MSc in Logic, ILLC, UvA

January 28, 2013

1 The World Color Survey and the facts to be explained

Berlin and Kay [2] pioneered a long tradition of research into color term systems of world's languages. This broad typological study pointed out that there clearly are universal tendencies concerning repertoires of basic color terms in particular languages. The essential claim for our purposes is an evolutionary one: there is a cross-linguistically universal order in which color vocabularies are enriched by new terms. The order of emergence is captured by the following implicational hierarchy of universals:

[white, black] < [red] < [green, yellow] < [blue] < [brown] < [purple, pink, orange, gray]

This hierarchy expresses a claimed complex constraint on all existing natural languages. It should be read as follows: if a language has a well-established term for red (that is, a term covering the point of the color space which is the prototypical denotation of the English "red"), it also has basic terms for white and for black, but not necessarily *vice versa*. Similarly, if it has one for green or for yellow, it also has one for red, and so on. According to this picture, color term systems are far from being relativistically arbitrary: whatever number of terms they contain, these terms can carve the color space only in certain ways. For example, in a language with three color terms, these are bound to cover, respectively, red, black, and white. Full partition of the color space is claimed, so the English terms are in fact misleading. In the last case, the term for red is likely to cover violet or orange as well and the other two terms would rather correspond to "dark" and "light".

In order to examine the main theses of the founding study, the World Color Survey (WCS, a comprehensive monograph is [7]) was conducted in the subsequent decades. It substantially broadened the empirical base and

improved the methodology of the previous work, performing field research for 110 unwritten languages (listed in [12]) with a negligible level of genetic interrelatedness, with 24 informants per language on average. (Cf. [8] for methodology.) The employed color system was one of 330 Munsell chips, 320 of them in the Lenneberg and Roberts array of 40 hue columns and 8 levels of lightness, at maximum saturation, plus an achromatic column of 10 chips from white to black.

The results of the WCS concerning the universality of color terms emergence can be found in Kay and Maffi [9]. Here, the empirically documented color term systems are classified into 9 types in 5 stages with respect to how six focal points of the color space (prototypical denotations of the English "white", "red", "yellow", "green", "blue", "black") are grouped by the vocabulary of each particular language. For instance, in Stage II (languages with 3 basic color terms) there is only one observed type, [white; red+yellow; black+green+blue]; in Stage III (4 color terms) there are the types [W; R; Y; Bk+G+Bu], [W, R+Y, G+Bu, Bk] and [W, R, Y+G+Bu; Bk]. The authors note that there are five possible evolutionary trajectories between stages I to V, assuming that any evolutionary step consists in splitting the denotation (the covered focal points) of one term of the previous system in two. The trajectory [W+R+Y; Bk+G+Bu] > [W, R+Y, Bk+G+Bu] > [W, R+Y, G+Bu, Bk] > [W; R; Y; G+Bu; Bk] > [W; R; Y; G; Bu; Bk] is the most represented one in the WCS, capturing 91 of the 110 languages.[1]

The way the empirical results are presented requires some discussion. The talk of evolutionary paths as instantiated by the languages is slightly misleading: what was really observed was in each case a static color term system belonging to one of the 9 types, or to a transition between two of them. These observed transitions are the maximum of diachrony captured in the WCS; apart from this we cannot infer anything about how a particular observed color vocabulary actually came about. This having been clarifed, the formulation in terms of these emergence trajectories and the representation of the particular types seems more appropriate than the original strong formulation in terms of implicational hierarchy. First, it draws attention to the almost universal[2] principle of partitioning the color space by the available terms, and consequently to the fact that if a color vocabulary is enriched with an additional term, the denotation of some or all of the already established terms is likely to be modified. Second, the gathered data do not seem

[1] As one type can be shared by more trajectories, the languages captured by all these trajectories do not add up to 110, or 100%.

[2] Rare exceptions are discussed in section 3 of [9].

univocal enough to formulate an implicational hierarchy as strong as that of Berlin and Kay [2], and anyway, in order to formulate any such generalization, the data would have to be statistically evaluated with this in mind. Admittedly, an overall quantitative evaluation of the WCS data has been done [10, 12] and it has shown a clear non-random match among color term systems of world's languages, thus refuting the position of full relativism in this respect. But whether *particular* generalizations (implicative or other) are valid is an altogether different question. We conjecture that a consequential part of the original hierarchy would not find a significant support in the data, since in the sample there are relatively little languages spread among the types with 4 or less color terms, as opposed to about 80 languages with 5 terms, 6 terms or inbetween.

