Towards more realistic modeling of linguistic color categorization

José Pedro Correia Radek Ocelák

Abstract

The ways in which languages have come to divide the visible spectrum with their color terminology, in both their variety and the apparent universal tendencies, are still largely unexplained. Building on recent work in modeling color perception and categorization, as well as the theory of signaling games, we incrementally construct a color categorization model which combines perceptual characteristics of individual agents, game-theoretic signaling interaction of these agents, and the probability of observing particular colors as an environmental constraint. We also propose a method of transparent evaluation against the data gathered in the World Color Survey.

The results show that the model's predictive power is comparable to the current state of the art. Additionally, we argue that the former is superior in terms of motivation of the principles involved, and that its explanatory relevance with respect to color categorization in languages is therefore higher. Our results suggest that the universal tendencies of color categorization cannot be explained solely in terms of the shape of the color space induced by our perceptual apparatus. We believe that only by taking the heterogeneity of the phenomenon seriously can we acquire a deeper understanding of why color categorization takes the forms we observe across languages.

Keywords: color categorization, color naming, World Color Survey, signaling games

1 Introduction

The decades following the landmark work by Berlin and Kay (1969) up to now have seen an immense research effort concerning the issue of color categorization, or naming, in the languages of the world. Are there universal patterns, or at least remarkable cross-linguistic tendencies, in how different languages categorize the spectrum of visible color by their "basic" color terminology? If so, how can we explain the existence of such patterns? For a long time, the discussion had been polarized between two opposing camps, the universalists (in particular, Paul Kay and colleagues) and the relativists (such as Barbara Saunders and John Lyons), who suggested different answers to both these questions.

The universalists defend the existence of variously strong universal patterns of color naming, from the firm implicational hierarchy of color terms as put forward by Berlin and Kay (1969), progressively mitigated later (Kay 1975; Kay and Maffi 1999), to weaker statements of non-randomly strong cross-linguistic patterns (see Kay and Regier 2003; Regier, Kay, and Cook 2005). The more recent of these claims are based on the results of the World Color Survey (Kay et al. 2009; Kay and Cook 2016), whereby color naming data for 110 unwritten languages of 45 language families have been gathered. The explanations proposed by the universalists (see Kay and McDaniel 1978; Kay and Maffi 1999; Kay et al. 2009) have been mainly in terms of the neurophysiology and psychophysics of color perception in human individuals, with the putative privileged status of black, white, red, green, yellow, and blue as the Hering primaries or the Fundamental Neural Response categories. For criticism, see Saunders and van Brakel (1997a), Jameson and D'Andrade (1997), Jameson (2010), Ocelák (2015), and Witzel (2018).

The relativists, on the other hand, highlight the differences and peculiarities of color naming in particular languages and criticize the methodology behind the universalist findings, including the procedures of the World Color Survey (Lyons 1995; Lucy 1997; Saunders and van Brakel 1997a; Saunders 2000, 2007). There is also a related but distinct position of relativism with respect to color categories as a perceptual phenomenon, rather than as a sole matter of color naming (Roberson, Davies, and Davidoff 2000; Roberson et al. 2005). As opposed to the perceptuo-biological constraints on color categorization, relativists of the

former sort emphasize that the evolution of a color naming system is essentially a socio-cultural process, and they note that if there actually are strong cross-linguistic patterns in color categorization, this may be caused partly by the history of language contact, notably by the influence of the Western colonial languages (see Saunders and van Brakel 1997a,b).

What is clear from the empirical data is that there are cross-linguistic statistical regularities (Jäger 2010), but there is also a lot of variation. Intra-linguistic individual fluctuation is patently clear (more on this in Section 2.3). And even in their strongest claims of universality, Berlin and Kay (1969) admit room for variation within some levels of their hierarchy[1]. The two camps focus each on the aspect of the data that supports their explanatory hypothesis. Around the turn of the century, many adopted the idea that it would take both perceptuo-biological constraints *and* cultural processes to explain the data (Dedrick 1998, 2006, Among others,)). Given the added difficulty of considering the effect and interaction of various constraints, this perspective has motivated the use of mathematical and computational models as exploratory tools.

One well known paradigm was introduced in Steels and Belpaeme's case study (Steels and Belpaeme 2005) on the evolution of shared color categories. They experiment with agent-based models that aim to capture adaptive processes of category formation under pressure for successful communication, incorporating nativist, empiricist, and culturalist assumptions in various degrees and exploring their implications within the models. Other agent-based models explore variations on the theme. Dowman (2007) drops the distinction between linguistic terms (shared) and concepts (internal to the agents) and uses Bayesian inference. Komarova, Jameson, and Narens (2007) (and more recently Park et al. 2018) consider various learning mechanisms applied to a game of categorization based on the notion of a similarity range. Baronchelli et al. (2010) and Loreto, Mukherjee, and Tria (2012) try to reduce the innate constraints of the agents in a similar model, and argue that using a real feature of the human perceptual system (the so-called Just Noticeable Difference) is sufficient to generate results that exhibit some statistical properties that are close to the empirical data of the WCS. All of these models[2]introduce perceptual constraints upon individual agents, and make use of principles of dynamic linguistic interaction within communities of such agents, thus to some extent incorporating both the universalist and the relativist insights.

A similar paradigm is the framework of *signaling games*, a game-theoretic approach to modeling communication as anchored in convention, introduced by Lewis (1969) and later revived by Skyrms (1996) and other authors[3]. One big difference with the agent-based approach is that signaling game models can be studied at a higher level of abstraction via the use of population dynamics. These are abstract equations which purport to capture, at the population level, processes of biological or cultural evolution, but have also been shown to adequately represent the dynamics of some learning mechanisms. A model that investigates the issue of color categorization along these lines was introduced by Jäger and van Rooij (2007), and this is the paradigm followed in the implementation of the models presented here.

A later contribution in the universalist line by Regier, Kay, and Khetarpal (2007) dropped the problematic assumption of the privileged status of particular colors in human color perception, and suggested that the universal patterns can be explained in terms of optimal, efficient partitions of the perceptual space of color as such. One of the methods used to support their explanatory claim is to derive those using an optimization procedure that maximizes a well-formedness measure based on perceptual constraints. These are in turn induced by the location of the color represented by each Munsell chip in the CIELAB color space, which is supposed to capture the relations of similarity and difference between colors as perceived by a standard (that is, normal trichromatic) human observer. The optimization procedure is, however, admittedly artificial and does not purport to model the categorical formation process in a realistic way. Therefore, their perspective is compatible with the seemingly more realistic processes of linguistic interaction which lead to the emergence of categories in the models mentioned above. The connection between these two perspectives is more explicitly elaborated by Regier, Kemp, and Kay (2015).

From both these branches of color categorization modeling, we adopt the following implicit assumption: if a fairly realistic model can produce categorical schemes that fit well with the empirical color naming systems, then such a model can claim explanatory relevance with respect to the cross-linguistic patterns. An explanation of both regularities and variance in color categorization is not an aim in itself. It would, in

[1]Stage III languages can differentiate either green or yellow, and Stage VII languages optionally carve out purple, pink, orange, or gray.

[2]See Belpaeme (2014) and Kallens, Dale, and Smaldino (2018) for more detailed overviews of the literature.

