. The receiver picks one of the (2 or 3) colors corresponding to the color term in its initial mapping. If the colors are the same, positive feedback is given, and the receiver proceeds with its current code. If colors differ, negative feedback is given. The receiver adjusts its current mapping as follows: a random permutation of the color set is performed, and a new association with color terms is computed in the same way as the initial code was. By chance, after a finite number of trials, the receiver will learn the sender's language. We define the learning point as t+s, where t is the first trial at which the receiver's code de facto coincides with the sender's, and s is the number of colors in a game. That is, we assume that the receiver will know he has learned the language after he has played with the final code in at least as many trials as there are colors.

Color frequency (1;..;5)	DL	OR	OI	RDL	ROR	ROI	S
0.20; 0.20; 0.20; 0.20; 0.20	73.75	0.4	0.4	3	7	7	17
0.15; 0.20; 0.30; 0.20; 0.15	71.18	0.39	0.5	6	6	6	18*
0.10; 0.20; 0.40; 0.20; 0.10	70.37	0.38	0.6	7	5	5	17
0.05; 0.20; 0.50; 0.20; 0.05	74.29	0.38	0.7	2	4	4	10
0.05; 0.15; 0.60; 0.15; 0.05	72.36	0.37	0.75	5	3	3	11
0.05; 0.10; 0.70; 0.10; 0.05	72.44	0.36	0.8	4	2	2	8
0.01; 0.09; 0.80; 0.09; 0.01	75.02	0.36	0.89	1	1	1	3

Table 2. Results of simulations of color categorization in signaling games with fixed roles. Each row is a different simulation cycle. The leftmost column shows the frequency with which each color is drawn from the array in each cycle (details in 2.2). Abbreviations: DL (duration of learning in trials); OR (optimal accuracy with random choices); OI (optimal accuracy with informed choices); RDL (ranking of DL values); ROR (ranking of OR values); ROI (ranking of OI values); S (rank sum score). The optimal parameter setting (i.e., the one with highest S) is marked with an asterisk (*).

2.2.2. Simulation cycles

The simulations were run on a laptop computer mounting a 2.9 GHz Intel Core i7 processor. We conducted 7 separate simulation cycles testing different frequency distributions on the 5-color set (Table 2). In the first cycle, all colors had the same probability (i.e., 0.2) of being drawn from the set. In each subsequent cycle, color frequencies were manipulated such that the middle color had increasingly higher chance, and terminal and transition colors had increasingly lower chance, of being drawn from the set. Each simulation cycle comprised 10000 games with 1000 trials per game. For each cycle, we computed: the duration of the learning phase in trials (DL; the time it takes for the player to learn the language); the optimal accuracy by random choices (OR; the highest accuracy the player can achieve if, upon observing a color term, she chooses at random among the 2 or 3 colors that may be associated to it, ignoring possible non-uniform frequency distributions of colors); the optimal accuracy by informed choices (OI; assuming that the receiver has access to frequency distributions on the colors: the highest accuracy he can attain by choosing the most probable of the 2 or 3 colors associated to the term he has observed). This resulted in 7 different values (one per cycle) of each of these three variables. Variable values were ranked according to three optimality criteria: (1) DL must be minimized, i.e., the duration of the learning phase should be as short as possible (a shorter learning period corresponds to a higher rank); (2) OR should be maximized, i.e., the optimal accuracy by random choices should be as high as possible: high accuracies should be attainable by selecting at random one of the colors associated to a given term (a higher accuracy corresponds to a higher rank);

and (3) OI must be minimized, i.e., optimal accuracy by informed choices must be as low as possible, as there should be no incentive to learn the frequency distribution of colors (a potential confound with learning color categories) and to use the strategy of choosing the most likely color given a term. The optimal parameters shorten learning periods and increase player accuracy on the assumption that he should have no access to the frequency distribution on colors as drawn from the color set, and that he should be able to play successfully by random choices. The optimal model parameters will be used to set stimulus presentation parameters in the behavioral learning protocol (2.8).

2.2.3. Optimal color frequencies

As reported in Table 2, the optimal frequency distribution over colors drawn from the 5-color set is: 0.15 for each of the terminal colors, 0.20 for each of the transition colors, and 0.30 for the middle color. This frequency pattern optimizes learning for the receiver, as learning time is minimized, the accuracy he can obtain by choosing randomly colors of the same category in response to each signal is maximized (i.e., he need not know the frequency distribution of colors), and the accuracy he would obtain with informed choices is minimized (i.e., there is little incentive to learn the color frequency distribution). This result was used to set the frequencies of colors as shown to the sender/computer in the SG training in the main experiment (2.8) and in the color discrimination studies (2.10).

2.3. Structure of the main study

The study took place during three consecutive days and were organized as follows (Table 1). On Day 1, participants were administered the Ishihara colorblindness test (2.6), followed by the forced choice labeling task (2.7) and by the first training session on artificial color categories with SGs (2.8). On Day 2, the second training session with SGs was followed by the EEG session

(2.9). On Day 3, the colorblindness test and the forced choice labeling task were repeated.

2.4. Participants

Twenty-five right handed native speakers of Italian took part in the study after giving written informed consent. Five volunteers were discarded either after EEG during Day 2 due to equipment failures or after preliminary data analysis because the data were contaminated by artifacts (i.e., more than 30% of EEG segments had to be discarded). Twenty volunteers (9 females, mean age 23.35) were included in the final data analysis. All participants had normal or corrected-to-normal visual acuity. Color vision was evaluated by means of the Ishihara colorblindness test (2.6). Participants received 25€ for participating. The study was approved by the Ethics Committee of SISSA.

2.5. Materials

From the labeling data (2.1), we identified two colors that were prototypical for two basic color categories. These were the colors that were labeled using the same color word across participants, and with the cognate color word across languages. In the case of Chinese, since it is not an Indo-European language like the rest of the languages in this experiment and hence there are no cognate words, we used the Chinese translation-equivalent terms for the different labels from English. These were the terminal colors in the array. Further, we identified a color that was labeled with two color words with approximately equal frequency across languages. This was used as the middle color in the array. Finally, we identified two transition colors by interpolating the terminal and the middle colors in color space.

2.5.1. Selection of terminal colors

We selected all the colors labeled with each of the 5 basic color terms (red, orange, brown, green and yellow) with a frequency equal or higher than 0.7, and we ran a between-subjects one-way ANOVA with the participant's L1

(between-subjects: 6 levels) as the independent variable, and the frequency with which each color was labeled by a basic color term as the dependent variable. We then selected the color name-color pairs that showed no significant effects of L1. Munsell values, labeling frequencies and ANOVA statistics were (N is the frequency with which a color is labeled using the relevant basic color word): 7.5R 5/18 for red (N=0.825, SD=0.278; $F(5,54)=0.611$, $p=0.692$), 5YR 6/12 for orange (N=0.776, SD=0.34; $F(5,54)=0.825$, $p=0.538$), 7.5Y 7/10 for yellow (N=0.797, SD=0.33; $F(5,54)=0.659$, $p=0.656$), 7.5YR 3/6 for brown (N=0.8, SD=0.341; $F(5,54)=0.755$, $p=0.586$), and 5GY 4/8 for green (N=0.878, SD=0.242; $F(5,54)=0.746$, $p=0.593$). Among these, 7.5YR 3/6 (brown) and 5GY 4/8 (green) were selected as the terminal colors. This choice was dictated by the middle color (2.5.2), which was intermediate between green and brown.

2.5.2. Selection of the middle color

We identified a set of middle color candidates by means of the following criterion: the color X is labeled using the basic terms A and B on average with frequency F; in the ideal case, there is no other label than A and B with which X is labeled, and F is 0.5. Any pair of terms A and B with approximately equal F would suggest X lies at the boundary between the lexical categories A and B. A mixed ANOVA model was used, with the independent factors L1 (between-subjects: 6 levels) and Color Terms (within-subjects: 2 levels), and as the dependent variable the frequency with which participants would label a particular color using either of the two color terms. As a potential middle color, we identified a Munsell color for which participants used the two names 'green' and 'brown' with approximately equal frequency (G is the mean frequency by which the color is labeled with 'green', and B with 'brown'): B=0.447, SD=0.34; G=0.38, SD=0.31 (7.5Y 4/6; L1: $F(5, 106)=0.871$, $p=0.503$; Name: $F(1, 106)=0.597$, $p=0.442$; L1xName: $F(5, 106)=1.526$, $p=0.188$).

2.5.3. Selection of transition colors

Transition colors were defined as those within our stimulus set that interpolated in color space the terminal and middle colors chosen based on statistical criteria: 2.5Y 4/6 as a brown (B=0.719; SD=0.34), and 2.5GY 4/6 as a green (G=0.848; SD=0.28). The final 5-color array was as follows (cells show mean labeling frequencies):

	7.5YR 3/6	2.5Y 4/6	7.5Y 4/6	2.5GY 4/6	5GY 4/8
'Brown'	0.8	0.719	0.447	-	-
'Green'	-	-	0.38	0.848	0.878

2.5.4. Artificial color terms and categorizations

We used the two syllables 'to' and 'nu' as artificial basic color terms to be used as signals in SGs. These two syllables are phonologically unrelated to the basic color terms for brown ('marrone') and green ('verde') in Italian, the native language of participants in the main experiments. There were four possible categorizations of the 5-color array, arising out of the combination of the 2 color terms and 5 colors, allowing only sets of either 2 or 3 colors. In categorization 1, 'to' was associated to colors 1 and 2, and 'nu' to colors 3, 4 and 5. In categorization 2, 'to' was associated to 1, 2 and 3, and 'nu' to 4 and 5. In categorization 3, 'nu' was associated to 1 and 2, and 'to' to 3, 4 and 5. In categorization 4, 'nu' was associated to 1, 2 and 3, and 'to' with 4 and 5. An equal number of participants (N=5) in the main experiments was randomly assigned to each categorization on Day 1.

2.6. Colorblindness test

Participants' color vision was evaluated by a computerized version of the Ishihara (1917) colorblindness test to exclude a possible color vision impairment. This was carried out on Days 1 and 3 before the forced choice labeling tasks in the main experiment, as well as on Day 1 preceding the forced choice labeling task in the two color discrimination studies (2.10). All participants included in the experiments passed the Ishihara test.

2.7. Testing native color categories: forced choice labeling task

On Day 1, after the colorblindness test, participants performed the same forced choice labeling task used in the 5-color array construction study (2.1). The purpose of this forced choice labeling task on Day 1 was to assess how participants categorize the browns and greens in the 5-color array, in order to determine whether and how learning in the SG would change native color categories. The stimuli and procedure where the same as in the preliminary forced choice labeling study (2.1.1-2.1.2). On Day 3, we administered once again the forced choice labeling task to determine whether learning of artificial color categories via the SG led to effective acquisition, i.e., whether learning modified native color categories. For each participant, we calculated the number of repetitions (maximum 10; 2.1.3) in which each of the five colors in the array was labeled with 'green' or 'brown'. We focused in particular on the middle color (7.5Y 4/6). Participants were classified as having either of two native color categorizations of the 5-color array, depending on whether the middle color was grouped with greens (N=17) or browns (N=3).

2.8. Learning artificial color categories: the Signaling Game

Each participant learned one of the 4 possible categorizations (2.5.4) of the 5-color array using the 2 artificial color terms, as follows. Participants played a Signaling Game (SG), identical to the SG in the computer simulations previously described (2.2). At the beginning of the game, participants were randomly assigned to one of 4 languages: the color categorizations and vocabularies of 2.5.4. During the game, in each trial, a color invisible to the player was randomly drawn from the 5-color set, according to the color frequency pattern: 0.15 for each of the terminal colors, 0.20 for each of the transition colors, and 0.30 for the middle color (see 2.2.3). The color term corresponding to the randomly-drawn color was shown to the player, whose task was to choose from the 5-color array one of the 2 or 3 colors belonging to the category denoted by the observed term. Feedback was given in each trial as to whether the random color and the color chosen by the player matched

or not. In this version of the SG, the computer plays as the sender, and the participant is the receiver, whose task is to decode the signals he receives from the computer, and infer the correct color categorization and color vocabulary.

2.8.1. Procedure

At the start of a game, the participant/player was shown both terms 'to' and 'nu' at the center of the screen for 300 ms, followed by the 5-color array with colors in random order. In each trial, the participant/player observed a color term (shown in capital letters with font Calibri size 48 in the middle of a mid-grey screen), and had to choose one of the colors associated to that term in the language by pressing on a full-size keyboard one of five keys, spatially associated to the 5-color array, which was shown on a single screen with colors in random order. The participant saw the array with colors, now ordered by hue (brown to green), as feedback, with two arrows: one pointing to the randomly drawn color, associated to the observed color term in the language (the arrow was marked with the label 'AS', 'associated'), the other pointing to the color selected by the player as a response to the observed term (marked with the label 'SC', 'selected').

Training with SGs consisted of two sessions, one on Day 1 and one on Day 2. Each session lasted until the participant had reached to 60 correct responses. At the end of the game, the participant had to select the categorization he thought he had learned. He was presented with all possible categorizations and vocabularies on a single screen, shown as four arrays with the terms 'to' and 'nu' indicating the different color-term combinations, and he had to press a key from 1 to 4 to choose one. They were asked to confirm their choice in a second trial identical to the first.

2.9. Neural consequences of learning color categories: EEG study

2.9.1. Stimulus presentation

The last task administered on Day 2 was passive visual exposure to color-term and term-color sequences, which were either Learned (agreeing with the categories and vocabulary learned in the SG) or Incongruent (deviating from the learned set). The stimuli were the 5 colors and the 2 color terms used in the training with SGs. There were two blocks, whose order in each session was randomized across participants: in one block, a color term was shown for 300 ms (capital letters, in Calibri font size 48, in the middle of a mid-grey screen), followed by a color for 400 ms (shown as a square of 120×120 pixels) in the middle of a mid-grey screen and by a fixation cross for 1250 ms (term-color or TC sequences); in the other block, a color was shown for 400 ms, followed by a color term for 300 ms and by a fixation cross (color-term, CT sequences). Each block had 200 Learned and 200 Incongruent trials. Across trials, all possible combinations of colors and color terms, in both CT and TC sequences, were used. Moreover, all colors were presented with equal frequency, in contrast with the non-uniform frequency distribution used in the SG training (2.2.3). This was effected in order to avoid that color-position effects (e.g., to terminal versus intermediate colors; see Introduction) were confounded with frequency effects.