In the general lack of statistically conclusive support for individual universal features of human color categorization, we will focus on the two of them that can be most reliably inferred from the absolute numbers reported in Kay and Maffi [9]. One is that 3-term vocabularies tend to partition the color space according to the scheme [W; R+Y; Bk+G+Bu], that is, to separate the warm colors while keeping the cool colors together with black. This is the only reported type for Stage II, instantiated by 6 languages. The other universal feature to focus on is the evolutionarily late division of green and blue: [W; R; Y, G+Bu; Bk] is by far the most represented type of 5-term systems (Stage IV), instantiated by 41 languages. The strong universal tendencies to carve up the color space in the described way when, respectively, three and five color terms are available will be in the following regarded as the most obvious, or secure, empirical facts to be explained. Besides qualitative evaluation based on this finding, we will evaluate our model in a more refined, quantitative way, against the detailed WCS data for particular languages [3].

2 Related work

The general debate on the nature of cognitive categories is dominated by three competing paradigms, nativism, empiricism and culturalism, the third often presented as a solution to the aged dilemma between the first two (cf. [16] and the references there). The debate on color categorization, specifically, is furthermore structured along the dimension of universality vs. relativity, which is arguably a distinct one, despite affinities such as that between nativism and universalism. The WCS has posed this question as straightforwardly empirical and provided data; as a result, recent posi-

tions on both the universalist [10, 12] and the relativist [14] side are rather moderate.

Granted that there *are* universal tendencies in color categorization, explanation of these (and of the remaining relativity) has been approached in several ways. Kay and Maffi [9] themselves present an updated version of a model that had been continuously developed by the WCS authors, on the (close to) nativist assumption of 6 naturally focal colors. The issue has been also studied within the broadly culturalist framework of the Iterated Learning Model [15, 4]. However, here we will only discuss in detail the approaches that directly motivate our own model, in which the emergence of categories is conceived in terms of cultural interaction on the basis of innate characteristics of human perception. In a nutshell, these are works [1] and [11], focusing on the impact of perceptual constraints on routinized cultural interaction; the more recent work [13] of the WCS authors investigating partitions of the perceptual space in terms of optimality; and Jäger and van Rooij's [5] proposal to treat the issue in game-theoretical terms. Moreover, in the final section on prospects we discuss one additional, empiricist approach, to motivate a possible extension of our model.

The first two of our motivating approaches jointly assume that universality of categorization might be explainable in terms of specific characteristics of human visual perception. We discuss, first, Baronchelli et al. [1] and Loreto et al. [11], who attempt to derive the universals from a particular formulation of the dependence of perception on the physical character of the input. Then we outline the explanatory strategy of Regier, Kay and Khetarpal [13], which appears to be a more general, though in a sense less elaborated, version of the former.

2.1 Just Noticeable Difference

[1], as well as [11], appeal in explanation to a simple characteristic of human visual perception, called Just Noticeable Difference (JND). The human JND is a psychophysiologically determined function which for any given wavelength from the extent of the visible spectrum gives the minimal difference in wavelength of two hues that are distinguishable by human eye in that particular region of the spectrum. This function is implemented as a constraint on cultural interaction of artificial agents, conceived roughly along the lines of Steels and Belpaeme's [16] "culturalist model". In this setting, color term vocabularies and categorical systems of individual agents in a population are made to co-evolve through their repetitive participation in standardized linguistic interaction over empirical input ("the Category

Game"). Similarity between the emergent systems and those observed empirically is then supposed to vindicate the explanatory role of the human JND.

Despite general formulations ("excellent quantitative agreement" with the WCS data), Baronchelli et al. are successful only in a specific sense. Their simulation does not demonstrate for particular universal features of human color categorization how these might have been arrived at. It shows only that categorical systems developed via cultural interaction constrained by the human JND are less dispersed across populations than when a flat, non-human JND is used. The only quantitative agreement, then, is between the ratio of the two respective values of dispersion, and the dispersion ratio of the actually observed categorical systems of the WCS, compared to a specific randomization of these (as in [10]). The agreement of these two ratios on $\sim$1.14 is remarkable, but hard to interpret in isolation.