[3]See Skyrms (2010), Huttegger (2014), and Franke and Wagner (2014) for overviews.

turn, constitute an important contribution to the more general debate concerning the origins of concepts or categories in human cognition, the aged "nature vs. nurture" debate[4].

However, the explanatory claims of the existing models (in so far as they directly raise such claims) appear inconclusive to us. In each case, we feel that serious improvements can be made as regards to one or both of the two crucial desiderata of any model that is to be explanatorily relevant: on the one hand, realism of the principles involved; on the other, a transparent evaluation procedure allowing for the decision whether (or to what extent) the categorical systems produced by the model in question resemble the color naming systems actually observed in human languages.

In the present paper, we put forward a new color categorization model, in order to examine several explanatory ideas concerning the patterns of color categorization in the languages of the world. Our model, in its successive variants, preserves the general structure of the models referred to above. That is, we employ a dynamic, game-theoretic interaction above perceptual constraints motivated on the level of individual agents. We nonetheless believe we advance beyond the existing work, both in the respect of the model's motivation and its evaluation. In the final stage, we also enhance the model with an additional, environmental constraint, namely with quasi-realistic probabilities based on the frequency of occurrence of each color in a natural environment. On a more general level, we hope in this paper to establish a transparent evaluation methodology that is so far missing in color categorization modeling, and which will sharpen the very notion of a model's predictive success. We believe the existing modeling approaches can and should provide results that are directly comparable to empirical data.

The paper is further organized as follows. In Section 2 we define our evaluation methodology and discuss the World Color Survey data that is used as its basis. In Section 3 we formulate a color categorization model based on similarity-maximization signaling games. Our objective is to suggest a more naturalistic alternative to the categorical optimization used by Regier et al. In Section 4 we introduce two incremental variants to our model, which are meant to improve upon the adequacy of its perceptual component. In Section 5 we discuss the explanatory significance of our model (in its subsequent variants) with respect to the empirical patterns of color naming, and we outline the room for future improvements. We make our concluding remarks in Section 6.

2 Evaluating color categorization models

As regards the original opposition between the universalists and the relativists, it needs to be noted that the tension is only partly mitigated by the fact that the more recent models tend to combine perceptual constraints with interactional ones. Another point of confrontation is, what are the empirical data on color naming which we should use in evaluating a model's ability to produce realistic categorical systems? The standard option (Regier, Kay, and Khetarpal 2007; Baronchelli et al. 2010; Loreto, Mukherjee, and Tria 2012) is to confront the model with the data provided by the World Color Survey (Kay et al. 2009; *WCS Data Archives*). Yet, to remind, the relativists tend to question the relevance of these very data with respect to the problem at hand, suggesting that the universal patterns therein are at least partly caused, first, by the influence of the Western colonial languages, and second, by the methodology of the WCS, which is itself claimed to be universalist in nature. In our evaluation procedure, we attempt to assume a considered position between universalism and relativism. Namely, we make use of the data gathered in the WCS, notably and indisputably the most extensive cross-linguistic research ever performed on color naming; but instead of taking the data at face value, we clean and reduce them in several ways, so as to neutralize the most serious objections raised by the relativists.

We start this section by discussing the World Color Survey data and how we handle them for our evaluation purposes. Further, we describe the methods employed to compute the similarity of color categorization schemes, which will in turn help us quantify the predictive performance of a color categorization model. In order to illustrate the working of the proposed methods, we perform a quantitative reassessment of the model by Regier, Kay, and Khetarpal (2007). Finally, we define a clear methodology to evaluate the results of the models proposed later in this study.

[4]For discussion of the conceivable positions within the dispute, see Steels and Belpaeme (2005).

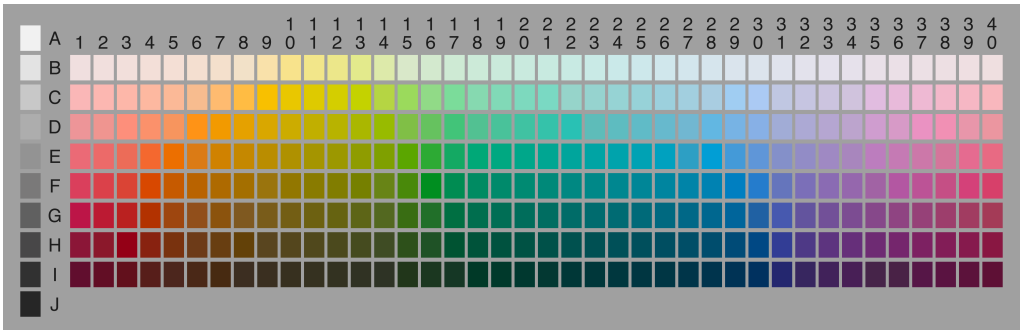


Figure 1: The array of 330 Munsell color chips used in the World Color Survey. Reproduced from Cook, Kay, and Regier (*WCS Data Archives*) with authors' permission. The figure is only for the sake of illustration; faithfulness to the standardized physical set of Munsell chips is generally not guaranteed with screening or printing devices.

2.1 The World Color Survey data

The World Color Survey (Kay et al. 2009; Kay and Cook 2016; *WCS Data Archives*) investigated color naming in 110 unwritten languages of 45 language families. On average there were 24 informants consulted per language and the modal number was 25, although for some languages the number of informants was as low as 6. Each informant was presented 330 Munsell color chips in a fixed random order, and was asked to provide a short name in response to each. The 330 individually presented color chips belong to an array (hereafter, "the Munsell color array", Figure 1) composed of 40 chromatic hues (columns) at 8 levels (rows) of lightness[5], plus a column of 10 achromatic chips from white to black. The Munsell color array, serving as a background, thus provides for convenient representation of the color categories in particular languages.

The relativist objections concerning the relevance of the color naming data provided by the World Color Survey are discussed in detail by Ocelák (2013, ch. 4), in part approvingly. The discussion results in several suggestions for a reduction of the body of the WCS data. It is meant to isolate a part of the data that can safely stand the relativist criticism and be considered representative of color naming in human languages, for the purpose of evaluating color categorization models. Accordingly, our evaluation procedure makes use of the data available in the WCS Data Archives; yet diverging from the previous evaluation practice (Regier, Kay, and Khetarpal 2007; Baronchelli et al. 2010; Loreto, Mukherjee, and Tria 2012), we reduce the data along two dimensions, prior to performing evaluation.

2.2 Language exclusions

Of the 110 languages covered by the WCS, we exclude the data on 29 languages, listed in the Table 1, because of their arguable lack of representativeness with respect to the empirical phenomenon at hand. More specifically, we used the following two criteria:

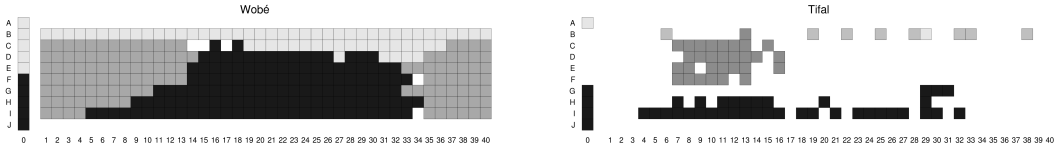
1. We excluded 26 languages with a non-negligible presence of (detectable) loanwords in their color terminology; more specifically, a language was excluded if it was a creole, contained at least one loanword used by more than half of the respondents, or showed evidence of a pressure of a different language on its color system (all according to the data and characteristics described by Kay et al. 2009). This is meant to do some justice to the relativist worry that the universal patterns in the WCS data may reflect the world-wide impact of the Western colonial languages; that is, an accidental historical fact for which a categorization model cannot be expected to account.