2.9.2. Data acquisition

The EEG was sampled from 128 scalp locations using a BioSemi system with active electrodes. Instead of a ground channel, BioSemi employs two electrodes, a Driven Right Leg (DRL) and a Common Mode Sense (CMS) channels, driving the average potential close to the amplifier AD-box reference voltage. This analogue-to-digital reference can be considered as the virtual ground of the amplifier. DRL and CMS were placed at symmetric side positions relative to the mid-point between A1/Cz and A19/Pz. An average reference was used during the recordings. The sampling rate was 1024 Hz. The data

were high-pass filtered at 0.1 Hz and low-pass filtered at 256 Hz. All filtering was digital.

2.9.3. Data analysis

EEG data were epoched from -200 ms to 1000 ms relative to the onset of the second stimulus (i.e., the color term in CT sequences, and the color in TC sequences), and were baseline corrected using data from the −200 to 0 ms interval. Segments were discarded if they contained activity exceeding ±100 μV thresholds in any channel. Eye movements in the bilateral fronto-polar channels C29 and C16 were identified by means of the following procedure: (1) EEG signals from electrodes C29 and C16 were band-pass filtered at 1-15 Hz; (2) Hilbert analytic amplitudes of these signals were derived, resulting in the envelope of each signal; (3) each envelope was then normalized by computing z-scores for each signal; (4) a z-value per time point was computed, summing z-scores from the C29 and C16 channels, and normalizing the sum by dividing it by the root of 2 (i.e., the number of channels involved); (5) EEG segments exceeding the threshold resulting from step (4) were discarded. A similar procedure was employed to identify and discard muscle artifacts in any of the EEG channels, with the only difference in step (1): data were digitally band-pass filtered at 100-125 Hz. A digital low-pass filter at 30 Hz was applied to segments surviving artifact rejection. ERPs were computed by averaging over artefact free epochs from each condition (Learned/Incongruent), in each block (CT/TC), for each participant separately. The definition of Learned and Incongruent trials for a given participant depended on the particular color categorization that he/she was trained on during the SG sessions. Finally, grand-average ERPs were computed by further averaging over participant specific averages.

Statistical analyses of ERP effects were conducted by means of the following procedure (Maris and Oostenveld 2007): (1) participant specific ERP averages were compared between the Learned and Incongruent conditions from each channel and time point with dependent samples *t*-tests; (2) data from neighboring time points and channels in which p-values were smaller

than 0.05 were clustered together; (3) the cluster-level *t*-statistics was computed as the sum of *t*-values from all samples belonging to the cluster; (4) the cluster-level p-value was estimated using a Monte Carlo simulation: participant specific ERP averages across all samples in a cluster from both experimental conditions were collected in a single set; this new set was randomly partitioned into two subsets of equal size; the subsets were compared by means of a *t*-test; these steps were repeated 1000 times; a cluster-level p-value was computed as the proportion of partitions that resulted in a larger *T*-statistic than in the observed ERP data.

2.10. Probing color categories: color discrimination studies

An independent sample of participants performed the same colorblindness test, forced choice labeling task and the two training sessions with SGs as in the main study (2.7-2.9). These were followed by a same/different color discrimination task modified from Liu et al. (2010), and by an odd-one-out discrimination task modified from Witzel & Gegenfurtner (2013).

2.10.1. Participants

Twelve right handed native speakers of Italian (6 female, mean age 23) took part in the studies during two consecutive days. On Day 1, the Ishihara colorblindness test (2.7) was followed by the forced choice labeling task (2.8) and by the first training with SGs (2.9). On Day 2, the second training with SGs was followed by the discrimination tasks.

2.10.2. Same/different discrimination task

Participants were presented with a colored square (155×155 pixels) surrounded by a colored frame (190×190 pixels) in the center of a mid-grey screen, and they were asked to judge, by pressing either of two keys on a full-size keyboard, whether the square and the frame were of the same or of a

different color. If the square and the frame were of the same color, this was considered a 'same' trial. If the square was of a different color than its surrounding frame, it was considered a 'different' trial.

In each trial, first a fixation-cross appeared for 1000 ms at the center of the screen, followed by a square surrounded by a frame shown for 200 ms, and by a blank mid-grey screen shown for 800 ms. The stimuli consisted of 25 color-frame combinations (5×5 colors) re-used in 300 trials. Of these 300 trials, 150 showed the inner square and the surrounding frame in the same color, and 150 showed them in different colors. In 50% of the different color trials, the colors of the square and of the frame were associated to the same term in the artificial color system that each participant had learned during the SG training ('within category' or WC trials). In the other 50% of the different color trials, stimuli were constituted by squares with their surrounding frames in a different color, which participants had learned to associate to different color terms ('across category' or AC trials).

2.10.3. Odd-one-out discrimination task

We used the spatial 4-Alternative Forced-Choice (4AFC) discrimination task from Krauskopf and Gegenfurtner (1992), also used by Witzel and Gegenfurtner (2013). In each trial, observers were shown 4 colored disks. One of these colors differed in chroma (i.e., in hue, but not lightness) from the other 3. Participants had to indicate which disk was different by pressing one of 4 keys corresponding to the 4 positions in space of the disks.

At the start of a trial, a fixation cross appeared for 1000 ms at the center of the screen, followed by the 4-disk stimulus shown for 500 ms against a mid-grey background, and by a blank mid-grey screen shown until a response was given. The odd disk occupied any of the 4 possible positions on the screen, randomized across trials. In 50% of the trials, the color of the odd disk was associated with the same color term as the color of the 3 other disks in the color categorization that the participant had learned during the SG training sessions ('within category' trials). In the other 50%, the color of the

odd disk was associated with a different color term than the color of the other 3 disks during the SG training ('across category' trials).

2.10.4 Analysis of color discrimination data

Paired *t*-tests were conducted to assess the differences in the performance of the participants in the two tasks across conditions. Differences in reaction times and accuracy were compared between across-category and within-category trials. Accuracy was determined by the number of correct trials in the last fourth part of each of the trainings (where coordination is expected to be present already) divided by a fourth of the total amount of trials the participant needed to end that training session. To analyze the contrast between within and across category trials, for each of the discrimination tasks (Same-Different and Odd One Out), we ran two-way ANOVAs on within- and across-category trials. For each task, 2 ANOVAs were run with the reaction times of subjects to the each of the two tasks and 2 ANOVAs with the accuracies obtained by subjects in each task. This was done first with all 5 colors in the array as factors and then with each position as factors: terminal colors (colors 1 and 5 together), transition colors (colors 2 and 4 together) and middle color (color 3). Moreover, Wilcoxon's rank sum tests were conducted comparing the means of the participants' performance in reaction times and accuracies, for all 5 colors separately and for the 3 positions.

3. Results

3.1. Discrimination studies

3.1.1. Learning color categories and vocabularies

Seven out of twelve participants learned the artificial color system during the first training session with signaling games, and eleven out of twelve learned it during the second training. The average number of trials it took participants to get to 60 correct responses and finish a session was higher for the second training than for the first: 214 trials (SD=69.06) and 245 (SD=134.44), respectively. Nonetheless, this difference was not significant (Paired *t*-test; $t(11)=1.2881$, $p=0.224$).

3.1.2. Same/different discrimination study

Reaction times in within-category (WC) (M=456.28) trials were significantly longer than in across-category (AC) trials (RT M=432.47) (Paired samples *t*-test, $t(11)=3.1582$ $p=0.0091$). There were no significant differences either between same-color trials (RT M=431.96) and WC trials (Paired *t*-test, $t(11)=1.5833$, $p=0.141$), or between same-color trials and AC trials (Paired *t*-test $t(11)=0.0391$, $p=0.969$), or between different color (i.e., WC+AC trials, RT M=444.37) and same-color trials (Paired *t*-test $t(11)= 0.9071$, $p=0.383$).

Accuracies did not differ across conditions (AC/WC trials: paired *t*-test, $t(11)=1.1747$, $p=0.265$; AC/same-color trials: paired *t*-test, $t(11)=0.1545$, $p=0.88$; WC/same-color: paired *t*-test, $t(11)= 1.3283$, $p=0.211$; different/same trials: paired *t*-test, $t(11)=1.1243$, $p=0.285$). The ANOVA showed no interactions between the 5 colors and condition (WC/AC) for either reaction times ($F(1,112)=0.007$, $p=0.934$) or accuracies ($F(1,112)=0.107$, $p=0.744$). When comparing WC to AC for each color, Wilcoxon rank sum tests showed significant differences in reaction times (Table 4) for all colors (color 1: $V=78$, $p=0.0004$; 2: $V=74$, $p=0.0034$; 3: $V=5$, $p=0.0049$; 4: $V=78$, $p=0.0005$; 5: $V=78$, $p=0.0005$), while in accuracies (Table 4) for color 4 only (1: $V=3$, $p=0.1411$; 2: $V=9$, $p=0.8335$; 3: $V=42$, $p=0.152$; 4: $V=0$, $p=0.009$; 5: $V=1$, $p=0.0523$). When

using the position of the color in the array (i.e., terminal, transition or middle color) as a factor, the ANOVA showed no significant interactions effects of reaction times ($F(2,60)=0.006$, $p=0.994$) or accuracies ($F(2,60)=0.012$, $p=0.988$). In all tests, the Bonferroni-corrected α is 0.01.

3.1.3. Odd-one-out study

Reaction times in WC (RT M=527.63) trials were significantly longer than in AC (RT M=503.34) trials (Paired *t*-test, $t(11)=3.6503$, $p=0.0038$). Accuracy was better in WC (mean=0.97) trials than in AC (mean=0.92) trials (Paired *t*-test, $t(11)=3.2387$, $p=0.0079$). The ANOVA showed no interaction between the 5 colors and condition for reaction times ($F(1,112)= 0.144$, $p=0.706$) or accuracies ($F(1,112)=0.002$, $p=0.963$). When comparing WC to AC trials for each color, Wilcoxon rank sum tests showed significant differences in reaction times for colors 1 and 4 (Table 4) (1: $V=78$, $p=0.0005$; 2: $V=68$, $p=0.021$; 3: $V=13$, $p=0.0425$; 4: $V=78$, $p=0.0005$; 5: $V=68$, $p=0.021$), as well as in accuracies (Table 4) for colors 2 and 4 only (1: $V=5$, $p=0.5896$; 2: $V=45$, $p=0.009$; 3: $V=13$, $p=0.1522$; 4: $V=45$, $p=0.0091$; 5: $V=5$, $p=0.2932$). Using the position of the color in the array (terminal, transition or middle) as a factor, the ANOVA showed no interactions of condition and position for reaction times ($F(2,60)=0.175$, $p=0.840$) and accuracies ($F(2,60)=0.290$, $p=0.749$). In all tests, the Bonferroni-corrected α is 0.01.

Same - Different	Reaction Times		Accuracies	
Color/Condition	WC	AC	WC	AC
1	446.7 (51.2)	422.5 (48.8)	0.94 (0.1)	0.94 (0.09)
2	462.9 (53.4)	440.4 (43.7)	0.9 (0.13)	0.93 (0.09)
3	436.7 (67.59)	457.7 (47.4)	0.89 (0.17)	0.87 (0.12)
4	451 (53.8)	425.2 (56)	0.85 (0.16)	0.92 (0.11)
5	459.1 (63.3)	436.3 (53.1)	0.91 (0.09)	0.95 (0.1)

Odd One Out	Reaction Times		Accuracies	
Color/Condition	WC	AC	WC	AC
1	499.5 (83.6)	471 (61.1)	0.98 (0.02)	0.98 (0.03)
2	552.7 (124.7)	530.6 (140.3)	0.91 (0.1)	0.99 (0)
3	564.9 (163.5)	542.8 (84.5)	0.94 (0.04)	0.95 (0.04)
4	548.3 (116.7)	500 (87)	0.93 (0.08)	0.96 (0.05)
5	505.5 (82.1)	476.9 (88.1)	0.97 (0.03)	0.98 (0.03)

Table 4. Results of color discrimination experiments. Mean values (and SD) are reported. Abbreviations: WC (within-category condition); AC (across-category condition).

3.2. Learning color categories

3.2.1. Forced choice labeling task on Days 1 and 3

On day 1, out of the 20 participants, 17 grouped the middle color in the array together with the greens (average probability across participants in this group of the color being called 'green', p=0.8) and 3 grouped it with the browns (average probability across participants in this group of the color being called 'brown', p=0.7) in at least 60% of the trials in which the middle color was shown. On day 3, only one person changed from grouping the middle color in the array together with the greens (probability=0.7) to grouping it with the browns (probability=1) after having played the signaling games with a

categorization that did not match the one they had originally used in the first forced choice labeling task.

3.2.2. Learning in the signaling games on Days 1 and 2

All participants but one (i.e., 19) learned the artificial color system during the first training session with signaling games. All participants (20) learned after the second day. The amount of trials required to finish the task was significantly higher in the first day (mean number of trials=191.05) than in the second (160.85; paired samples *t*-test $t(19)=4.5893$, $p=0.0002$; Fig. 2b). Accuracy during training on the second day (mean=37.780) was significantly better than during the first (mean=32.241; paired *t*-test $t(19)=4.0671$, $p=0.0007$). Accuracies during both days differed significantly from chance level and from optimal accuracies by informed choices (one-sample *t*-tests, $p<0.001$; Fig. 2c). Accuracies during both days were not different from optimal accuracies by random choices, and from the accuracies obtained in the computer simulations (one-sample *t*-tests, $p>0.1$; Fig. 2c).