Loreto et al. [11] come somewhat closer to explanation of particular universals of color categorization. The human JND as a constraint on routine language interaction over empirical input is sufficient for them to derive a hierarchy of color terms according to the time it takes for color terms in various regions of the visible spectrum to be agreed upon within the population. The announced "excellent quantitative agreement with the empirical observations of the WCS" is concealed from the reader. But the authors rightly point out that their hierarchy, [red, (magenta)-red] < [violet] < [green/yellow] < [blue] < [orange] < [cyan], is similar to the implicational hierarchy of Berlin and Kay [2]. Let us discuss the relevance of this finding.

First, there seems to be a methodological problem with choice of color terms and their matching to regions of the spectrum. This should, arguably, have been done either by selecting a set of cross-linguistically basic colors and locating them in the spectrum, or by selecting important points or sections of the JND function and reading off the respective colors; but an opaque combination of both seems to have taken place. In the first case we would expect both green and yellow in the selection, instead of green/yellow, and we might challenge the inclusion of cyan and (magenta)-red. In the second case, while most of the selected points reflect peaks and valleys of the function, (magenta)-red does not, violet and red are disputable, and there is an unreflected valley between violet and blue. Some of this could be actually resolved in favor of the parallel between the achieved hierarchy and Berlin and Kay's hierarchy; first of all, there are reasons to pick only red for the experiment, instead of red, (magenta)-red and violet. But there remains the problem that green/yellow in the achieved hierarchy is a single transitional color, while in Berlin and Kay's hierarchy green and yellow are

two distinct colors occupying the same position.

Moreover, let us remind that Berlin and Kay [2] is a dated reference and there is little point in evaluating explanatory proposals concerning universals of color categorization against the hierarchy stated there, in presence of the WCS data, the superiority of which is both empirical and methodological. Our conclusions in Section ... indicate that the mismatch between Loreto, Mukherjee and Tria's [11] actual findings and the cross-linguistic reality would have been magnified by an up-to-date evaluation, rather than attenuated. While we believe that the features of human perception that are captured by the JND function should play an important explanatory role regarding linguistic universals, the two papers just discussed do not more than indicate so.

2.2 The CIELAB space

The previous approach appeals to a particular feature of human perception (the resolution power in different frequencies of visible light). The explanatory strategy adopted by Regier, Kay and Khetarpal [13], with reference to [6], is a more general version of that. Instead of carving up a physically defined space (one-dimensional in the previous case), they consider partitions of the psychologically relevant, 3D color space CIELAB, which is designed so that standard Euclidean distance of two hues corresponds to their psychological dissimilarity. In this, the human JND is encompassed rather than discarded as a source of explanation, for what it expresses has to be involved also in construction of any psychologically relevant space. When the Munsell color palette used in the WCS is projected into the CIELAB space, its chips mark the surface of an irregular sphere there. What is then discussed are partitions of the set of color points thus arranged. The authors convincingly show a strong preference of the WCS languages for efficient partitions, efficient in terms of maximizing the compactness of their color categories in the CIELAB space. What this means is that the closer two chips are in the perceptual space, the more likely they will be lumped under the same color term.

This is clearly an important result, pointing to optimality as an essential factor of color categorization. However, this line can be drawn further. Given the specific way of evaluation (each language's actual partition vs. its various rotations around the sphere), the results cannot directly account for any particular linguistic universal in question. For instance, we do not see whether the most efficient ways of partitioning the figure into 5 regions involve keeping blue and green together. Another issue is that optimality or

efficiency is a static feature of a categorical system (of an individual speaker or of a language *in abstracto*), without it being clear how it might have come about. In our approach we adopt the idea of the overall character of human visual perception, reflected in the CIELAB color space, as the likely source of universals of color categorization. For sake of comparability with the WCS data we also work with the projected Munsell palette. Instead of static assessment of optimality, though, we will be interested in an evolutionary, agent-based dynamic of cultural interaction, in the game-theoretic formulation proposed in [5], over empirical input located in the defined space.

2.3 Similarity-maximization games

Jäger and van Rooij [5] construe the problem as a similarity-maximization signaling game. Nature picks a point from the color space as the meaning to be conveyed; the sender sends one term from a finite set to signal the chosen meaning to the receiver; the receiver interprets the received signal by choosing a point from the color space. The payoff this signalling action brings to both the sender and the receiver is a monotonically decreasing function of the distance of the receiver's interpretation from the intended meaning in the color space. In general, the sender strategy is a function from the set of points of the color space to (a probabilistic distribution over) the given set of terms, and the receiver strategy is a function from the set of terms to (a probabilistic distribution over) the set of points of the color space. If we let the game be played repetitively and relate payoffs from each particular game to the "fitness" of the sender and the receiver strategy employed in that game,[3] or the probability that they will be employed in the next run, we get an evolutionary process with a specific dynamic. This process can be, in principle, viewed as a model of evolution of color categories in a community. How various parameters of such a model are to be set up is, of course, subject to discussion.