2. We excluded 3 other languages based on apparent or suspected failures in the application of the standard WCS procedure in their case, and the following doubts regarding the reliability of the collected data. For Karajá, the data were elicited in a group setting, and are therefore probably significantly

[5]Each chip is on the maximal level of saturation that is available, for that particular hue-lightness combination, in the Munsell color order system; cf. Fairchild 2005

Reason for exclusion	Languages
Spanish loanwords	Agta, Aguacatec, Amuzgo, Cakchiquel, Camsa, Chinantec, Chiquitano, Garífuna, Huastec, Huave, Mazahua, Mazatec, Nahuatl, Tarahumara (Central), Tarahumara (Western)
Other loanwords	Agta (Tagalog and English), Guaymí (probable English loanword), Gunu (probable loanword of unknown origin), Halbi (Hindi and probably Oryian), Yupik (English loanword)
Other linguistic interactions	Chavacano (Spanish creole), Djuka (Dutch creole), Kriol (English creole), Mixtec (pressure from Spanish), Saramaccan (English-Portuguese creole)
Potential methodological issues	Gunu ("local assistants of undocumented training"), Cree (suspicious outliers in the 50% agreement array), Didinga (suspicious outliers in the 50% agreement array), Karajá (group elicitation)

Table 1: Excluded languages



(a) **(b)**

Figure 2: Two examples of majority maps. Plots are created using the 'ggplot2' R package (Wickham 2009). Coloring is done automatically and in grey scale to avoid inadvertent connotations of native color terms to English ones.

interdependent between individuals. The data on Cree and Didinga show several strong outliers surrounded by a different category in the 50% agreement array (see below); we consider it unlikely that a majority of respondents would agree upon naming of a single chip in the surround of another category if the standard WCS procedure had been followed, so we exclude these two languages as well.

We keep the remaining 81 languages.

2.3 Per-language aggregation

The available WCS data make it possible to retrieve, for each of the 330 chips of the Munsell color array, how that particular color was named by each individual speaker of each WCS language. Regier, Kay, and Khetarpal (2007), in their evaluation, represent the categorical system of a WCS language by what they call a "mode map". In a mode map, each of the 330 chips is assigned to the category corresponding to the term which was used for naming of that chip by the highest number of speakers. It follows that each language is presented by its mode map as completely carving the Munsell array with its color categories. In contrast to Regier, Kay, and Khetarpal (2007), but in accordance with the "higher agreement arrays" in Kay et al. (2009), we do not assign a chip to a category unless a substantial part of the informants used the corresponding term in naming it. Consequently, we represent the categorical systems of languages with higher agreement arrays which typically contain "gaps", that is, regions of chips that are not assigned to any category. Since we set the necessary level of agreement at strictly more than 50%, in the following we will call these higher agreement arrays "majority maps". Figure 2 shows two examples of a majority map, namely for the languages with the lowest (5) and highest (250) number of missing data points. The median number of missing data points in our reduced subset of the WCS is 78, or approximately 23.6%. Majority maps for all languages in the WCS are available online (Correia 2015h).

This step reacts to another relativist objection that seems pertinent to us, namely that the method of mode maps, by definition, presents any language as standardly categorizing color, regardless of the fact that some languages apparently do not have abstract color terminology at all, or have established color categories only in *some* regions of the color space. Our method of representation conservatively assumes that a color can be assigned to a linguistic category only when there is a sufficient agreement[6] among speakers upon the naming of that color.

Note that, knowing the categorization made by each individual speaker of the World Color Survey, we could drop the abstract level of languages altogether and directly use the individual categorizations for our evaluative purposes. That would make the aggregation step described in this section unnecessary. However, the little theoretical gain would be more than balanced by the loss of comparability to other literature on color naming, and this approach would not spare us the problem of missing data (as discussed below). Moreover, the traditional notion of language has been involved in the very choice of WCS informants. We thus prefer keeping this notion in our analysis.

2.4 Quantitative measurement of color-naming scheme similarity

The problem presented to each subject going through the WCS data elicitation process can be seen as a simple categorization task: given a number of data points (s)he is to indicate which category each point belongs to, with data points being the 330 Munsell chips and categories being the salient color terms in their language. This is the same task that a (computational) model of color categorization should be able to perform in order to enable evaluation against the WCS empirical data. If a model generates an output in the same format, it is thus possible to evaluate it against a majority map by a measure of similarity.

Computing cluster similarity Although we talk about a *categorization* as the assignment of each Munsell chip to a color category, formally the data is equivalent to what would be called a *clustering* in the context of cluster analysis. Therefore, we searched the literature looking for a measure that we could employ to compare two categorizations. Vinh, Epps, and Bailey (2010) make an in-depth analysis of measures for the comparison of clusterings and suggest the use of a variant of the Normalized Mutual Information (NMI). They also stress the importance of correcting for chance "when the number of data items is relatively small compared to the number of clusters" (Vinh, Epps, and Bailey 2010, p. 2847) and propose an adjusted form of the measure. Given that the number of data items in our case is 330 (Munsell chips) and the number of clusters is often over 3 (color terms), based on their indicative values (Vinh, Epps, and Bailey 2010, p. 2846) we should use adjustment for chance.

Unfortunately, we were unable to find an efficient implementation of the adjusted NMI, and a naive implementation had prohibitive computational costs. Therefore, we resorted to using the Adjusted Rand Index (Rand 1971) (ARI) instead, which although not possessing all of the nice theoretical properties of the adjusted NMI, still works quite well and is a well-known and widely used measure of cluster similarity. The ARI takes into account both the amount of agreement and of disagreement between two clusterings (ignoring category labels) to compute a number which is upper bounded by 1 and has an expected value of 0 for two random independent categorizations. It is, however, not lower bounded by 0 and can take negative values under certain conditions. An advantage of this measure is that it can be used even when the number of categories between two clusterings is different, allowing us to compare any two color categorization schemes. All calculations were performed in R (R Core Team 2014) using the implementation of the ARI in the 'mclust' package (Fraley, Raftery, and Scrucca 2014).

Dealing with missing data Given our decision to represent the categorical system of each language with a majority map, we are faced with an issue in the direct application of this measure of similarity. Namely, the ARI is defined assuming that each categorization is *complete*, *i.e.* that every data point is assigned to one category. However, when data points do not achieve a certain level of agreement, we represent them as missing data, thus not assigning them to any category. Since the ARI is computed based only on points that are well-defined in both clusterings, we can obtain somewhat misleading results with this measure. For

[6] Our setting of the necessary agreement at strictly more than 50% is arbitrary, apart from the fact that it is the lowest value that avoids the risk of having to assign a chip to one of more equally suitable categories.

example, the two languages shown in Figure 2 have an ARI of 0.79. This is a very high number, given that we assign no category to approximately 75% of the data points for Tifal.