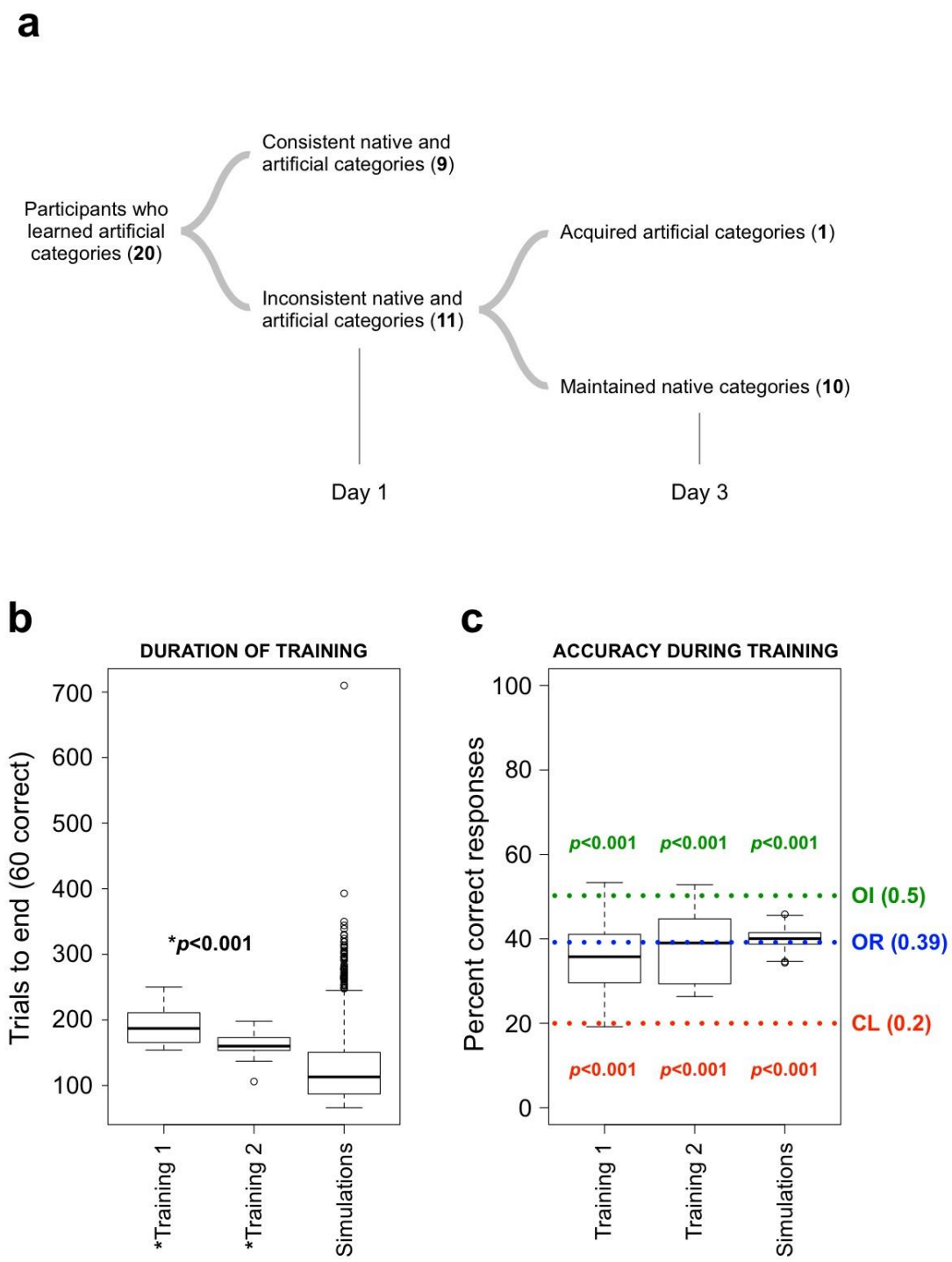

a
Participants who learned artificial categories (20)
Consistent native and artificial categories (9)
Inconsistent native and artificial categories (11)
Acquired artificial categories (1)
Maintained native categories (10)
Day 1
Day 3
b
DURATION OF TRAINING
Trials to end (60 correct)
*p<0.001
*Training 1
*Training 2
Simulations
c
ACCURACY DURING TRAINING
Percent correct responses
p<0.001
OI (0.5)
OR (0.39)
CL (0.2)
Training 1
Training 2
Simulations

Figure 2. Results of the main behavioral experiments and of computer simulations. (a) Tree diagram illustrating the learning trajectories of participants (N=20). During the training on Day 1, 9 participants learned artificial color categories that were consistent with their own native color categories, and 11 learned artificial categories that were inconsistent with their own native categories. Of these 11 participants, 1 acquired the artificial color categories of the training set, and the remaining 10 participants maintained their native color categories. (b) Duration of training for participants on Day 1 (Training 1) and Day 2 (Training 2), and for agents in computer simulations. The ordinate shows the total number of trials to end a session at 60 correct trials: i.e., 60 exact matches between the color randomly drawn from the 5-color array, and associated to the color term seen by the player, and the color chosen by the player in response to the color term. (c) Accuracies (percent correct trials relative to the total number of trials to end a session) in Training 1 and Training 2, and in computer simulations, are significantly above chance level (CL, 0.2, red line) but significantly below the optimal accuracy level with informed choices (OI, 0.5, blue line). Actual accuracies do not differ from the optimal accuracy level with random choices (OR, 0.39, green line). In all boxplots, the thick line inside the box is the median, box height is equal to the interquartile range, whiskers indicate adjacent values, and empty circles are outliers.

3.3. EEG

In ERPs we found no incongruency effects in either CT or TC trials that involved the terminal colors, i.e., the best exemplars of each category. However, the middle and transition colors did show clear ERP effects. We found a dissociation between CT and TC sequences, with the strongest effects for CT at the middle color, and the strongest effect for TC at transition colors (Fig. 3). Both effects are negative-going ERP waves that appear around 550-580 ms from stimulus onset and last until the end of the epoch (Table 3). These effects are accompanied by earlier and smaller clusters which fail to reach statistical significance (Table 3). We tested whether the average amplitude of ERP effects across all channels for each given color depends on the frequency with which that color was presented in the training sessions, and we found no correlation (CT: r=-.53, p=0.3615; TC: r=-0.79, p=0.1114)

Term-Color (TC)	Early clusters				Late clusters			
	Latency	S	Tsum	P-value	Latency	S	Tsum	P-value
Terminal colors	-	-	-	-	-	-	-	-
Transition colors	134-296 ms	1276	-3222	0.129	550-1000 ms	13045	-34791	<0.001
Middle color	-	-	-	-	-	-	-	-

Color Term (CT)	Early clusters				Late clusters			
	Latency	S	Tsum	P-value	Latency	S	Tsum	P-value
Terminal colors	-	-	-	-	-	-	-	-
Transition colors	-	-	-	-	-	-	-	-
Middle color	237-337 ms	665	1810	0.177	580-1000 ms	1963	-5010	0.039

Table 3. Results of cluster-based permutation statistics of ERP data comparing between Incongruent and Learned association trials in Term-Color and Color-Term blocks. The table shows the latency of clusters, cluster size (S) in number of adjacent time-channel samples, the sum of *T*-statistics (Tsum) in each cluster, and Monte Carlo p-values. Marked empty cells (-) indicate absence of clusters.

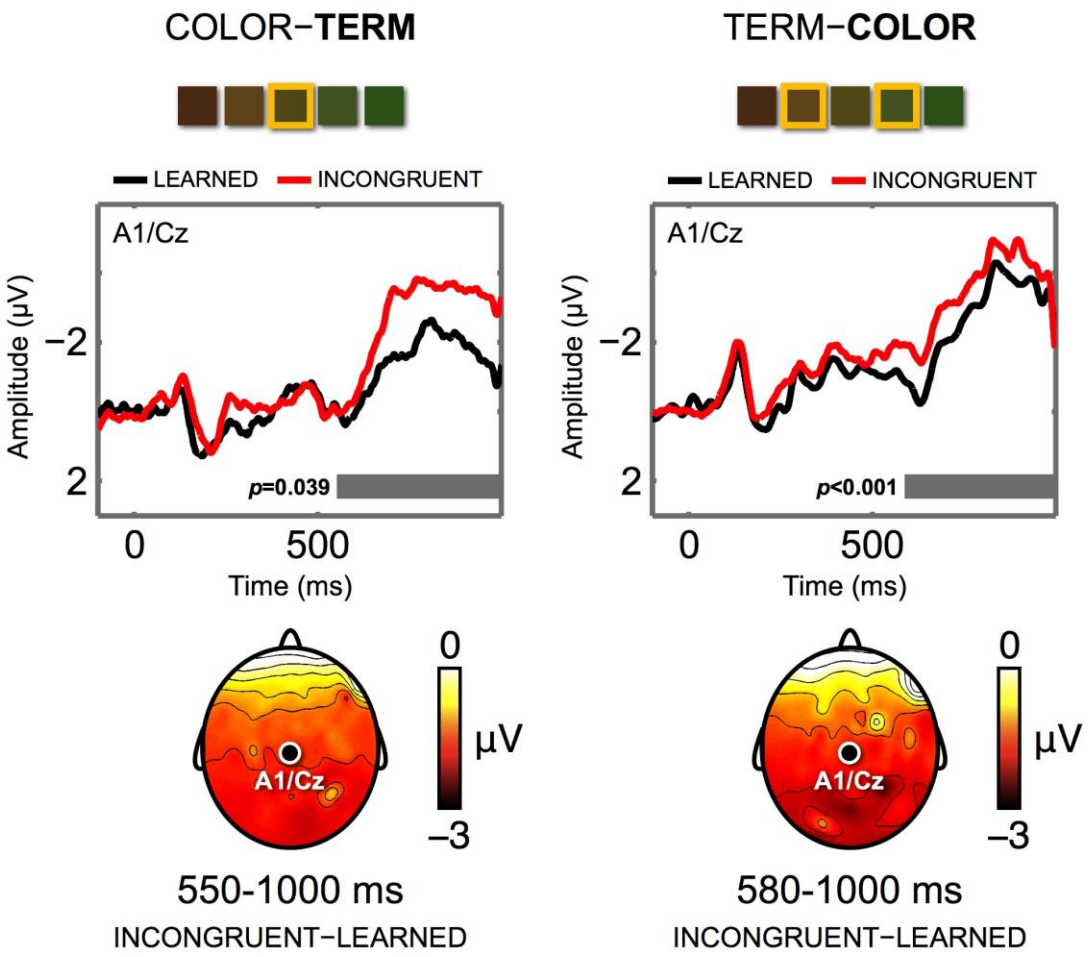


Figure 3. Event related brain potentials (ERPs) evoked by the color term presented after the middle color in CT sequences (left), and by transition colors presented after the color term in TC sequences (right). The onset of the eliciting stimulus (i.e., the color term in CT sequences, or the color in TC sequences) is at 0 ms. Negative amplitude values are plotted upwards. Single-channel ERP plots show grand-average (N=20) waveforms from Learned (black) and Incongruent (red) trial s from the vertex channel A1/Cz. Dark grey bars in the lower portion of each plot represent the highest ranked cluster of adjacent time points at which statistically significant effects were found (Table 3). Topographic maps display grand-average (N=20) differences between Incogruent and Learned trials in the time window of the highest ranked cluster.

4. Discussion

The aim of this study was to investigate the cortical plastic changes produced by the acquisition of new color categories and color terms. Our results suggest there may be a common neural substrate for color categories, shared by representations of colors and color terms. Moreover, regardless of whether a color or a color term is processed, the underlying representation seems to be of a high-level/semantic kind (late ERP clusters) and only to a lesser extent of a low-level/perceptual kind (early ERP clusters). Thus, our results stand in contrast with Fonteneau & Davidoff (2007), Holmes et al. (2008), and Kwok et al. (2011), whose findings pointed instead to a primarily perceptual basis of color categories. A possibility here is that the way in which a novel color system is learned, whether in a linguistic or communicative setting as in the present experiments, or by individual associative learning as in the previous studies, may determine what cognitive systems are recruited to represent and process stimuli belonging to each color category. This hypothesis is not itself a variant of the Sapir-Whorf hypothesis (i.e., that non-linguistic cognitive differences co-vary with structural differences between language systems, and that the latter tends to determine the former; Kay & Kempton 1984), but rather highlights one of its preconditions: that cognitive effects at category boundaries can be produced by learning (artificial) languages.

Importantly, the effect we found in incongruent trials, is not confounded with either consistency or inconsistency between these newly learned categories and native color categories: half of the participants in our sample learned categories that were consistent with their native color system, and half of them learned an artificial color system that was inconsistent with their native categorization. However, the incongruency effect was significant at the sample level.

We calculated and obtained optimal frequencies of presentation of the colors by means of a series of computer simulations, and this enabled us to minimize the possibility of participants learning the probabilities of presentation of each color, another potential confound to studying the

mappings between colors and terms. Moreover, we found no correlation between the mean amplitude of the ERP effect across all channels for a given color and the frequency of appearance of the color in the signaling game. Finally, we randomized the order of the colors in the array to insure that participants were not just learning positions in space or an abstract order of cells in the array.

4.1. An N400 effect to color category incongruencies?

The polarity and topographical distribution of the ERP effects found in this study are similar to those of N400 effects (Kutas & Hillyard, 1980). However, the latency in our study is delayed. This phenomenon is in line with previous work. Aparicio (2012) presented trilinguals with unrelated non-cognate words in the 3 languages they spoke, using a semantic categorization task: words in L1 elicited earlier N400 peak amplitudes than L2 and L3 words. Any natural or artificial semantic system that is added to one's L1 may therefore elicit N400 effects whose onset and peak latencies are delayed. Moreover, N400 effects are elicited by all content bearing stimuli, including words and pictures (Nigam et al., 1992). Hence, the similarity between the negative ERPs evoked by colors and color terms in this study may provide another reason to consider them as N400 effects.

4.2. Category ambiguity and ERP effects

Possibly the most striking result to emerge from our study is that incongruency effects in ERP are strongest not at terminal colors, i.e., for the best exemplars of each color category, but at the intermediate colors, i.e., at the boundary between categories. This appears to be inconsistent with a neural model in which the best exemplars (or prototypes) of each category are approximately equidistant from the boundaries, and occupy approximately the 'center' of a category. When presented with a color term, an observer would activate a representation of the prototype as a prediction for the upcoming color stimulus. So, the farther the color stimulus from the prediction, the stronger the ERP effect (possibly reflecting prediction *error*), as in predictive coding models of neural processing (Friston 2010). ERP effects

should therefore be largest at the opposite terminal, e.g., at the terminal brown, if the color term for green is presented. Overall, this should result in a U-shaped distribution of ERP effects over the 5-color array. The fact that a different pattern was observed suggests a different neural explanation must be sought. A hypothesis is that ERPs do not reflect prediction errors, or possibly distance of deviants from the center of a category, but *learning effort*. Terminal colors can be effortlessly assigned to either color category, first because they do not pose any special discrimination problems relative to the other colors in the array, and second because they are unambiguous members of either category also in the L1 of participants (see our forced choice labeling data). In contrast, transition and middle colors may belong to one or two color categories in the native language of participants, and the artificial color system they are trained on may not follow that pattern. Furthermore, middle colors, and to a lesser extent transition colors, are likely to be lexically ambiguous (see our forced choice labeling data), and therefore *constitute* a boundary, rather than occupying a category-internal position just past a boundary. Consequently, imposing a category boundary *between* colors may be a relatively demanding rewiring process for neural circuits, which would respond with stronger signals to stimuli that are incongruent with the outcome of such plastic changes at the boundary. However, this account does not explain the fact that ERP effects were stronger at the middle color in CT sequences, whereas they were stronger at transition colors in TC sequences, and is in fact consistent also with a reversed pattern of effects, or patterns in which only middle or only transition colors are the loci of the strongest ERP effects in both sequence types. Boundary effects were only partly seen in the present color discrimination studies. Statistical analyses show that only for the transition color 4 in both discrimination tasks, and for colors 2 and 4 in the Odd One Out task, was accuracy significantly better in across-category trials than in within-category trials. Although our data provide preliminary evidence in that direction, further research is necessary to establish how exactly ERP and discrimination effects depend on the presence of a boundary between categories.