We chose to base our evolutionary model in similarity-maximization signaling games, rather than in the Category Game of Steels and Belpaeme [16], adopted in Baronchelli et al. [1] and Loreto et al. [11]. In the former setting, categories are inherently linguistic and can be unproblematically called

[3] Jäger and van Rooij motivate this by appeal to priming effects between sufficiently close coding and decoding strategies. We found their reasoning unconvincing: reinforcement via priming can take place between the sender and the receiver strategy of a single individual, but not interindividually, since the sender in a game has no access to the strategy of the receiver and *vice versa*.

"concepts" as well. The latter approach, on the other hand, makes the conceptual distinction between perceptual and linguistic categories. Each agent, based on empirical input, individually divides a continuous perceptual space into regions (perceptual categories) within which she cannot further distinguish. A linguistic category then emerges through subsuming of adjacent perceptual categories under a single term. As perceptual categorization independent of language seems to be a problematic notion to us, we prefer the simpler formulation in terms of signaling games. Admittedly, we consider only the 320 (330?) Munsell chips as the color space, instead of the continuous space. This choice, motivated by its simplicity and good evaluability against the WCS data, can be seen as a preliminary perceptual categorization of the continuous space. However, a difference is that in our case the perceptual space is not carved up arbitrarily by individual agents, but uniformly and in roughly homogeneous way with respect to human resolution abilities.

References

[1] A. Baronchelli, T. Gong, A. Puglisi, and V. Loreto. Modeling the emergence of universality in color naming patterns. *Proceedings of the National Academy of Sciences*, 107:2403–2407, 2010.

[2] B. Berlin and Kay P. *Basic Color Terms*. University of California Press, Berkeley, California, 1969.

[3] R. Cook, P. Kay, and T. Regier. Wcs data archives. Available at http://www.icsi.berkeley.edu/wcs/data.html (January 2013).

[4] M. Dowman. Explaining color term typology with an evolutionary mode. *Cognitive Science*, 31:99–132, 2007.

[5] G. Jäger and R. van Rooij. Language structure: psychological and social constraints. *Synthese*, 159:99–130, 2007.

[6] K. A. Jameson and R. G. D'Andrade. It's not really red, green, yellow, blue: An inquiry into perceptual color space. In C. L. Hardin and L. Maffi, editors, *Color categories in thought and language*, pages 295–319. Cambridge University Press, 1997.

[7] P. Kay, B. Berlin, L. Maffi, W. R. Merrifield, and R. Cook. *The World Color Survey*. Center for the Study of Language and Information, Stanford, 2009.

[8] P. Kay and R. Cook. The world color survey (encyclopedia chapter). In R. Luo, editor, *Encyclopedia of Color Science and Technology*. Springer, in press.

[9] P. Kay and L. Maffi. Color appearance and the emergence and evolution of basic color lexicons. *American Anthropologist*, 101:743–760, 1999.

[10] P. Kay and T. Regier. Resolving the question of color naming universals. *Proceedings of the National Academy of Sciences*, 100:9085–9089, 2003.

[11] V. Loreto, A. Mukherjee, and F. Tria. On the origin of the hierarchy of color names. *Proceedings of the National Academy of Sciences*, 109:6819, 2012.

[12] T. Regier, P. Kay, and R. S. Cook. Focal colors are universal after all. *Proceedings of the National Academy of Sciences*, 102:8386–8391, 2005.

[13] T. Regier, P. Kay, and N. Khetarpal. Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, 104:1436–1441, 2007.

[14] D. Roberson, J. Davidoff, I. R. Davies, and L. R. Shapiro. Color categories: evidence for the cultural relativity hypothesis. *Cognitive psychology*, 50:378–411, 2005.

[15] K. Smith, S. Kirby, and H. Brighton. Iterated learning: A framework for the emergence of language. *Artificial Life*, 9:371–386, 2003.

[16] L. Steels and T. Belpaeme. Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, 28:469–529, 2005.