Our solution to this problem is to perform a form of *data imputation*. Namely, the similarity between two categorizations is calculated by first generating a certain number of complete cases for each categorization, where data points with no assigned category are attributed one at random (uniformly selected from the existing categories), then calculating the ARI between each complete case for the first categorization and each complete case for the second categorization, and finally taking the median of those values. This procedure allows us to quantify the overall similarity more reliably since the intermediary ARIs are only calculated for complete categorizations. The similarity value thus obtained, henceforth called **S**, becomes 0.05 between Wobé and Tifal, a much lower value than before.

Interpreting results The approach has certain implications that one should be aware of when interpreting results. As was mentioned above, the ARI is designed to have an expected value of 0 when applied to random independent categorizations. Because categories are assigned to data points at random during the imputation process, **S** will be more biased towards 0 the more missing data points a language has. Going back to the examples in Figure 2, whilst the median similarity between Wobé and the rest of the languages in the corpus drops from 0.58 to 0.32 when introducing imputation, for Tifal the drop is from 0.66 to 0.03.

We accept these consequences, since it reflects a conservative approach. Using this definition of similarity, a language with many missing data points is bound to have low similarity against any other categorization system. If we insisted on the possible intuition that Tifal instantiates a categorization system similar to that of Wobé, notwithstanding the imperfect data, that would clearly invite the objection of undue universalism in assumptions. However, when interpreting a certain value of **S**, it can be useful to also look at the value of the ARI. Since the latter compares only defined data points, it can help one understand to what extent the value of **S** is influenced by the amount of missing data or a mismatch between known values. This will become clearer when we apply the metrics in practice in the following section.

2.5 Revisiting the results of Regier et al.

In order to demonstrate the application of our quantitative evaluation method, we revisit part of the results of Regier, Kay, and Khetarpal (2007), namely what they call theoretically optimal color-naming schemes. These schemes are optimal, or maximally well-formed, in that they maximize the spatial compactness of categories, as measured in the perceptual color space CIELAB. In other words, they minimize the average perceptual distance of two chips within a category, and maximize the average perceptual distance of two chips from different categories. We will henceforth refer to these results as RKK-*n*, with *n* corresponding to the number of color categories. Therefore, RKK-3 is their theoretically optimal color-naming scheme for 3 color categories, and so forth. These results are plotted in Figure 3.

We calculated our similarity measure **S** (using 100 complete cases per language for the imputation process) and the pure ARI between each result and each language in the WCS, as well as between each pair of languages in the WCS. Equipped with these quantitative results, we can revisit the first prediction of Regier, Kay, and Khetarpal (2007, p. 1438), namely that "[a]rtificially generated color-naming schemes that lie at global well-formedness maxima should resemble the natural color-naming schemes found in some of the world's languages." In Table 2 we present, per RKK result, similarity values for both the three WCS languages which rank highest in terms of **S** and the examples provided by Regier et al. that are not in that group[7]. We emphasize in italics all the examples mentioned by Regier et al. In Table 3 we present the most similar WCS language according to our metric **S** for each of the WCS languages mentioned in Table 2. The full data is available online (Correia 2015g).

First, we should observe that it is not always the case that the examples presented by Regier et al. as WCS languages that resemble their simulations are the best matches according to our measure of similarity. One factor that certainly impacts this has to do with our use of majority maps with missing data, rather than

[7]We present the numbers for the additional languages given as examples of good matches by Regier et al. for a matter of completeness. For RKK-6 we do not show numbers for Aguacateco since we chose to exclude this language from the data set, for reasons discussed in Section 2.2. Also, note that we compare each optimal color-naming scheme to all languages in the WCS, not only those with the same number of color terms. The reason is that there are also "residual" terms covering only a few of the 330 color points, which makes the groups of n-term languages not quite distinct.

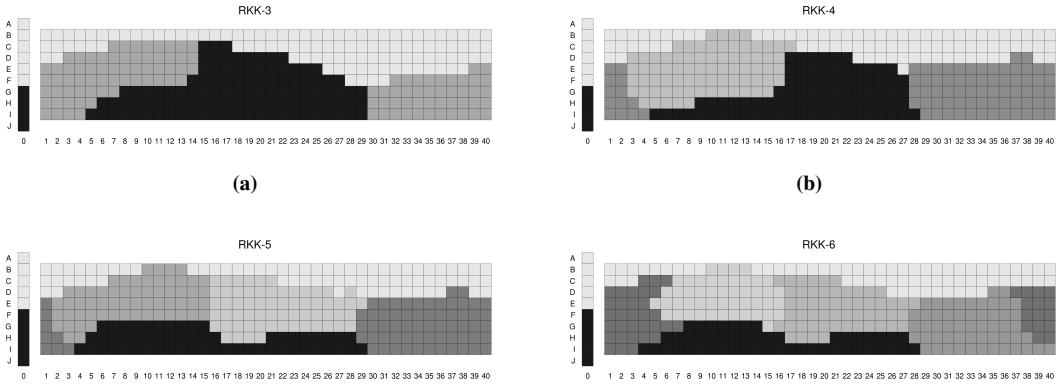


Figure 3: The theoretically optimal color-naming schemes as calculated by Regier et al.

	RKK-3			RKK-4			RKK-5			RKK-6		
	Language	**S**	ARI	Language	**S**	ARI	Language	**S**	ARI	Language	**S**	ARI
Top 3	*Wobé*	**0.48**	0.49	Wobé	**0.37**	0.38	Bauzi	**0.43**	0.50	Bauzi	**0.45**	0.51
	Ejagam	**0.39**	0.41	Colorado	**0.34**	0.45	*Colorado*	**0.40**	0.54	Colorado	**0.43**	0.56
	Bauzi	**0.38**	0.44	Bauzi	**0.33**	0.37	Múra Pirahá	**0.40**	0.44	Ocaina	**0.38**	0.46
Additional	*Bété*	**0.35**	0.48	*Culina*	**0.32**	0.50	*Iduna*	**0.32**	0.51	*Cofán*	**0.38**	0.47
							Cayapa	**0.31**	0.54	*Buglere*	**0.37**	0.55

Table 2: Similarities between RKK results and some languages in the WCS.

	Language	**S**	ARI
Wobé	Ejagam	**0.71**	0.77
Ejagam	Nafaanra	**0.79**	0.86
Bauzi	Múra Pirahá	**0.63**	0.78
Bété	Ejagam	**0.66**	0.92
Colorado	Bauzi	**0.60**	0.85
Culina	Vagla	**0.41**	0.74
Múra Pirahá	Bauzi	**0.63**	0.78
Iduna	Colorado	**0.51**	0.92
Cayapa	Cofán	**0.53**	0.92
Ocaina	Cofán	**0.64**	0.89
Cofán	Ocaina	**0.65**	0.89
Buglere	Zapotec	**0.53**	0.83

Table 3: Most similar languages in the WCS corpus for languages figuring in Table 2.

mode maps. This can be corroborated by comparing the values of **S** with those of the ARI. Most languages in the additional set have an ARI higher than some languages in our top 3. But then again, so does a language like Tifal with ARI values against RKK-3, RKK-4, RKK-5, and RKK-6 of respectively 0.64, 0.77, 0.79, and 0.81. Given the lack of agreement among its speakers (see Figure 2 again), it would be rather bold to claim a good match with any RKK result. We can only be aware of such situations, we believe, by both taking into account the problem of intra-linguistic variation as we do with the missing data, and analyzing the problem equipped with an quantitative measure of similarity, rather than relying on visual comparisons.