4.3 Languages selection for the construction of the array

Although in order to build our stimuli we tested speakers of 6 different languages, it is important to point out that apart from Chinese, they were all languages of the same family (Indo-European). This makes the language sample not representative of different language families. Since languages from a same family will tend to categorize the colors in a similar way, we cannot be sure of whether using a truly representative sample of languages would have resulted in a different effect.

5. Conclusion

Consider again the fictional scenario described in the Introduction to this paper. Signaling theory predicts that the botanist and the natives will coordinate on the meaning of the signal 'duwi'. In this case coordination is unilateral as the botanist will learn the range of shades of green (or a subset thereof) the natives call 'duwi'. Similar situations, in which translation is required, but must start from an initial state of seemingly irreducible indeterminacy, are common in human interactions. Second language acquisition is a case in point, but language contact also involves the construction of a (possibly temporary) parallel semantic system to allow some degree of communication among individuals belonging to different populations.

Signaling games appear to be a suitable laboratory model of such situations. We used signaling games to train participants to an artificial color system featuring 5 colors and 2 color terms, and we tried to pinpoint the neural consequences of the learning process using EEG. Our results indeed suggest that ERP responses track learning effort and are strongest at the boundary between categories.

Chapter 3

Rewiring color categories in the lightness dimension: An EEG study

1. Introduction

In Chapter 1, we described a study in which people from two populations that are part of the same geographical region but speak different languages, Galician and Spanish, are asked to name colors varying along the lightness and hue dimensions. The results showed differences in how speakers of these two languages categorize color space, in particular how they set the boundaries between the yellows and greens in the hue dimension, and within the color blue in the lightness dimension. This led us to the conclusion that different regional color meanings can co-exist even if two populations share geographical space and interact frequently.

In our previous experiment (see Chapter 2) we re-created, through the use of signaling games, the situation in which two languages interact and, starting from a point of an apparently irreducible indeterminacy, a receiver begins to find meaning and make use of new terms presented by a sender, i.e., a speaker of a different language, with a potentially different semantics for color terms. We used a color space limited by a prototypical (see Chapter 2, 2.5) brown color and by a prototypical green color. The intermediate colors varied in the hue dimension, from brown to green. In this study we want to validate and extend the results of the previous one, in another dimension of color space (lightness). To this end, we constructed a color array with Munsell colors that are all considered 'blue' by English speakers.

Several research groups have studied differences across languages in how people categorize and name colors that corresponds to the term 'blue' in English. These are briefly reviewed below. Languages like English or German use only one basic color term, to refer to the blue color space. The basic color terms are monolexemic terms that are used with relatively high-frequency in the language and are consistently agreed upon by speakers of a certain language in forced choice labeling tasks, Terms like "salmon pink",

"turquoise", "lime green", etc. are not used with much consensus among English speakers or very consistently, so they are considered non-basic colors. Also, applications must not be restricted to a narrow class of objects (e.g. blond: human hair) and the colors have to be psychologically salient for the informants. However, languages like Korean, Japanese, Greek, and Russian, have two basic blues. These languages were studied in Athanasopoulos et al., (2010) using a color oddball detection task , in Thierry et al., (2009) with a shape oddball detection task Roberson et al. (2009) with a same-different judgment task, and Winawer et al. (2007) with a color matching task. For these languages these studies found higher accuracies and/or faster reaction times when the stimuli being presented belonged to different linguistic categories than if the colors were from the same category, compared to languages which have only one term for the blues.

Gonzales Perilli et al (in prep) studied the case of the blues in Spanish spoken in Spain and Spanish spoken in Uruguay ('Rioplatense Spanish'). In Spain 'azul celeste' (light blue) is a subcategory of the color blue whereas in Uruguay, as well as in Argentina, Chile and Paraguay, 'celeste' is a basic color term (i.e. azul and celeste are independent color categories). When tested in discrimination tasks containing items from the categories azul and celeste, participants from Uruguay performed better in across-category trials (Winawer et al., 2007). This effect that was not found in the Spanish participants, who performed roughly equally in both within and across category trials.

One possible explanation for the effect in Rioplatense Spanish may be language contact. The color 'celeste' as a basic color term could have been borrowed by Rioplatense Spanish from Italian due to the high rate of Italian immigration in the southern South American countries during the World Wars. Paramei & Menegaz (2013), when comparing Italian to English found that Italian speakers, monolingual and bilingual, make use of three color terms to name the blue area: celeste, azzurro and blu. The Italian prototypical 'blu' was found to be darker than the English prototypical 'blue'. When they tested Italian-English bilinguals the location of bilinguals' 'blu' deviated from that of Italian monolinguals: it shifted towards the English 'blue' (Athanasopoulos,

2009) which was related to the bilinguals' proficiency in English and duration of immersion in the English speaking country.

With this experiment we try to model a situation in which a speaker of a language that has a single term for the color blue (e.g. English) learns a language, like the ones mentioned before, which partitions the blue color space into two different categories using two terms, different from what occurs in their native language, as described below.

1.2. Constructing the color space

In our previous experiment (see Chapter 2) we found that incongruency effects in ERPs when participants saw color terms preceded or followed by colors were stronger at the middle and transition colors, i.e. at the boundary between categories, and not at terminal colors, i.e. at the best exemplars of each color category. One of our hypotheses was that middle colors are likely to be lexically ambiguous (see Chapter 2, 2.1), and therefore constitute themselves a boundary, rather than occupying a category-internal position just past a boundary. Imposing a category boundary between colors may be a relatively demanding rewiring process for neural circuits, and this effort might explain the lager ERP effect seen in the middle and transition colors. To further test this hypothesis, we wanted to try to produce a similar ERP boundary effect at the terminal colors of the array, using an ambiguous color as an extreme (terminal color) of the arrays instead, and more prototypical colors in the rest of the array. Color 5 lays within the blue hue (the boundaries of this category were determined based on the results of a forced choice labeling study with English speaking participants; details below) but the It is ambiguous, i.e., close to the boundary with other categories (in this case, color 5 is close to black). In the new task, participants are forced to partition a uniform portion of color space in their native language, that is, the blue space, into subcategories, using two artificial terms (pseudo-words).

2. Methods

We will first explain how we constructed and used a 5-color array as a stimulus set in the main behavioral and EEG experiments using the results from a forced choice labeling study we conducted (2.1). Afterwards, the structure of the main experiment on two successive days (2.2; Table 1), the participants (2.3), and the materials and methods (3.4) involved in the main experiments are described.

2.1. Constructing color stimuli: forced choice labeling study

We conducted a forced choice labeling study to identify five colors that would form an ordered array with a light blue and a dark blue as extremes, and transition colors in between. The resulting array would sample a homogenous portion of color space varying along the lightness dimension in the natural category 'blue' for English speakers. The array was intended to serve as a stimulus set in the behavioral and second EEG experiment. Again in this experiment, whether the middle color in the 5-color array belongs to one or to the other artificial category is entirely determined by our training protocol with signaling games (SGs). Therefore, all 4 artificial categorization possibilities, with each category comprising either 2 or 3 colors, involving the 5 colors in the array and 2 artificial color terms, could be included in the experimental design. The aim of this forced choice labeling study was to identify 5 colors in the blue hue labeled by participants as 'blue' with a frequency close to 1, and 'light blue' or 'dark blue' with a frequency greater than 0.7 each. Then an array is built from the lighter dark blues or the darker dark blues with 5 colors differing in lightness, arranged in linear lightness order (from lighter to darker). If an array of 5 'light blue' colors was used, the first color in the array should be labeled with the basic terms 'blue' and 'white' with equal frequency (ideally ~0.5), and the fifth color should be labeled 'blue' with a frequency greater than 0.7 *and* with the non-basic terms 'light blue' and 'dark blue' with equal frequency too. If instead a 'dark blue' array was used, the first color in the array should be labeled 'blue' with a frequency greater than 0.7 *and* with the non-basic terms 'light blue' and 'dark blue' with equal

frequency, and the fifth color should be labeled ‘blue’ and ‘black’ also with equal frequency. Two more transition colors were found by interpolating the terminal colors with the middle color in color space. The results of the forced choice labeling study are reported in section 2.4.

2.1.1. Participants

Fourteen native speakers of English performed the task (mean age 28, 46; 9 female). Participants had normal or corrected-to-normal vision. Their color vision was evaluated by the Ishihara Colorblindness Test (2.5.).

2.1.2. Materials

The stimuli consisted of 42 Munsell (1912) colors varying in the lightness dimension in the blue hue range. Colors were selected using a standard Munsell color chart. Translation to the RGB systems was done using a standard conversion table.

In all experiments we used the same hardware and software settings. The stimulus presentation monitor was a 22-inch LCD Samsung Syncmaster 2233RZ with 120 Hz refresh rate (frame duration 8.33 ms). The screen resolution was 1680×1050 pixels, and each pixel subtended ~1.7 arcmin. The minimum and maximum luminances of the screen were 0.19 and 134 cd/m2. Luminances were measured using a Minolta LS-100 photometer. A gamma-corrected lookup table (LUT) was used to ensure that luminance was a linear function of the digital representation of the image.

2.1.3. Procedure

Participants sat in front of the computer monitor and a standard full-size keyboard (viewing distance ~80 cm). Their heads were fixed by means of a chin rest. In each block, participants saw each of the 42 colors 10 times in random order. Each color was presented for 400 ms, followed by a mid-grey screen (RGB [128 128 128]) for 1750 ms. Each color was shown as a square

(size 120px × 120px) in the center of a mid-grey screen. There were two forced choice labeling tasks: forced choice labeling task 1 consisted in pressing one of 4 keys to label the color just shown. The keys were labelled 'W' for white, 'G' for grey, 'BLU' for blue and 'BLA' for black. In forced choice labeling task 2, they pressed one of 2 keys labelled 'LB' for light blue and 'DB' for dark blue. Forced choice labeling task 1 was followed by forced choice labeling task 2 for all the participants. The position of the labels on the keyboard was randomized and counterbalanced across participants. Participants had to use their index finger to answer and, after having pressed the key, to return with the finger to a marked point outside the keyboard equidistant to all keys.

2.2. Structure of the main study

The study took place during two consecutive days and were organized as follows (Table 1). On Day 1, participants were administered the Ishihara colorblindness test (3.5.), followed by the forced choice labeling task 1 and 2 (2.6.) and by the first training session on artificial color categories with SGs (2.8.). On Day 2, the second training session with SGs was followed by the discrimination tasks (Same Different task and Odd One Out task) and finally the EEG session (2.9).

Session	Task	Methods	Results	Tables	Figures
Day 1	Colorblindness test	2.5	-	-	-
	Labeling	2.6	3.1	-	-
	SG training	2.8	3.2	-	1
Day 2	SG training	2.8	3.2	-	1
	Discrimination tasks:	2.10	3.3	3	
	Same/Different	2.10.1	3.3.1		
	Odd one out	2.10.2	3.3.2		
	EEG session	2.9	3.4	2	3

Table 1. Structure of the main study: organization of the tasks on two successive days. The table indicates the Methods section in which each task is described, and the Results section and the Tables and Figures in which results are reported.

2.3. Participants

28 native English speaking students from The University of Auckland participated in this study and received a NZD 20 voucher for their participation. 9 were excluded: 4 participants left the experiment after the first day, 2 due to a technical problem with the EEG, 2 due to an error in the configuration of the task and 1 because she consumed psychoactive substances the day before. Data from the remaining 19 participants were analyzed (M= 21.22 years, S.D.=3.42 years, 10 women). All participants reported normal or corrected-to-normal vision and they were right handed. Color vision was evaluated by means of the Ishihara colorblindness test (3.5.).Participants provided written informed consent and all research was approved by The University of Auckland Human Participants Ethics Committee.

2.4. Materials

From the forced choice labeling data (2.1.), we identified a color that was called light blue and dark blue with equal frequency across participants. We decided to use this as the first color in the 5 color array, and constructed it with 5 colors considered as 'dark blue' by the group of English speaking participants of our forced choice labeling study. We also identified another color which was labeled 'dark blue' and 'black' with a frequency of 0.5 each. This was selected as the fifth color in the array. Finally, we identified three transition colors by interpolating the terminal colors in color space.

2.4.1. Selection of terminal colors

For each participant, we calculated the number of repetitions (maximum 10) in which each of the five colors in the array was labeled with 'white', 'grey',

'blue' or 'black' in forced choice labeling task 1 and 'light blue' or 'dark blue' in forced choice labeling task 2.

Since we wanted to form a uniform color space, we needed to use colors in the blue hue that differ in lightness but this difference had to be minimal: just enough to be perceived as a different color. For this reason we decided to build an array of either light or dark blue colors.

If we used dark blue colors, the criterion for the selection of the first color in the array was the following: for color X to be the first of the 5 colors, it has to be labeled using the basic term 'blue' on average with a frequency F of approximately 1 and with the non-basic terms 'light blue' and 'dark blue' on average with equal frequency (ideally F=0.5). If 'light blue' and 'dark blue' had approximately equal F it would suggest X lies at the boundary between the lexical categories 'light blue' and 'dark blue'. For color Y to be the fifth of the 5 colors, it has to be labeled using the basic terms 'blue' and 'black' or 'grey' on average with a frequency F of approximately 0.5.