In order to further put the numbers in perspective, we can compare them with numbers from real languages in the WCS. We may think of this as a pragmatic baseline for a theoretical color-naming scheme: if we conceive the scheme as a hypothetical language on its own, we can interpret its similarity to a given language in comparison with the similarity of other real languages to that language. This is of course dependent on how languages in the WCS compare to each other (we can imagine a situation of a loner language, totally different from all the others), but the point here is simply to contextualize the numbers for a simulation by providing further examples.

For RKK-3, 2 of the 3 best matching languages, namely Wobé and Ejagam, were presented by Regier, Kay, and Khetarpal (2007, p. 1439) as examples that visually appear to match their results well. The similarity of 0.48 between RKK-3 and Wobé might appear considerably good and lead us to the conclusion that RKK-3 adequately captures the solution achieved by this empirical language. However, we can note that, for example, the similarity between Wobé and Ejagam, which is the best WCS match for Wobé, is 0.71. Despite the apparent similarities, the difference between 0.48 and 0.71 gives the indication that there are also significant differences between RKK-3 and Wobé which are much less prominent between, say, Wobé and Ejagam. On visual inspection one would say that this has to do with the boundary of the term used in the "red" region being moved further into the "blue" region in the simulation results (an observation also made by Regier, Kay, and Khetarpal (2007, p. 1438)), and with the "white" category being extended into the darker regions of the spectrum.

RKK-4 does not seem to be a very good representative of any WCS language in our reduced set, given the relatively low values of even the top three most similar languages and the significant differences between it and each of those languages. The best match is with Wobé again, at 0.37, substantially lower than the similarity value achieved by RKK-3 and much lower than the value between Wobé and Ejagam at 0.71. Culina, the only language presented by Regier, Kay, and Khetarpal (2007, p. 1439) as an example of a good match, ranks as the 4th most similar language at 0.32. If we look at the ARI value for this language, we understand why Regier et al. present it as an example of a good match. Visually, it does appear closer to RKK-4 than Wobé in the tendency to center a separate category in yellow. Using our proposed metric **S**, this apparent similarity is attenuated because of the lower agreement among the speakers of Culina.

RKK-5 and RKK-6 have similarity values for their three most similar languages which are comparable to RKK-3 in absolute terms. Top two positions are occupied by the same two languages in both cases, Bauzi and Colorado. Both Bauzi's and Colorado's highest similarity (respectively, 0.63 against Múra Pirahá, and 0.60 against Bauzi) are significantly lower than the highest similarity for Wobé; the absolute values nonetheless indicate some significant differences. Visual inspection shows that RKK-5 overestimates (against Bauzi and Colorado) the size of the color categories centered in the "red/violet" and in the "yellow" region, and that RKK-6 does not quite capture the empirical solutions particularly in the "green" and "blue" regions.

Altogether, it may seem surprising that one language, Bauzi, can figure as one of 3 most similar WCS languages for *each* of RKK-3 to RKK-6, and that another, Colorado, achieves the same for three of these optimal schemes. The possibility of such a result follows from our method of measuring similarity. Bauzi and Colorado are both 5-term languages. Even if, by the same token, their similarity to a language with, say, 4 or 6 terms cannot be perfect, they and such a language can still be mutually fairly informative as to how they split the color spectrum (at least in quantitative terms).

In conclusion, with quantitative data we can say with more confidence that the simulation results of Regier, Kay, and Khetarpal (2007) do indeed approximate some of the natural color-naming schemes, at least to a degree. However, we also observe that the approximation is still not at the same level as seen

between some naturally occurring color-naming schemes.[8]

2.6 Evaluating models

The exploratory analysis in the previous section serves to demonstrate that, even when equipped with a quantitative measure of similarity, the assessment of a color categorization model based on the color-naming schemes it produces is a non-trivial task. Our particular stance towards the models that we introduce in this paper is that each simulation result should be seen as a color-naming scheme in its own right, albeit an artificial one. Each output is seen as a valid potential reflection of the assumptions put into the models. We produce many simulations per configuration, and consider all of them relevant. That further complicates the analysis.

The question can be phrased as follows. Given a model that produces artificial categorical partitions of the Munsell color array, and a method of determining the similarity between two arbitrary partitions (artificial or empirical, total or partial), how do we confront the set of partitions provided by the model with the empirically observed color-naming schemes, so as to assess the model's predictive power?

We do not expect the result of a particular simulation to be a good prediction for the empirical categorical systems in general, as these can differ wildly from each other. Rather, a good simulation result is one that matches well with *some* of the existing languages, similarity with the others notwithstanding. For each simulation, we therefore focus on the *maximal* similarity achieved when compared with the 81 empirical languages in consideration.

Going beyond the level of assessing a particular simulation result, one could conceive of a model's overall predictive success in terms of precision and recall. Achieving maximal precision would mean that each categorical system generated by the model would resemble an existing language. Maximal recall would be achieved when each of the existing languages would be reflected in a categorical system generated by the model. In practice, however, achieving both good precision and recall goes far beyond our expectations given the current state of the art. In the following, the particular stages of our model will be assessed primarily in terms of precision. That is, we ask how many of our simulations find a sufficiently similar counterpart among the WCS languages.

It needs to be emphasized that this is a method to compare the relative predictive power of different models, rather than a method of quantifying how good a model is in absolute terms. All our quantifications of the efficiency of a particular model are strictly relative to our sample of 81 WCS languages. Besides the obvious issue of how representative the sample is, note that if our particular sample were enlarged, the same simulations could be expected to reach higher maximal similarities on average.

Now that we have established a methodology to quantitatively assess the accuracy of color categorization models, let us bring forward our first proposal to increase realism in this kind of modeling, and evaluate its results accordingly.

3 On the evolution of color categories

Regier, Kay, and Khetarpal (2007) obtain their theoretically optimal color naming schemes by taking the best of 20 runs of an algorithm whereby categories are formed through steepest ascent in well-formedness, starting from an initial random categorization. The authors do not claim this process to model the actual ways of the evolution of color categories in human languages, in our opinion rightly so. Namely, the procedure explicitly operates with the distance of various colors in the perceptual color space, that distance being a rather abstract characteristic. An explicit knowledge of it cannot be assumed on the part of the language users, through whose communicative behavior language evolution proceeds. In our first proposal towards more realistic modeling of color categorization, we keep the perceptual setting of the model of Regier et al., namely by using the 330 color chips of the Munsell array represented in the perceptual color space CIELAB as the percepts to be categorized. However, we implement a more naturalistic process

[8] Regier, Kemp, and Kay repeat the experiment with a slightly adjusted procedure and obtain 4 new optimal color naming schemes which differ mildly from those discussed here. Given the illustrative character of this section, we do not inspect the results in similar detail here, but give an analogical table of achieved similarities in Appendix A.1. In general, these new optimal schemes perform better for 5 terms, worse for 3 terms, and comparably for 4 and 6 terms.

of forming categories, one that we believe better reflects how particular color naming systems of human languages may have emerged.