On the other hand if we used the light blue colors, for color X to be the first one in the array, it has to be labeled using the basic terms 'blue' and 'white' with a frequency F of approximately 0.5. This would suggest that X lies at the boundary between the lexical categories 'blue' and 'white' or 'grey'. For color Y to be the fifth of the 5 colors, it has to be labeled 'blue' on average with a frequency F of approximately 1 and with 'light blue' and 'dark blue' on average with equal frequency (ideally F=0.5).

As potential colors, we identified two Munsell color that met these criteria inside the dark blues category: for color 1 in the array, Munsell color 5PB5/12 (B is the mean frequency by which the color is labeled 'blue', LB is the mean frequency by which the color is labeled with 'light blue', DB with 'dark blue'): B= 0.84, SD=0.28; LB= 0.01, SD=0.38, DB= 0.9,SD=0.30. For color 5 in the array, Munsell color 5PB1/4: Blue= 0.56, SD= 0.32; Grey= 0.21, SD=0.38, Black= 0.19, SD= 0.26.

2.4.2. Selection of the middle and transition colors

Transition colors were defined as those within our stimulus set that interpolated in color space the terminal and middle colors chosen based on statistical criteria: 5PB4/10 as a lighter dark blue (frequency with which participants called it 'dark blue'=0.57; SD= 0.31.), 5PB3/8 as middle dark blue ('dark blue'=0.81 SD=0.29) and 5PB2/6 as a darker dark blue ('dark blue'=0.93; SD= 0.07). The final 5-color array was as follows (cells show mean forced choice labeling frequencies):

	5PB 5/12	5PB 4/10	5PB 3/8	5PB 1/4	2.5PB 2/6
'light blue'	0.51	0.37	0.13	0.11	-
'dark blue'	0.4	0.58	0.83	0.86	0.92
'grey'	0.09	0.08	0.12	0.14	0.20
'black'	-	0.02	-	-	0.19

2.4.4. Artificial color terms and categorizations

We used the two syllables 'to' and 'nu' as artificial basic color terms to be used as signals in SGs. There were four possible categorizations of the 5-color array, arising out of the combination of the 2 color terms and 5 colors, allowing only sets of either 2 or 3 colors. In categorization 1, 'to' was associated to colors 1 and 2, and 'nu' to colors 3, 4 and 5. In categorization 2, 'to' was associated to 1, 2 and 3, and 'nu' to 4 and 5. In categorization 3, 'nu' was associated to 1 and 2, and 'to' to 3, 4 and 5. In categorization 4, 'nu' was associated to 1, 2 and 3, and 'to' with 4 and 5. Five participants were randomly assigned to categorization 1, six to categorization 2, four to categorization 3 and five to categorization 4 on Day 1 (the number of participants assigned to each was equal before discarding participants from the original 28 volunteers group).

2.5 Colorblindness test

Participants' color vision was evaluated by a computerized version of the Ishihara (1917) colorblindness test to exclude a possible color vision impairment. This was carried out on Day 1 before the forced choice labeling tasks. All participants included in the experiments passed the Ishihara test.

2.6 Testing native color categories: forced choice labeling task

On Day 1, after the colorblindness test, participants performed the same forced choice labeling task used in Experiment 1 (see Chapter 2) but this time with the 5 colors array of blues. The purpose of this forced choice labeling task on Day 1 was to assess how participants categorize the blues in the 5-color array, in order to determine whether and how learning in the SG would change native color categories. The stimuli and procedure where the same as in the preliminary forced choice labeling study (2.1.). For each participant, we calculated the number of repetitions (maximum 10; 2.1.3.) in which each of the five colors in the array was labeled with 'white', 'grey', 'blue' or 'black' in forced choice labeling task 1 and 'light blue' or 'dark blue' in forced choice labeling task 2. We focused in particular on the middle color (5PB 3/8). Participants were classified as having either of two native color categorizations of the 5-color array, depending on whether the middle color was grouped with the 'light blues' (N=2) or the 'dark blues' (N=17).

2.7. Optimal color frequencies

The optimal frequency distribution over colors drawn from the 5-color set is: 0.15 for each of the terminal colors, 0.20 for each of the transition colors, and 0.30 for the middle color. This frequency pattern optimizes learning for the receiver, as learning time is minimized, the accuracy he can obtain by choosing randomly colors of the same category in response to each signal is maximized (i.e., he need not know the frequency distribution of colors), and the accuracy he would obtain with informed choices is minimized

(i.e., there is little incentive to learn the color frequency distribution). This result was used to set the frequencies of colors as shown to the sender/computer in the SG training in the main experiment (2.8.) and in the color discrimination studies (2.10).

2.8. Learning artificial color categories: the Signaling Game

Each participant learned one of the 4 possible categorizations (2.4.4.) of the 5-color array using the 2 artificial color terms, as follows. Participants played a Signaling Game (SG), identical to the SG in the computer simulations previously described (Chapter 2, 2.2). At the beginning of the game, participants were randomly assigned to one of 4 languages: the color categorizations and vocabularies of 2.4.4. During the game, in each trial, a color invisible to the player was randomly drawn from the 5-color set, according to the color frequency pattern: 0.15 for each of the terminal colors, 0.20 for each of the transition colors, and 0.30 for the middle color (see 2.7.). The color term corresponding to the randomly-drawn color was shown to the player, whose task was to choose from the 5-color array one of the 2 or 3 colors belonging to the category denoted by the observed term. Feedback was given in each trial as to whether the random color and the color chosen by the player matched or not. The computer played as the sender, and the participant is the receiver, whose task is to decode the signals he receives from the computer, and infer the correct color categorization and color vocabulary.

2.8.1. Procedure

Participants were seated in a darkened booth. Stimuli were presented on a 21- inch LCD monitor (60 Hz refresh rate) using Presentation® software (Version 0.70). The monitor was positioned at 57 cm from the participant.

At the start of a game, the participant/player was shown both terms ‘to’ and ‘nu’ at the center of the screen for 300 ms, followed by the 5-color array with colors in random order. In each trial, the participant/player observed a color term (shown in capital letters with font Calibri size 48 in the middle of

a mid-grey screen), and had to choose one of the colors associated to that term in the language by pressing on a full-size keyboard one of five keys, spatially associated to the 5-color array, which was shown on a single screen with colors in random order. The participant saw the array with colors, now ordered by lightness (light blue to dark blue), as feedback, with two arrows: one pointing to the randomly drawn color, associated to the observed color term in the language (the arrow was marked with the label 'AS', 'associated'), the other pointing to the color selected by the player as a response to the observed term (marked with the label 'CH', 'chosen').

Training with SGs consisted of two sessions, one on Day 1 and one on Day 2. Each session lasted until the participant had reached to 60 correct responses. At the end of the game, the participant had to select the categorization he thought he had learned. He was presented with all possible categorizations and vocabularies on a single screen, shown as four arrays with the terms 'to' and 'nu' indicating the different color-term combinations, and he had to press a key from 1 to 4 to choose one. They were asked to confirm their choice in a second trial identical to the first (Fig.1).

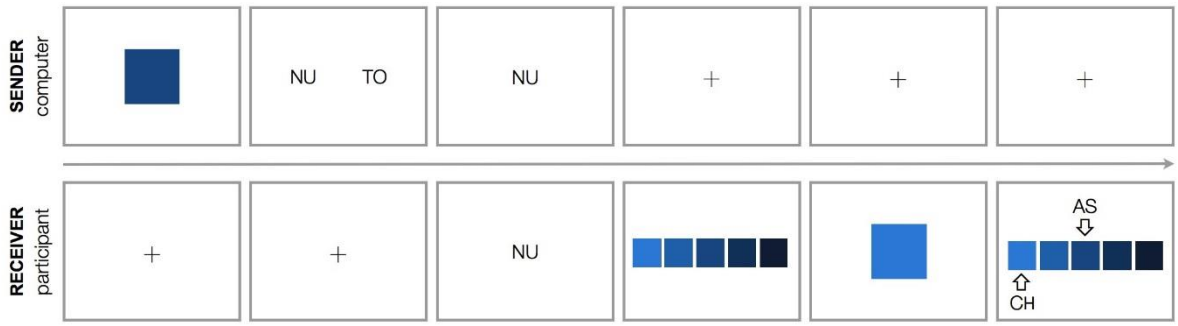


Figure 1. A trial of the signaling game used in the behavioral training sessions. The top and bottom rows show what the sender/computer and the receiver/participant see on successive screens, respectively. Roles are fixed at the start of the game. Time flows from left to right. The sender/computer plays according to a fixed categorization of the 5-color array, which must be learned by the receiver/participant. In each trial, the sender sees one color, and sends the signal associated to the category the color belongs to. The receiver sees the signal, and must choose from the 5-color array the one that may have been seen by the sender by pressing one of five keys on a standard full-size keyboard. The feedback shown to the receiver

indicates, by means of two sets of arrows and labels, the color associated to the signal (AS) and privately seen by the sender, and the color chosen by the receiver as a response to the signal (CH). Over trials, the receiver will learn what colors are associated to each signal, i.e., the color categorization the sender is playing by.

2.9. Neural consequences of learning color categories: EEG study

2.9.1. Stimulus presentation

The last task administered on Day 2 was passive visual exposure to color-term and term-color sequences, which were either Learned (agreeing with the categories and vocabulary learned in the SG) or Incongruent (deviating from the learned set). The stimuli were the 5 colors and the 2 color terms used in the training with SGs. There were two blocks, whose order in each session was randomized across participants: in one block, a color term was shown for 300 ms (capital letters, in Calibri font size 48, in the middle of a mid-grey screen), followed by a color for 400 ms (shown as a square of 120×120 pixels) in the middle of a mid-grey screen and by a fixation cross for 1250 ms (term-color or TC sequences); in the other block, a color was shown for 400 ms, followed by a color term for 300 ms and by a fixation cross (color-term, CT sequences). Each block had 200 Learned and 200 Incongruent trials. Across trials, all possible combinations of colors and color terms, in both CT and TC sequences, were used. Moreover, all colors were presented with equal frequency, in contrast with the non-uniform frequency distribution used in the SG training (2.8.). This was effected in order to avoid that color-position effects (e.g., to terminal versus intermediate colors; see Chapter 2) were confounded with frequency effects.

2.9.2. Data acquisition

The electroencephalogram (EEG) was recorded using a 128-channel Geodesic Sensor Net and NetAmps 300 amplifier (Electrical Geodesics Inc., EGI). It was digitized at 1000 Hz and acquired referenced to the vertex electrode. Individual sensor impedance was kept below 40 kΩ and measured

both prior to and halfway through the experiment. Offline, the data was analyzed using EEGLab toolbox (Delorme & Makeig, 2004) and Fieldtrip (Oostenveld et al, 2011) for Matlab (The MathWorks Inc., Natick, Massachusetts). Data were re-referenced to the average. The data were segmented into 1200 ms segments, from 200 ms before stimulus presentation to 1000 ms after. The sampling rate was 1024 Hz. All filtering was digital.

2.9.3. Data analysis

EEG data were epoched from -200 ms to 1000 ms relative to the onset of the second stimulus (i.e., the color term in CT sequences, and the color in TC sequences), and were baseline corrected using data from the −200 to 0 ms interval. Segments were discarded if they contained activity exceeding ±100 μV thresholds in any channel. Eye blinks in the bilateral fronto-polar channels E8, E25, E127 and E126 were identified using a blink detection function by means of the following procedure: the function finds artifacts that are shaped like a typical blink, which is represented by a Chebyshev function. To find this shape, the function computes the covariance between a segment (window) of the epoch and the Chebyshev function. The larger the covariance, the more evidence there is that a large blink-shaped voltage deflection is present in that window. We specified a Blink Width of 400 ms and a Test Period of -200 to 1000 ms. The function first computes the covariance between a 400-ms wide Chebyshev function and the EOG waveform from -200 to +400 ms and continues this process for successive 400-ms windows. It then compares the largest of these covariance values to the threshold, and the epoch is marked for rejection if the largest covariance exceeds the threshold.

ERPs were computed by averaging over artifact free epochs from each condition (Learned/Incongruent), in each block (CT/TC), for each participant separately. The definition of Learned and Incongruent trials for a given participant depended on the particular color categorization that he/she was trained on during the SG sessions. Finally, grand-average ERPs were computed by further averaging over participant specific averages.

Statistical analyses of ERP effects were conducted by means of the following procedure (Maris and Oostenveld 2007): (1) participant specific ERP averages were compared between the Learned and Incongruent conditions from each channel and time point with dependent samples *t*-tests; (2) data from neighboring time points and channels in which p-values were smaller than 0.05 were clustered together; (3) the cluster-level *t*-statistics was computed as the sum of *t*-values from all samples belonging to the cluster; (4) the cluster-level p-value was estimated using a Monte Carlo simulation: participant specific ERP averages across all samples in a cluster from both experimental conditions were collected in a single set; this new set was randomly partitioned into two subsets of equal size; the subsets were compared by means of a *t*-test; these steps were repeated 1000 times; a cluster-level p-value was computed as the proportion of partitions that resulted in a larger *T*-statistic than in the observed ERP data.

2.10. Probing color categories: color discrimination studies

After the second day of training, participants performed a same/different color discrimination task modified from Liu et al. (2010), and an odd-one-out discrimination task modified from Witzel & Gegenfurtner (2013).

2.10.1. Same/different discrimination task

Participants were presented with a colored square (155×155 pixels) surrounded by a colored frame (190×190 pixels) in the center of a mid-grey screen, and they were asked to judge, by pressing either of two keys on a full-size keyboard, whether the square and the frame were of the same or of a different color. If the square and the frame were of the same color, this was considered a 'same' trial. If the square was of a different color than its surrounding frame, it was considered a 'different' trial.

In each trial, first a fixation-cross appeared for 1000 ms at the center of the screen, followed by a square surrounded by a frame shown for 200 ms,

and by a blank mid-grey screen shown for 800 ms. The stimuli consisted of 25 color-frame combinations (5×5 colors) re-used in 300 trials. Of these 300 trials, 150 showed the inner square and the surrounding frame in the same color, and 150 showed them in different colors. In 50% of the different color trials, the colors of the square and of the frame were associated to the same term in the artificial color system that each participant had learned during the SG training ('within category' or WC trials). In the other 50% of the different color trials, stimuli were constituted by squares with their surrounding frames in a different color, which participants had learned to associate to different color terms ('across category' or AC trials).

2.10.2. Odd-one-out discrimination task

We used the spatial 4-Alternative Forced-Choice (4AFC) discrimination task from Krauskopf and Gegenfurtner (1992), also used by Witzel and Gegenfurtner (2013). In each trial, observers were shown 4 colored disks. One of these colors differed in lightness from the other 3. Participants had to indicate which disk was different by pressing one of 4 keys corresponding to the 4 positions in space of the disks.

At the start of a trial, a fixation cross appeared for 1000 ms at the center of the screen, followed by the 4-disk stimulus shown for 500 ms against a mid-grey background, and by a blank mid-grey screen shown until a response was given. The odd disk occupied any of the 4 possible positions on the screen, randomized across trials. In 50% of the trials, the color of the odd disk was associated with the same color term as the color of the 3 other disks in the color categorization that the participant had learned during the SG training sessions ('within category' trials). In the other 50%, the color of the odd disk was associated with a different color term than the color of the other 3 disks during the SG training ('across category' trials).

2.10.3 Analysis of color discrimination data

Paired *t*-tests were conducted to assess the differences in the performance of the participants in the two tasks across conditions. Differences in reaction times and accuracy were compared between across-category and within-category trials. Accuracy was determined by the number of correct trials in the last forth part of each of the trainings (where coordination is expected to be present already) divided by a fourth of the total amount of trials the participant needed to end that training session. To analyze the contrast between within and across category trials, for each of the discrimination tasks (Same-Different and Odd One Out), we ran two-way ANOVAs on within- and across-category trials. For each task, 2 ANOVAs were run with the reaction times of subjects to the each of the two tasks and 2 ANOVAs with the accuracies obtained by subjects in each task. This was done first with all 5 colors in the array as factors and then with each position as factors: terminal colors (colors 1 and 5 together), transition colors (colors 2 and 4 together) and middle color (color 3). Moreover, Wilcoxon's rank sum tests were conducted comparing the means of the participants' performance in reaction times and accuracies, for all 5 colors separately and for the 3 positions.

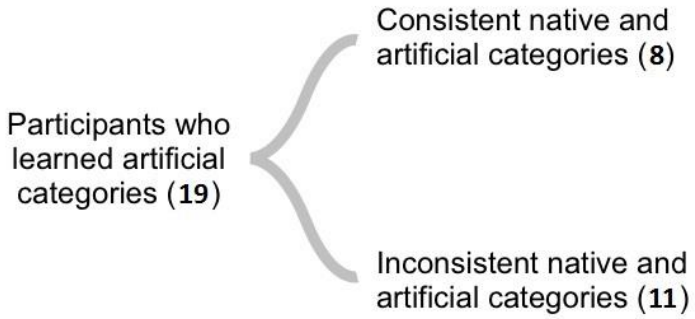
3. Results

3.1. Forced choice labeling task

All colors were labeled 'blue' with an average frequency F of 0.86 (color 1 F= 0.99;color 2 F=0.99; color 3 F=0.96; color 4 F=0.935; color 5 F=0.425). When asking whether the color was dark blue or light blue (forced choice labeling task 2), participants were grouped according to whether they had categorized the middle color (color 3) into either 'light' or 'dark '. Choices had to be stable in at least 60% of all trials. Out of the 21 participants, 18 grouped the middle color together with the dark blues (average frequency across participants in this group of the color being called 'dark blue', F=0.88), 1 grouped it with the light blues (frequency of the color being called 'light blue'

by this participant, F =0.6). Two participants labeled it 'dark blue' and 'light blue' with equal frequency (F =0.5). We ended up with 2 groups, one of 8 participants who learned a system consistent with their native category, and the other of 11 participants who learned a categorization inconsistent with their native one (Fig. 2).

Figure 2. Tree diagram illustrating the categorizations learned by participants (N=19). During the training on Day 1, 8 participants learned artificial color categories that were consistent with their own native color categories, and 11 learned artificial categories that were inconsistent with their own native categories.



3.2. Learning color categories and vocabularies

Nineteen out of twenty-one participants learned the artificial color system during the first training session with signaling games, and 18 chose the right categorization when they were asked at the end of the second training session. The average number of trials it took participants to get to 60 correct responses and finish a session was higher for the first training than for the second: 199.62 trials (SD=109.96) and 183.52 (SD=42), respectively. Nonetheless, this difference was not significant (Wilcoxon rank sum tests V= 110, p=0.8649).

Accuracy during training on the first day of training (mean= 36.54 SD=9.75) was similar to the accuracy level on the second training day (mean=

36.77 SD=6.69; paired *t*-test *t*(20)= 0.0944, *p*= 0.0944). Accuracies during both days differed significantly from chance level and from optimal accuracies by informed choices (one-sample *t*-tests, $p<0.001$). Accuracies during both days were not different from optimal accuracies by random choices, and from the accuracies obtained in the computer simulations (one-sample *t*-tests, $p>0.2626$).

3.3. Discrimination studies

3.3.1 Same/different discrimination study

Reaction times in within-category (WC) (RT M=510 SD= 56.53) trials were longer but difference was not statistically significant than in across-category (AC) trials (RT M=484.79 SD=52.47) (Paired samples *t*-test, *t*(20)= 1.8858 *p*= 0.074). There were no significant differences either between same-color trials (RT M=457.25 SD=110.49) and WC trials (Paired *t*-test, *t*(20)=1.7067, *p*=0.1034), or between same-color trials and AC (Paired *t*-test *t*(20)=0.8738, *p*=0.3926), or between different color (i.e., WC+AC trials, RT M=464.369 SD=127.9)and same-color trials (Paired *t*-test *t*(20)= 1.3137, *p*=0.2038).

Participants were more accurate in the across category trials (accuracy M=0.92 SD=0.014) than in the within category trials (accuracy M=0.88 SD=0.083)(AC/WC trials: paired *t*-test, *t*(20)= 2.1898, *p*=0.0406; AC/same-color (accuracy M=0.88 SD=0.14) trials: paired *t*-test, *t*(20)=1.19, *p*=0.248; WC/same-color: paired *t*-test, *t*(20)=0.0522,*p*=0.9589; different/same trials: paired *t*-test, *t*(20)2.3154=, *p*=0.0313). When comparing WC to AC for each position, Wilcoxon rank sum tests showed statistical differences in reaction times only for the terminal colors (Terminal: *V*=44, *p*=0.011; Transition: *V*=105, *p*=0.733; Middle: *V*=81, *p*=0.2428). They also showed statistical differences in accuracies for the terminal and the transition colors (Terminal: *V*=188, *p*=0.002; Transition: *V*=166, *p*=0.0045; Middle: *V*=108, *p*=0.8077) (Table 4). The ANOVA showed no interactions between the position of the color and condition (WC/AC) for either reaction times (*F*(144)=0.122, *p*=0.885) or accuracies (*F*(144)=0.060, *p*=0.942).

When comparing WC to AC for each color, Wilcoxon rank sum tests showed significant differences in reaction times only (Table 3) for color 5 (color 1: V=75, p=0.1678; 2: V=106, p=0.7593; 3: V=81, p=0.2428; 4: V=118, p=0.9457; 5: V=42, p=0.009016), while in accuracies (Table 3) for colors 1 and 2 only (1: V=189.5, p=0.001688; 2: V=127, p=0.002433; 3: V=108, p=0.8077; 4: V=145, p=0.313; 5: V=154, p=0.06966). The ANOVA showed no interactions between the 5 colors and condition (WC/AC) for either reaction times (F(190)=0.187, p=0.945) or accuracies (F(190)=0.105, p=0.981).

In all tests, the Bonferroni-corrected α is 0.01.

3.3.2. Odd-one-out study

Reaction times in WC (RT M=538.33 SD=249.28) trials were longer than in AC (RT M=495.41 SD=180.79) trials (Paired t-test, t(104)= 2.7497, p=0.007). Accuracy was better in AC (mean=0.43 SD=0.13) trials than in WC (mean=0.40 SD=0.16) trials (Paired t-test, t(104)=3.4343, p=0.0009). When comparing WC to AC for each position, Wilcoxon rank sum tests showed statistical differences in reaction times only for the terminal colors (Terminal: V=44, p=0.01135; Transition: V=67, p=0.0958; Middle: V=127, p=0.7079). They also showed statistical differences in accuracies only for the transition colors (Terminal: V=127, p=0.7079; Transition: V=173, p=0.001816; Middle: V=166.5, p=0.3979) (Table 4). The ANOVA showed no interactions between the position of the color and condition (WC/AC) for either reaction times (F(144)=0.967, p=0.383) or accuracies (F(144)=1.061, p=0.3493).

When comparing WC to AC trials for each color, Wilcoxon rank sum tests showed no statistical differences in reaction times (Table 3) (1: V=48, p=0.01755; 2: V=67, p=0.0958; 3: V=127, p=0.7079; 4: V=76 , p=0.179; 5: V=56, p=0.03844) and in accuracies (Table 3) for colors 1 and 2 only (1: V=142 , p=0.002088; 2: V=168, p=0.003521; 3: V=116.5, p=0.3979; 4: V=134, p=0.1212; 5: V=90, p=0.5883). The ANOVA showed no interaction between the 5 colors and condition, for reaction times (F(190)=01.041, p=0.387) or accuracies (F(190)= 0.682, p=0.60501).

In all tests, the Bonferroni-corrected α is 0.01.

Same - Different	Reaction Times		Accuracies	
Color	WC	AC	WC	AC
1	464.64 (143.28)	451.97 (127.88)	0.82 (0.30)	0.91 (0.2)
2	464.95 (151.1)	464.81 (199.68)	0.82 (0.27)	0.91 (0.19)
3	475.04 (123.45)	461.15 (136.14)	0.87 (0.24)	0.85 (0.27)
4	461.03 (134.39)	467.87 (126.08)	0.86 (0.29)	0.89 (0.26)
5	480.72 (124.54)	455.28 (113.65)	0.84 (0.31)	0.89 (0.22)

Odd One Out	Reaction Times		Accuracies	
Color	WC	AC	WC	AC
1	519.9 (231.67)	461.85 (171.99)	0.39 (0.44)	0.43 (0.47)
2	540.72 (416.99)	471.67 (182.08)	0.38 (0.44)	0.42 (0.47)
3	515.84 (361.34)	529.34 (236.53)	0.41 (0.45)	0.46 (0.49)
4	556.44 (300.21)	506.19 (293.26)	0.42 (0.45)	0.42 (0.46)
5	558.75 (233.29)	508.02 (235.49)	0.42 (0.44)	0.42 (0.46)

Table 3. Results of color discrimination experiments. Mean values (and SD) are reported. Abbreviations: WC (within-category condition); AC (across-category condition).

Same Different task

Position	Reaction times		Accuracy	
	V	p-value	V	p-value
Terminal	44	0.01135	188	0.002067
Transition	105	0.7335	166	0.004537
Middle	81	0.2428	108	0.8077

Odd One Out task

Position	Reaction times		Accuracy	
	V	p-value	V	p-value
Terminal	44	0.01135	127	0.7079
Transition	67	0.0958	173	0.001816
Middle	127	0.7079	166.5	0.3979

Table 4. Results of the Wilcoxon rank sum tests comparing WC to AC for each position. V and p values are reported.

3.3. EEG

In ERPs we found no incongruency effects in either CT or TC trials that involved the middle or transition colors, i.e., the best exemplars of each category. Instead it was one of the terminal colors, the fifth in the array, which showed ERP effects. The effect is a positive-going ERP wave that appears at 500 ms from stimulus onset and lasts until end around of the epoch (Fig. 3). The cluster size was 15666 (number of adjacent time-channel samples); the sum of *T*-statistics in the cluster was 48555 and the Monte Carlo p-value=0.001. Interestingly this effect was found only in the Color – Term (CT) condition and not in the Term Color for color 5. We did not find any significant clusters for color 1, 2, 3 and 4 in either of the condition (in all clusters $p>0.05$).

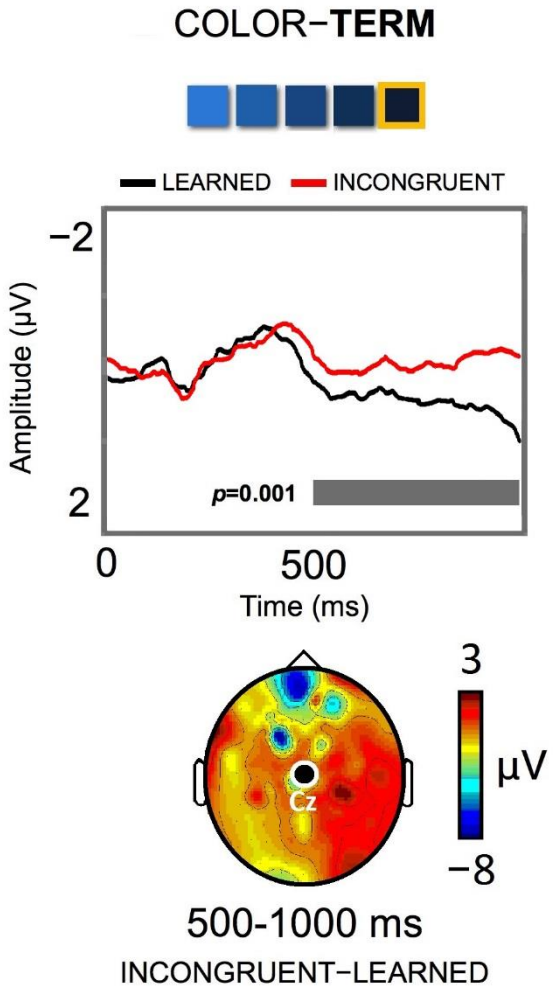


Figure 3. Event related brain potentials (ERPs) evoked by the color term presented after the fifth color in the array in CT sequences. The onset of the eliciting stimulus (i.e., the color term in CT sequences) is at 0 ms. Negative amplitude values are plotted upwards. Single-channel ERP plots show grand-average (N=19) waveforms from Learned (black) and Incongruent (red) trials from the vertex channel Cz. Dark grey bars in the lower portion of the plot represents the highest ranked cluster of adjacent time points at which statistically significant effects were found. The topographic map displays grand-average (N=19) differences between Incogruent and Learned trials in the time window of the highest ranked cluster.