3.1 Proposal 1: Evolution of categories as dynamics of signaling

We formalize this process in game-theoretic terms using similarity-maximization signaling games (henceforth *sim-max games* for short), in the spirit of Jäger and van Rooij (2007). The idea is as follows:

1. "Nature" picks a point from the color space;
2. Based on this point, a sender chooses one message from a finite set and communicates it to a receiver;
3. The receiver then picks a point from the color space based on the signal received;
4. This signaling interaction brings a payoff—which is a monotonically decreasing function of the distance in the color space between the receiver's interpretation and the original point picked by "Nature"—to both sender and receiver;
5. Sender and receiver independently adjust their behavior for future interactions.

If we let the game be played repeatedly and relate payoffs from each particular interaction to the "fitness" of sender and receiver strategies, that is, to the probability that these strategies will be employed in the next interaction, we get an evolutionary process with a specific dynamic.

This process can be viewed as an idealized model of how color categorization might evolve in a community. The cognitive demands on any particular agent may be relatively low in this model: the agent only needs to be able to match perceived colors with linguistic terms, as well as terms with perceived colors, and to update the pattern of matching based on the communicative success achieved by her strategy as well as by other strategies in the game. The similarity of particular colors remains an objective perceptual characteristic, but nothing depends on whether the agent is able to report on it for two arbitrary colors, or whether she consciously reflects on color similarity at all. Instead, we assume a naturalistic connection between the perceptual dissimilarity of two colors and the importance of distinguishing them in communication, which is reflected in the payoff function.

Of course, the level of idealization is high. We certainly do not claim that the existing color naming systems have actually evolved by routine repetition of game-like language exchanges within one homogeneous generation of speakers. We nonetheless believe that the principles involved in our model have played a major role in the evolution of color naming systems, whatever the time span and the social setting in which the actual process took place.

3.2 The model

Let us go into a more detailed formulation of the model. We use the chips of the Munsell array, with coordinates in the perceptual color space CIELAB, as the set of percepts, as is done by Regier, Kay, and Khetarpal (2007). This constitutes our state space, thus composed of 330 points, which can be indexed by hue (levels from 0 to 40) and lightness value (10 levels for the achromatic chips, 8 for the others) or by coordinates in the CIELAB space L, a, and b. We use the mapping from hue and value to CIELAB coordinates provided with the WCS data[9]. For illustration purposes, this state space is plotted in Figure 4a.

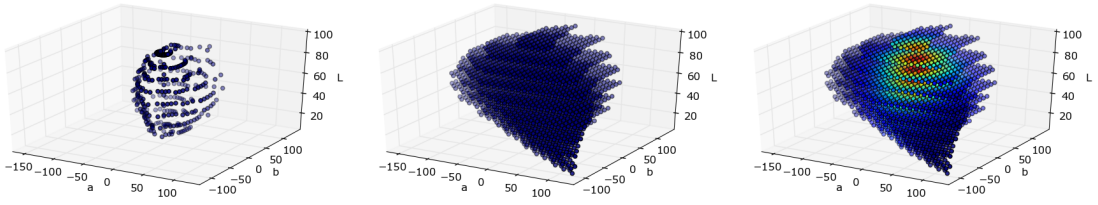
Given the CIELAB coordinates, the distance between two points in the state space $x_1 = \langle L_1, a_1, b_1 \rangle$ and $x_2 = \langle L_2, a_2, b_2 \rangle$ is simply given by their Euclidean distance, *i.e.*:

$$\mathrm{dist}(x_1, x_2) = \sqrt{(L_1 - L_2)^2 + (a_1 - a_2)^2 + (b_1 - b_2)^2}$$

The state space is thus effectively a subset of the CIELAB color space. As is done by Regier et al., on top of this we define a similarity metric as:

$$\mathrm{sim}(x_1, x_2) = e^{-c \cdot \mathrm{dist}(x_1, x_2)^2}$$

[9] Obtained from `http://www1.icsi.berkeley.edu/wcs/data/cnum-maps/cnum-vhcm-lab-new.txt`.



(a) **(b)** **(c)**

Figure 4: Different state spaces used in our models: (a) is based on the Munsell array, (b) and (c) are based on an estimate of the visible spectrum by Masaoka et al.; (a) and (b) have uniform priors, whereas (c) has more realistic priors (closer to red indicating higher probability here). Given that state spaces are three-dimensional, these plots can never provide a full picture of the state spaces. They only serve to illustrate differences in shape and prior probabilities between them. The full data is available online (Correia 2015b).

To be in line with their work, we use $c = 0.001$ for all simulations. Practically, this means that the similarity is decreasing almost linearly from 1 at distance 0 to about 0.8 at distance 50, and is virtually 0 for distances above 80.

The sim-max game is characterized by the tuple $\langle T, \Pr, M, U \rangle$, where T is the perceptual space described above, $\Pr \in \Delta(T)$ is the prior probability distribution over T, M is the set of messages available, and $U \in T \times T \to \mathbb{R}$ is the utility function for both sender and receiver (we assume "cheap talk", *i.e.* there are no message-dependent costs (see Stalnaker 2006)). We define only one utility function for both sender and receiver, thus we assume perfectly cooperative interests of both. Utility is equated with similarity: $U(x_1, x_2) = \mathrm{sim}(x_1, x_2)$. Unless otherwise stated, by default we use a uniform Pr, *i.e.* each point in the state space is *a priori* equally likely to be selected by "Nature".

Regarding game dynamics, we use the discrete-time replicator (Taylor and Jonker 1978). This equation was conceived to capture, in abstract terms, *differential reproduction*, that is, the idea that an individual's relative success in a population will have an impact on the likelihood that his traits are passed on to a new generation. This is a key notion in the theory of evolution and the replicator equation is the formalization most widely used in evolutionary game theory. The mathematical formulation is, however, agnostic of interpretations, and it can also be seen as capturing *differential imitation*, which is a form of cultural, rather than biological, evolution (Skyrms 2010, p. 55). When it comes to linguistic categorization, we are thus not claiming that the process of its evolution is tied to biology. It is a process where linguistic behaviors are selected with respect to how well they enable communication, and this happens by individuals adopting and favoring more successful strategies in place of less successful ones, not by passing actual genes on to biological generations.

Given that we are modeling differential imitation, rather than differential reproduction, we will use behavioral strategies. The main difference between these and mixed strategies is that, whereas a mixed strategy associates probabilities to whole pure strategies, a behavioral strategy associates probabilities to the set of possible choices (messages for senders, actions for receivers) at each choice point (states for senders, messages for receivers). This allows a behavioral strategy to evolve locally, at each choice point, which is more plausible if we are modeling imitation: an agent does not need to know the whole strategy of another to adopt his behavior for a given state, if that appears more successful. Formally, a sender strategy $\sigma \in T \to \Delta(M)$ associates with each point in the state space a probability distribution over the set of messages. A receiver strategy $\rho \in M \to \Delta(T)$ associates with each message a probability distribution over all points in the state space. Probability values can be interpreted as representing our uncertainty about an agent's behavior, an actual stochastic behavior, or percentages of a hypothetical population. In the context of this work, the latter is the most natural interpretation: think of $\sigma(x_1, m_1) = 0.7$ as representing that "70% of the population uses message m_1 when observing point x_1".