4. Discussion

In this study we intended to further investigate the functional reorganization of brain circuits, observed in the previous experiment (Chapter 2), through the acquisition of a new way of categorizing colors. This second experiment used an array of blue colors varying along the lightness dimension, whereas the first experiment used colors from the brown to green range, thus varying along the hue dimension. In the previous EEG study (Chapter 2), we found the strongest differences in ERP signals between learned and incongruent trials at the middle color in the array, that is, at the color that constitutes the category boundary in participants' native categorization of browns and greens. That middle color is ambiguous between two lexical categories, and the artificial color categorization was precisely designed to override that ambiguity by placing the ambiguous green-brown color into either the green or the brown sets. The observed ERP effects can therefore be taken to reflect the neural outcomes of a costly learning process, which consists precisely in classifying an originally ambiguous color into a sharply defined category. In the first experiment, however, the boundary of the native categories of brown and green, and the boundary of the new artificial categories, are very close and (in some participants) may even coincide. Here, we wanted to tease apart these two aspects, and produce a color array in which the artificial color category still sets a boundary in the middle of it, but the ambiguous color in the participants' original categorization lies at the terminal position, instead of at the middle one. If indeed ERPs reflect rewiring effort, then they should be largest at the terminal color in the second experiment. The results of this experiment confirm our predicted effect for terminal colors. They are moreover qualitatively in line with the results of our previous experiment, both in terms of latency and topographical distribution. Clifford et al. (2012), Özgen and Davies (2002) and Goldstone (1994) proposed that category learning enhances sensitivity around new category boundary regions. However, we did not have an effect at the color that sets the new category boundary. Instead, the most ambiguous color in their native category was the one what showed a cluster of significance.

The fact that we did not find any category effects in early ERP components but instead observed a late N400-like effect, is consistent with Clifford et al. (2012) and He et al. (2014), as well as with the findings presented in Chapter 2, suggesting that the effects of acquiring a new color categorization are not limited to perceptual processing alone.

Interestingly, when analyzing color by color we found that this effect is only statistically significant for one of the two terminal colors (i.e. only for color 5) that formed the extremes of our color array. It is important to notice that although color 1 is not ambiguous, it was labeled with the non-basic terms light blue and dark blue with roughly the same frequency (≈ 0.5). As we needed as stimuli 5 colors with minimal perceptual differences between one color and the one(s) next to it, we did not use the whole range of blues, but only the darkest blues of the pool of Munsell colors (we also selected a possible array with lighter blues but after statistical analyses of the labeling frequencies, the dark blues resulted more suitable for our purpose). If the 5 colors array continued before color 1 and after color 5, to the left we would have light blue colors and to the right black colors. For English speakers, a light blue color will still be called with the basic color term "blue" because it is only a lighter version of the prototypical blue, but color 5 was called "blue" and "black" with approximately equal frequency. Therefore, color 1 could be considered ambiguous, but between two non-basic color categories. Color 5 stands out for being the only color in the array that lies between the two basic color categories of blue and black, and was thus the most likely candidate to trigger the learning-related ERP effects observed in the first experiment (Chapter 2). This finding highlights the neural reality of the distinction between basic and non-basic color categories and terms. Rewiring effort was observed in both experiments for colors that lie at the boundary between basic color categories.

The results of the discrimination tasks also point to a special status of the terminal colors in the array. The participants' responses to them, in contrast with the transition and middle colors, were faster and more accurate when the target and the distractors were of a different artificial category rather than of the same one. We believe this may also to a certain degree support our hypothesis that ERPs reflect learning effort of the tasks, but in a

different way than experiment 1. Terminal colors are ambiguous: they may belong to one or two color categories for English speakers. Assigning an artificial color term to a color that constitutes a boundary between two colors in the native language of the participants may be more demanding and bring along discrimination problems. Therefore the system could be responding with stronger signals when the stimuli are incongruent with these more difficult colors to learn, relative to the most prototypical colors like the middle color (color 3). Still, our data cannot answer the question of how the presence of a boundary between two color categories affects the ERP and discrimination effects.

The effect we find for the ambiguous colors also confirms the fact that the average amplitude of ERP effects we found in experiment 1 does not depend on the frequency with which that color was presented in the training sessions, since we did not find any effects in the middle color which was the one most frequently presented in experiment 2.

Crucially, the fact that we see differences between learned and incongruent trials indicates that ambiguity of the color per se is not what drives the ERP effects. Furthermore, if ambiguity was effectively the critical feature of the stimuli, color 5 would show the largest ERP effect in Term-Color condition trials as well, which is not the case.

5. Conclusion

In chapter 1 we mentioned a hypothetical situation in which a group of monolingual tourists on an island, who have different native languages, are observing the horizon on the sea. We wondered whether they would look at it as the boundary between one color category and another or think of the sea and the sky as being of the same color category. Now consider one of them as having to different colors for the sky and the sea in her native language. She points to the sky and teaches a man in the group the name for that color in her language. Then the same for the color of the sea, and does this throughout the day. Will this man *see* these colors the next day differently? On possibility is that they would *perceive* the colors in the same way as before but they would

pay more attention to the newly learned color categories and process more effortfully colors that are ambiguous in their native languages.

On the other hand, Thierry et al.(2009) studied the effects of color terminology in Greek and English on early stages of visual perception using the vMMN, an electrophysiological index of perceptual deviancy detection. Greek speakers use two terms that are equivalent to the English term blue. The visual mismatch negativity was similar for blue and green in English participants, but there was greater distinction between different shades of blue than different shades of green in Greek participants. Future investigations should be done on what the effect of learning this artificial categorization of the blues would be on speakers of languages that use more color categories inside the blue hue.

General discussion

The aim of our research was to recreate in the laboratory a situation in which speakers of different languages, with possibly different color systems, interact, and one of them acquires a new color categorization from the other in a communication game (signaling). To this end, we set out by investigating two actual language groups co-existing in the same region.

Fieldwork

We analyzed how speakers of Galician and speakers of Spanish named 84 different Munsell colors, and we studied the similarities and differences in color categories within and across the two language groups. We observed differences in how the two languages set boundaries between the categories of yellow and green (hue), and within the blue category (lightness). These kind of findings are not new in the field. The EoSS project has enabled research on closely related languages such as Galician and Spanish. By using an objective referential grid of comparison across languages it is possible to investigate how the boundaries of words change across languages. In one of the studies that stems from this project, Majid et al (in press) analyzed data from 12 Germanic languages on 4 domains: color, body parts, containers, and spatial relations. Even if languages were very similar in the color domain, differences across languages were found in the choices of speakers on the best example of the color category. Also, there was a difference of 33 color terms between the language that had the most distinctions (English) and the one with the fewest (Faroese).

One interesting finding of our study is that Galician is a language spoken in a country where Spanish (Castilian) is the dominant language. Most inhabitants of Galicia are competent Galician speakers, but only 44% use the language regularly and 45% only occasionally[2]. Moreover, our participants were young adults. The percentage of young adults speaking Galician has

[2] I.G.E. Instituto Galego de Estatística. © Xunta de Galicia.

dropped from 28.49% in 2003 to 18.59% in 2008. This suggests Galician is becoming more and more a minority language. Importantly, we still found differences in how Galician speakers categorize colors, relative to Spanish speakers. However, these differences require further research.

In Chapter 1 we asked whether there would be differences in the way a Spanish speaker and a Galician speaker, who are looking at the same green tree, process the colors of leaves given they categorize this color differently. Our answer may be: possibly so, if only for a narrow set of colors in the relevant color range. Such differences are however more likely to play out in communication than in perception. Our laboratory experiments made use of signaling games to re-create a situation in which, from an initial condition of potential disagreement among speakers, one of them (the participant) will learn the color categories of the other (the computer) by partly rewiring her own native color system.

Signaling games

Moreno & Baggio (2014) showed that signaling games are a viable laboratory model of the acquisition of signals and their meaning. Here we used signaling games with fixed roles where the receiver (participant) adjusts her mapping and the sender (computer) sticks to its original categorization. An important step in our laboratory experiments was to determine whether the color categories we constructed were learnable. The nature of feedback participants obtain is potentially misleading: even if the color they chose was indeed one of those associated with the relevant color term, their exact color choice might not match the particular color as seen by the sender. We had to have a confirmation that this kind of feedback would actually be effective, allowing participants to actually learn the new categories, and not only finish the game. Our computer simulations allowed us to do so, and moreover to set optimal accuracy levels to which our behavioral results can be compared. Participants' accuracies were below optimal accuracies by informed choices: participants were not aware they could perform better by choosing the most probable of the 2 or 3 colors associated to the term they had observed. They

were instead close to optimal accuracies by random choices: they chose at random among the 2 or 3 colors that were associated to the terms, showing they had learned the structure of the color category and not the frequency by which each color appeared on the screen.

We could argue that our games were not actual 'signaling games': in a way they lack strategic uncertainty and therefore real interaction. However, our participants were instructed to learn the artificial language only, they did not know the sender (computer) always played with the same strategy. In fact many participants spontaneously expressed they thought the computer was now and then changing the color categorization.

EEG experiments

The aim of the EEG experiments 1 and 2 was to explore the cognitive and brain processes underlying the acquisition of new color categories. In experiment 1 we used an array of 5 colors that varied in hue from brown to green. The intermediate color in the array was the most ambiguous in both the native color categorization of the participants, and in the newly learned categorization: this color constituted the boundary between two categories.

The results of experiment 1 suggest there is a neural representation of color categories, which contains information on color sets and color terms associated to them. This representation seems to be of a high-level/semantic kind, as suggested by the presence of late endogenous ERP effects. This suggests that the neural changes that accompany the acquisition of color categories are semantic in nature and not strictly perceptual. This is further supported by the fact that early exogenous ERP effects were relatively weak, and not statistically significant.

The fact that we found effects only for color 3 (the most ambiguous one) in the color-term condition and in colors 2 and 4 (the transition colors) in the term-color condition is intriguing. One hypothesis is that ERPs reflect the outcomes of learning effort. Terminal colors can be effortlessly assigned to one or another color category also because they are unambiguous in their

native categorization (this confirmed by our labeling data). The other colors are more ambiguous and more difficult to discriminate. Color 3 in particular is the color that sets the boundary between one and the other category in the artificial language as well as in the native language of participants. Forcing a shift in the category boundary may be challenging. However, this hypothesis does not explain the fact that ERP effects were stronger at the middle color in CT sequences, whereas they were stronger at the transition colors in TC sequences. Moreover, are these ERP results due to the fact that color 3 is ambiguous in the native categorization of participants, and is therefore harder to rewire, or to the fact that it lies at the boundary of two categories in the artificial system? We tried to address this question in experiment 2.

In experiment 2, our array of colors varied in lightness instead, within the blue hue, and the position of the ambiguous color was at one extreme of the array (color 5). This color was labeled as 'blue' and 'black' with equal frequency by an independent group of English speakers. Colors 2, 3 and 4 were non-ambiguous blue colors and color 3 was the most prototypical blue (derived from our labeling results). Although color 1 was considered blue by all our participants in our forced choice labeling study, it was labeled with the non-basic color terms ‘light blue’ and ‘dark blue’ with equal frequency. This makes it a non-basic ambiguous color in spite of it being still ‘blue’.

With this distribution of the colors, we produced a color array in which the artificial category sets a boundary in the middle of it (as in experiment 1), but the most ambiguous color in the participants' native categorization lies at the terminal position, instead of at the middle one.

We did not find an ERP effect at the color that sets the new category boundary but at the most ambiguous color in participants' native category. Again this effect was a late N400-like negativity, only in the CT condition. We did not find earlier low-level/perceptual effects. These ERP results confirm our prediction of an ERP effect at the terminal color and is in accordance with experiment 1. This suggests that the effects we observe reflect the effort of rewiring or translating the native categorization into to the artificial color

system learned in the experiment, a task that is especially demanding for the more ambiguous colors in participants' native color system. However, this line of reasoning cannot explain why we do not see the same effect in the TC condition.

Further research is needed to get a better understanding of the contrasting results we found across conditions in experiment 1 and 2. Before the EEG sessions, in order to keep the participants' attention to the task, we told participants they should keep focused because we would ask questions about the task at the end of it (which we did not). We feel curious about the fact that some participants spontaneously expressed at the end of the experiment that they perceived a difference between the color-term condition and the term-color condition in the EEG task. One of them said "I found the color-term arrangement made much more sense than the term-color. This was because the color coming first was asking a question "which word describes this?" and I could then think of the answer, then I would get told if I was right. Each colour is assigned only one word so I can only be right or wrong. With the other way around, I get the term first so the 'answer' would be any of the set of colours that were assigned to that word. Since each word is assigned many colours, this is difficult to grasp and I felt less engaged." If it was the case for most of the participants to feel lest engaged to one condition than for the other, this task will be a crucial point to revise for future experiments.

Table 1 shows a comparative scheme of the results of experiments 1-2.

Table 1. Comparative scheme of the neurophysiological results of experiment 1 and experiment 2. The colors for which we found significant ERP effects have a yellow frame. In experiment 1 we found a significant cluster of activation in the color-term (CT) condition for color 3, which is the most ambiguous (constitutes a boundary between categories) in the artificial color categorization (indicated with A) participants learned and in the native language (N) of the participants. ERP effects were also found in color 2 and color 4 in the term-color (TC) condition. In experiment 2 an ERP effect was found in the CT condition only for color 5, which is the most ambiguous color in the native categorization (N) of the participants and none was found for color 3, the most ambiguous color in the artificial categorization (A).

Additionally, we took a series of measurements to rule out possible confounding variables affecting the results of our experiments:

- We randomized the order of the colors in the array we presented when participants gave their color choice to insure they were not only learning positions in space.
- We run computer simulations to calculate the optimal frequencies of presentation of the colors. These simulations predict our experimental data in the sense that the fixed roles forced the participant to effectively change her native categorization momentarily, meaning the message gets through. This also helped minimize the possibility that participants learned color frequencies instead of mappings: they actually learned the color categorizations and not only that color 3 was more frequent so they had to choose it more often to end the game;
- We found no correlation between the mean amplitude of the ERP effect across all channels for a color and its frequency during training in the signaling game;
- Roughly half the participants learned an artificial color categorization that was consistent with their native categorization and for half the artificial categorization was inconsistent with their native one. This suggests the observed ERP effects are not directly due to the structure of the native categories themselves, but to how these interact with the learned artificial system.

Our results may shed light into the neural processes underlying the acquisition of a second language. Malt & Majid (2013) highlight after reviewing several studies that the long learning process a person undergoes to master her own native language's naming pattern suggests that learning two or more languages in parallel or a second language implies greater challenge. Research such as Paramei & Menegaz (2013) and Athanasopoulos (2008) have shown that the amount of exposition a speaker has to a new language can affect the way she categorizes colors in her own native language. The effect we found may be the result of a temporary reorganization of the representation of color categories. However, if a longer training was applied, it may be the case this ERP effect changes as the new categorization starts to compete with the native one, mirroring the process of shifting category boundaries mentioned in the studies above. Further research has to be done in order to reach a better understanding the neural consequences of this phenomenon.