Replicator dynamics update the behavioral sender and receiver strategies according to their expected

utility. The state of each strategy at time instant $t+1$ is defined as follows:

$$\sigma_{t+1}(x,m) = \sigma_t(x,m) \times \frac{\mathrm{EU}_\sigma(x,m,\rho_t)}{\sum_{m' \in M} \sigma_t(x,m') \times \mathrm{EU}_\sigma(x,m',\rho_t)}$$

$$\rho_{t+1}(m,x) = \rho_t(m,x) \times \frac{\mathrm{EU}_\rho(m,x,\sigma_t)}{\sum_{x' \in T} \rho_t(m,x') \times \mathrm{EU}_\rho(m,x',\sigma_t)}$$

where expected utilities are defined as:

$$\mathrm{EU}_\sigma(x,m,\rho) = \sum_{x' \in T} \rho(m,x') \times U(x,x')$$

$$\mathrm{EU}_\rho(m,x,\sigma) = \sum_{x' \in T} \Pr(x') \times \sigma(x',m) \times U(x',x)$$

3.3 Results

Using the model described above, we ran 20 simulations[10] for 3, 4, 5, and 6 messages. Starting conditions σ_0 and ρ_0 are initialized with random values for every simulation. All simulations ran until a convergence criterion was met. Namely, simulations were stopped when the total absolute change in both sender and receiver strategy was under 1%, *i.e.* $\sum_{x \in T}\sum_{m \in M} |\sigma_{t+1}(x,m) - \sigma_t(x,m)| < 0.01$ and $\sum_{m \in M}\sum_{x \in T} |\rho_{t+1}(m,x) - \rho_t(m,x)| < 0.01$. Implementation of the model, code to run the simulations, and data analysis scripts are available online (Correia 2015a).

The part of the final state of a simulation that is used for evaluation is the sender strategy, which represents the task we are focusing on: assigning linguistic categories to colors. Given the interpretation of probabilities in the strategies described above, one can simply produce a majority map directly from the information in the sender strategy, much in the same way that is described for the WCS languages in Section 2.3. In practice, the majority maps obtained for our simulation results, unlike the majority maps for WCS languages, contained no gaps or missing data at all. The full data is available online (Correia 2015c).

Similarly to the notation used for the results by Regier et al., we denote these results as COM1-*n.i*, where n is the number of messages and i an incremental index to identify each result[11]. As discussed in Section 2.6, we will focus on analyzing the precision of the model in terms of the maximal similarity of each simulation result against the WCS languages. Furthermore, we will use the results of Regier, Kay, and Khetarpal (2007) and the discussion in Section 2.5 as somewhat of a baseline and a point of reference[12]. In Figure 5 we plot histograms of the maximal similarities obtained, as well as the value obtained by RKK's optimal color-naming scheme.

What these plots show is that for each value of n there is at least one color-naming scheme whose maximal similarity value is higher than the maximal similarity achieved by Regier et al. For 3, 5, and 6 terms, there is exactly one of 20 simulations outperforming their optimal scheme. On the face of it, this can be thought to be related to the fact that they obtain their optimal schemes as the best result of the 20 runs of their optimization algorithm, for each number of terms. It could indicate that, in terms of generating categorization schemes, sim-max games provide a valid alternative to the optimization procedure. However, the resemblance may be also accidental and misleading, and we should be careful in making such a claim. In the work of Regier et al. only one trial result per number of terms is reported, and it is not selected based on our measure of similarity. Given that optimality (in the authors' sense) and similarity with WCS languages are not necessarily correlated, we cannot discard the possibility that in the 19 unreported trials there were some that were less theoretically optimal, but more empirically successful.

[10] This number was chosen simply to be in line with Regier, Kay, and Khetarpal (2007). Although a study of the impact of this choice in the results could be performed, it would be outside the scope of this paper.

[11] The M in COM stands for Jan Mašek, who also contributed to our work. See the Acknowledgments section at the end of the paper.

[12] Note that we are comparing 20 of our runs against 1 of theirs, which is admittedly problematic. Ideally we would use all of their results, but we do not have access to them. Therefore, we would like to stress that this comparison is not fully conclusive.

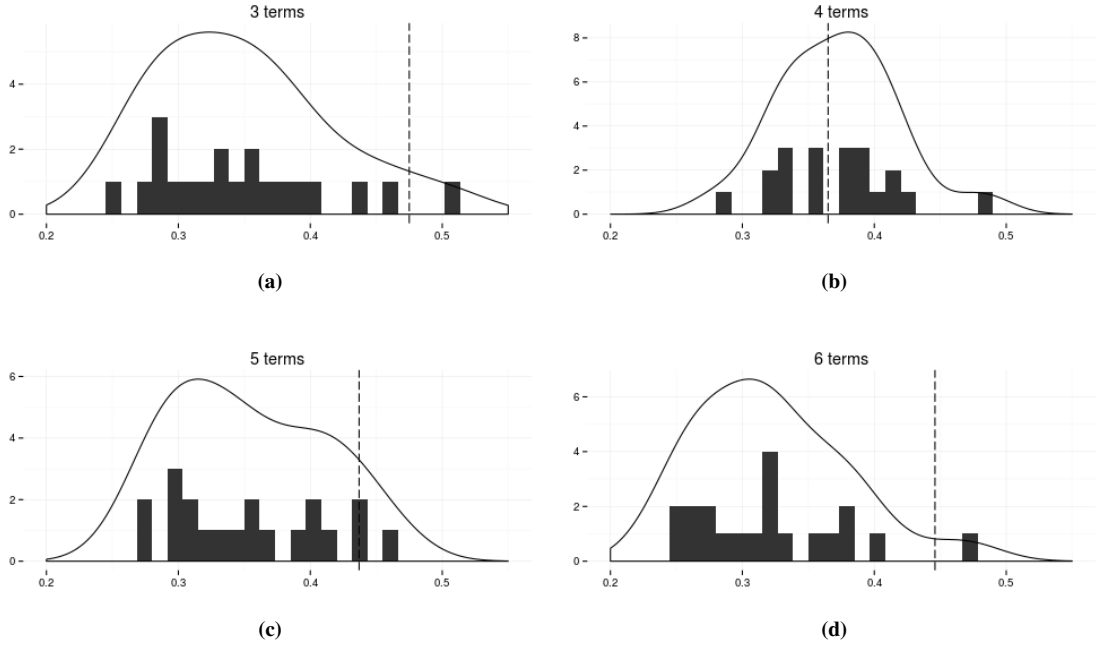


Figure 5: Histogram of maximal similarities of the resulting color-naming schemes of the first variant of the model: COM1. Solid line corresponds to a 1d kernel density estimate as produced with default parameters by the 'ggplot2' R package (Wickham 2009). Dashed line indicates the value obtained by RKK.