One question raised in Chapter 1 was whether there would be any changes in the brain of a Spanish monolingual speaker associated with learning the full meaning of 'azul' in Galician, were he exposed to these colors after during communicative interactions. We may conclude that could be the case. Neural reorganization is expected to take place. Our results are in line with Goldstone (1994) and with Davidoff (2001) that suggest that cultural and linguistic training can affect low-level perception. However, probably not so much at the perception level but at a later more semantic level. Also, a new categorization for the most ambiguous colors in the Spanish speakers' native categorization might be most difficult to learn. However, the mere fact that the new 'azul' is difficult to 'translate' to what in Spanish is an ambiguous color should not be what produces an ERP effect in Spanish speakers as, according to our results, we would still see difference between learned and incongruent trials and no effects in the TC condition.

For further research, it would also be interesting to investigate the possible plastic changes in the brain following training with our paradigm, using magnetic resonance imaging similar to Kwok et al. (2011).

In conclusion, our work shows that, by using laboratory models, one can successfully investigate the neural precursors of the acquisition of novel lexical and semantic categories in language interaction scenarios, involving coordination and communication among agents by means of signals.

Bibliography

Abrams & Strogatz (2003). Modelling the dynamics of language death. *Nature* 24, 900.

Ameel, E., Storms, G., Malt, B.C., & Sloman, S. A. (2005). How bilinguals solve the naming problem. *Journal of Memory and Language*, 53, 60–80.

Aparicio, X., Midgley, K.J., Holcomb, P.J., Pu, H., Lavaur, J.M., Grainger, J. (2012) Language Effects in Trilinguals: An ERP Study. *Frontiers in Psychology* 3:402.

Athanasopoulos,P (2009). Cognitive representation of color in bilinguals: The case of Greek blues. *Bilingualism: Language and Cognition* 12 (1), 83–95 Cambridge University Press.

Bakker, P., & Mous, M. (1994). Mixed languages: 15 case studies in language intertwining. Studies in Language and Language Use 13. *Amsterdam: Institute for Functional Research into Language and Language Use.*

Baronchelli, A. Gong, T., Puglisi, A., Loreto, V. (2010). Modelling the emergence of universality in color naming patterns. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(6), 2403–2407.

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

Bird, C., Berens, S., Horner, A. & Franklin, A. (2014) Categorical encoding of color in the brain. *Proceedings of the National Academy of Sciences, 111,* 4590-4595.

Blume, A. DeJong, D.V., Kim,Y.G., Sprinkle, G.B. (1998). Experimental Evidence on the Evolution of Meaning of Messages in Sender-Receiver Games. *American Economic Review* 88: 1323-1340.

Bowerman, M. & Pederson, E. (1992). Topological relations picture series. In S. C. Levinson (ed.), Space stimuli kit 1.2: November 1992, 51. *Nijmegen: Max Planck Institute for Psycholinguistics.*

Burenhult (2006) N. Body part terms in Jahai. *Language Science*, 28:162–180

Clifford A, Holmes A, Davies IRL, Franklin A (2010) Color categories affect pre-attentive color perception. *Biological Psychology* 85(2):275–282.

Crystal, David. 2000. Language Death. *Cambridge: Cambridge University Press.*

Czigler, I., Balázs, L., Winkler, I., submitted. Pre-attentive formation of feature-conjunction: II. Visuli stimuli.

DeValois R.L., Abramov, I., Jacobs, G.H. (1966). Analysis of response patterns of LGN cells. *Journal of the Optical Society of America*, 56 (1966), pp. 966–977

Dunn, M. Majid, A. & Jordan F. (2010) Biographical Questionnaire for the EoSS Project. Incorporates the Edinburgh Handedness Inventory (Oldfield, 1971).

Dunn, M., Greenhill, S.J., Levinson, S.C., Gray, R.D. (2011) Evolved structure of language shows lineage specific trends in word-order universals. *Nature* 473: 79–82.

Eitan, Z., & Timmers, R. (2010). Beethoven's last piano sonata and those who follow crocodiles: Cross-domain mappings of auditory pitch in a musical context. *Cognition*, Vol. 114, No. 3, pp. 405-422.

Fonteneau, E., & Davidoff, J. (2007). Neural correlates of color categories. *Neuroreport* 18, 1323–1327.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience* 11, 127-138.

Gleason, H.A. (1961). An introduction to descriptive linguistics. *New York: Holt, Rinehart & Winston.*

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology*: General, 123(2), 178–200.

González Perillli, F.,Rebollo, I., Morales-Geribón, N., Maiche, A., Arévalo, A. Blues across two different Spanish-speaking populations (in prep).

Greenberg, J.(1999). Are there mixed languages? In *Essays in Poetics, Literary History and Linguistics Presented to Viacheslav Vsevolodovich Ivanov on the Occasion of his Seventieth Birthday.* Lazar Fleishman et al. (eds), 626-633. Moscow: *United Humanities Press.*

He, X., Witzel, C., Forder, L., Clifford, A. & Franklin, A. (2014). Color categories only affect post perceptual processes when same- and different category colors are equally discriminable. *Journal of the Optical Society of America A* Vol. 31, Iss. 4, pp. A322–A331

Holmes, A., Franklin, A., Clifford, A. & Davies, I. (2009). Neurophysiological evidence for categorical perception of color. *Brain and Cognition* 69, 426–434.

Ishihara S. (1917). *Tests for colour-blindness*. Handaya, Tokyo: Hongo Harukicho.

Jameson, K. A. & Komarova, N. L (2009a). Evolutionary models of color categorization. I. Population categorization systems based on normal and dichromate observers. *Journal of the Optical Society of America A* 26, 1414–1423.

Jameson, K. A. & Komarova, N. L (2009b). Evolutionary models of color categorization. II. Realistic observer models and population heterogeneity. *Journal of the Optical Society of America A* 26, 1424–1436.

Jordan, F., Dunn, M. & Majid A. (2009). Body Part naming booklet. Developed for the EoSS project. Nijmegen: Max Planck Institute for Psycholinguistics.

Kay, P. & Kempton, W. (1984). What is the Sapir – Whorf hypothesis? *American Journal of Physical Anthropology* 86, 65 – 79.

Kay, P. & Regier, T. (2003). Resolving the question of color naming universals. *Proceedings of the National Academy of Sciences*15, 100.

Kay, P. & Regier, T. (2006) Language, thought and color: recent developments. *Trends in Cognitive Sciences*. 10(2):51-4

Komarova, N.L., Jameson, K.A., Narens, L. (2007). Evolutionary models of color categorization based on discrimination. *Journal of Mathematical Psychology* 51 (6), 359–382.

Krauskopf, J. & Gegenfurtner, K. (1992). Color discrimination and adaptation. *Vision Research* 32(11):2165-75.

Kwok, V.,Niu, Z.,Kay, P., Zhou, K.,Mo, L.,Jin, Z., So, K., Tan, L.H. (2011). Learning new color names produces rapid increase in grey matter in the intact adult human cortex. *Proceedings of the National Academy of Sciences of the United States of America* 108, 6686–6688.

Lewis, D. (1969). *Convention: A Philosophical Study*. Cambridge, MA*: Harvard University Press.*

Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (eds.). (2014). Ethnologue: Languages of the World, Seventeenth edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Liberman, A.M., Harris, K.S., Hoffman, H.S., Griffith, B.C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology* 54, 358 – 368.

Liberman, A.M., Lane, H., Harris, K.S., Kinney, J.A. (1961). Discrimination of relative onset time of components of certain speech and nonspeech patterns. *Journal of Experimental Psychology* 61, 379.

Liu, Q., Li, H., Campos, J.L., Teeter, C., Tao, W., Zhang, Q., Sun, H.J. (2010). Language suppression effects on the categorical perception of color as evidenced through ERPs. *Biological Psychology* 85(1):45-52.

Loreto, V., Mukherjee, A., Tria, F. (2012). On the origin of the hierarchy of color names. *Proceedings of the National Academy of Sciences* 109: 6819–6824.

Majid A, Jordan F, Dunn M (2011) Evolution of semantic systems procedures manual. Nijmegen: Max Planck Institute for Psycholinguistics.

Majid, A. & Levinson, S. C. (2007). The language of vision I: Color. In A. Majid (ed.) Field Manual Volume 10, 32–35. Nijmegen: Max Planck Institute for Psycholinguistics.

Majid, A. (2008). Focal colors. In A. Majid (ed.) Field Manual Volume 11, 8–10. Nijmegen: Max Planck Institute for Psycholinguistics.

Majid, A. (2014). Comparing lexicons cross-linguistically. In Oxford Handbooks Online.

Majid, A., Jordan, F., & Dunn, M. (Eds.). (in press). Semantic systems in closely related languages [Special Issue]. *Language Sciences.*

Maris, E. & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods* 164(1):177-90.

Moreno, M. & Baggio, G. (2014). Role asymmetry and code transmission in signaling games: An experimental and computational investigation. *Cognitive Science*. 1551-6709.

Mufwene, S. (2002). Pidgin and creole languages. *International Encyclopedia of the Social and Behavioral Sciences,* 11440-11445.

Munsell, A.H. (1912). A Pigment Color System and Notation. *The American Journal of Psychology (University of Illinois Press)* 23 (2): 236–244.

Muysken, P. (1997). Media Lengua, in Thomason, S. G. Contact languages: a wider perspective *Amsterdam: John Benjamins*. 365-426

Nigam, A., Hoffman, J.E., & Simons, R.F. (1992). N400 to semantically anomalous pictures and words. *Journal of Cognitive Neuroscience* 4, 15-22.

Oldfield, R.C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, 9, 97-113

Oostenveld, R., Fries, P., Maris, E., Schoffelen, JM (2011) FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience* Volume 2011.

Özgen, E., & Davies, I. R. L. (2002). Acquisition of categorical color perception: A perceptual learning approach to the linguistic relativity hypothesis. *Journal of Experimental Psychology*: General, 131(4), 477–493.

Paramei G. V. & Menegaz G. (2013). 'Italian blues': A challenge to the universal inventory of basic color terms. In M. Rossi(Ed.), *Colour and Colorimetry: Multidisciplinary Contributions* (Vol. IX B, pp. 164-167). Rimini: Maggioli Editore.

Peters, J. (1845). Miscellaneous remarks upon the government, history, religions, literature, agriculture, arts, trades, manners, and customs of the Chinese: as suggested by an examination of the articles comprising the Chinese museum, in Marlboro' Chapel, *Boston, Boston: Eastburn's Press.*

Pinheiro, H., Seillier-Moiseiwitsch, F. & Sen, P.K. (1998) Analysis of Variance Based on Hamming Distances, International Biometric Conference, Cape town, South Africa.

Quine, W.V.O (1960). *Word and Object*. Cambridge, MA: MIT Press.

Roberson D, Hanley JR, Pak H. (2009). Thresholds for color discrimination in English and Korean speakers. *Cognition*. 2009;112(3):482–487.

Roberson, D. & Davidoff, J. (2000). The Categorical Perception of Colours and Facial Expressions: the Effect of Verbal Interference. *Memory and Cognition* 28.977-986.

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, 369–398 4

Roberson, D., Pak, H., Hanley, J.R. (2008). Categorical perception of color in the left and right visual field is verbally mediated: evidence from Korean. *Cognition* 107, 752 – 762.

Roberson, D., Davidoff J., Davies I.R., Shapiro L.R. (2005) Color categories: evidence for the cultural relativity hypothesis. *Cognitive Psychology*. 50, 378–411

Shayan, S., Ozturk, O., & Sicoli, M. A. (2011). The thickness of pitch: Crossmodal metaphors in Farsi, Turkish, and Zapotec. *The Senses and Society*, 6(1), 96–105.

Skyrms, B. (1996). Evolution of the Social Contract. *Cambridge University Press*. KJS

Skyrms, B. (2010). Signals. Evolution, Learning, & Information. *Oxford University Press*.

Tallon-Baudry, C., Bertrand,O., Delpuech, C., Pernier, J.(1996) Stimulus Specificity of Phase-Locked and Non-Phase-Locked 40 Hz Visual Responses in Human. *The Journal of Neuroscience*, 16(13): 4240-4249.

Tan, L. H., Chan, A. H., Kay, P., Khong, P-L., Yip, L. K. C., & Luke, K-K (2008). Language with perceptual decision. *Proceedings of the National Academy of Sciences*, 105, 4004-4009.

Terrill (2006) A. Body part terms in Lavukaleve, a Papuan language of the Solomon Islands. *Language Science*, 28:304–322.

Thierry, G., Athanasopoulos, P., Wiggett, A., Dering, B., & Kuipers, J. R. (2009). Unconscious effects of language-specific terminology on pre-attentive color perception. *Proceedings of the National Academy of Sciences*, 106(11), 4567–4570.

Waggoner, T. L. (2002). Quick six color vision test pseudoisochromatic plates. From Color vision testing by GoodILite Company.

Wang & Minett (2005). The invasion of language: Emergence, change and death. *Trends in Ecology and Evolution* 20, 263-269.

Whorf, B. L. (1936). "The punctual and segmentative aspects of verbs in Hopi". Language 12(2): 127-131.

Whorf, B. L. (1936). In E. C. Parsons (Ed.), Hopi journal of Alexander M. Stephen (Vol. 2, pp. 1198–1326). Columbia University contributions to anthropology (No. 23). *New York: Columbia University Press.*

Whorf, B.L. (1956). Discussion of Hopi linguistics. In J. B. Carroll (Ed.), Language, thought, and reality: Selected writings of Benjamin L. Whorf (pp. 102–111). *New York: John Wiley.*

Winawer, J., Witthoft, N., Frank, M.C., Wu,L., Wade, A.R., Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences of the United States of America* 104, 7780–7785.

Witzel, C. & Gegenfurtner, K.R. (2013) Categorical sensitivity to color differences. *Journal of Vision*. 13(7):1.

Xu, J., Dowman, M., Griffiths, T.L. (2013) Cultural transmission results in convergence towards color term universals. *Proceedings of the Royal Society B: Biological Sciences* 280.

Zollman K.J.S. (2005). Talking to Neighbors: The Evolution of Regional Meaning. Philosophy of Science 72: 69-85. A