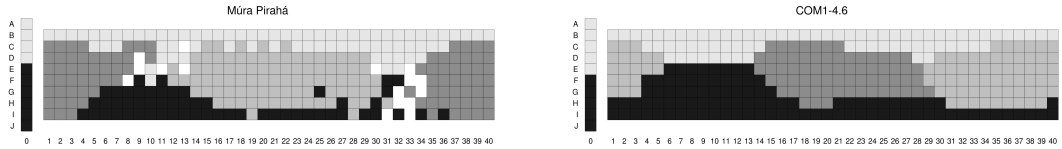


Figure 6: A 4-term WCS language with a term in the green region, and a similar result from the first variant of our model.

That could be related to why, for 4 terms, our procedure outperforms their optimal scheme in about half of the 20 cases, instead of just one. Their optimal scheme, with the "red" term extended into the blue region and with the vast "yellow" term reaching from reddish orange to yellowish green, does not find a very good counterpart in our WCS sample. Some of our simulations apparently do, without necessarily reaching the same level of theoretical optimality. Inspecting the results more up close reveals that many simulation outcomes have Múra Pirahá as their most similar language in the WCS. This is a language that has 4 color terms, but rather than having a term covering the yellow region of the space, it has one covering the green and blue region, much like our second highest scoring result (see Figure 6). Interestingly, our model also produced color-naming schemes that are similar to RKK-4, including the ones ranking first and third in maximal similarity values.

The fact that our model can produce results that are at least on par with those by Regier et al. means that sim-max games are not only a more plausible mechanism to explain the evolution of linguistic color categorization, but also one that makes reasonable predictions. However, both these aspects are far from perfection. Looking at the low end of the maximal similarity values, we find that many simulation results are not a good representative of any WCS language. Furthermore, some principles and assumptions behind the model are still far from realistic. In the following section we present more proposals to improve on

the theoretical foundations of our model; more specifically, we reconsider the perceptual constraints on the above described game-theoretic interaction.

4 Perceiving in full color

Regier et al. present their work as an elaboration of the proposal by Jameson and D'Andrade (1997) that the universal tendencies of color naming could be explained in terms of efficient divisions of the irregularly shaped perceptual space of color. However, what is actually partitioned in their model, as well as in our own previous proposal, is not very close to the actual spectrum of colors that humans can perceive. Rather, the models provide categorical partitions of the figure that results when the 330 chips of the Munsell array are represented in CIELAB. The figure has roughly the shape of the surface of a bumped sphere plus a middle axis formed by the achromatic chips from white to black (see Figure 4a). The relevance of this irregular figure with respect to the problem at hand seems limited in at least two respects. First, the set of points is neither representative of the full range of the visible spectrum, as per state-of-the-art research in color perception, nor does it cover the whole Munsell color system evenly, using only hue-value pairs at maximum saturation. Second, it does not in any way reflect the natural distribution of colors, *i.e.* how likely humans are to observe one color more often than the other.

As an instrument of the WCS, the Munsell array provided for the collection of invaluable evidence about color naming in human languages, and we will keep evaluating the outcomes of our model against this data. More specifically, we will represent any simulated partition of the color space in terms of how it would divide the Munsell figure, and compare these divisions to the categorical systems actually observed in the WCS. However, for the first reason stated above, it is not clear why efficient divisions of that particular figure should have any explanatory value with respect to the evolution of color categorization in languages. In order to examine the hypothesis by Jameson and D'Andrade (1997), we need to consider categorical partitions of the full spectrum of colors visible to humans, or at least a better approximation thereof.

4.1 Proposal 2: A more realistic perceptual space

Professor Albert H. Munsell devised his color order system in the early 20th century and its subsequent improvements took place in its first half. The abundant use of the Munsell array in the past color naming research derives from the popularity of Berlin and Kay (1969), who in turn follow Lenneberg and Roberts (1956) in this. Since Munsell's times, however, many advances have been made in the scientific understanding and characterization of human color perception, notably the formulation of the CIE XYZ system in 1931, of the CIELAB and CIELUV color spaces in 1976, and more recently of the color appearance model CIECAM02 (see Fairchild 2005). As noted above, we acknowledge the function of the Munsell array as a tool of empirical description. We, however, deprive it of the *explanatory* function it has been given in the previous models. The Munsell figure (the result of representing the array in the CIELAB color space) thus ceases to be the body which is primarily to be partitioned in our simulations. We are then looking for a more up-to-date approximation of the spectrum of visible color.

Masaoka et al. (2013) investigate the problem of estimating the number of human-discernible colors in light of state-of-the-art research in color perception. They produce estimates along various color models and several illumination conditions. Although they conclude that "the number of discernible object colors remains a conundrum" (Masaoka et al. 2013, p. 275), for the purposes of our modeling efforts, their estimates are a major step up in realism. Where they are concerned with calculating very accurate numbers, we merely require a representation of the spectrum of visible colors that is closer to reality than what was previously used. We understand and acknowledge the dimensions of variation that condition the estimation of the shape of the visible spectrum, from the theoretical limitations of our understanding (reflected in different models and representation techniques), to the choice of a model's particular variant based on specific viewing conditions, or even the inter-individual differences in visual abilities. However, for the sake of simplicity we will bite the bullet and make some pragmatic choices in order to focus on the problem at hand.

With that in mind, we used one of the estimates of the boundaries of the visible spectrum calculated by Masaoka et al. and reconstructed a color solid that could fit our modeling purposes. Namely, we took their

estimate in the CIELAB space for the 6500K illuminant(Masaoka et al. 2013, see first row, third column of Figure 10). (a good approximation of the daylight of a sunny day) and calculated a grid of points, with minimum distance 10 in each dimension, that would fit those boundaries. The result is a set of 2296 points ranging from 5.5 to 95.5, -160 to 130, and -120 to 140, respectively in the L, a, and b dimensions (see Figure 4b). Contrast this with the state space based on the set of Munsell chips, which consists of 330 points ranging from 15.6 to 96, -63.28 to 61.57, and -49.63 to 109.12, respectively in the L, a, and b dimensions. Clearly, according to more up-to-date estimates, there is much more to the spectrum than what the Munsell array can capture.

Regarding the concrete model presented in Section 3.2, the change we propose is simply to replace state space T. All formulations of the sim-max game are generic and thus remain the same. The change has an impact in evaluation: since T is no longer the Munsell array, a sender strategy $\sigma \in T \to \Delta(M)$ can no longer be used to directly map the array to categories. The solution we chose was to reconstruct a categorization for the Munsell array that represents a given sender strategy by taking, for each of the 330 Munsell points, the category assigned by that strategy to the point of the 2296 which is most similar to the Munsell point in the CIELAB space. Formally, we can think of it as follows: let T_M be the state space based on the Munsell array and T_S the new state space we are proposing, rather than producing a majority map from $\sigma : T_S \to \Delta(M)$ we produce it from a strategy $\sigma' : T_M \to \Delta(M)$, where for $t \in T_M$:

$$\sigma'(t) = \sigma(\arg\max_{t' \in T_S} \mathrm{sim}(t,t'))$$

Despite the added realism, this change to the model by itself does not make it perform too well. We will discuss the results in more detail in Section 4.3. Given our confidence that the change in state space is justified, we looked for additional important ingredients that could be missing. Naturally, there are many aspects that could be pushed to greater realism. From the options we considered, the frequency with which various colors occur in the natural environment was the one we felt could be taken into account with the greatest degree of objectivity.

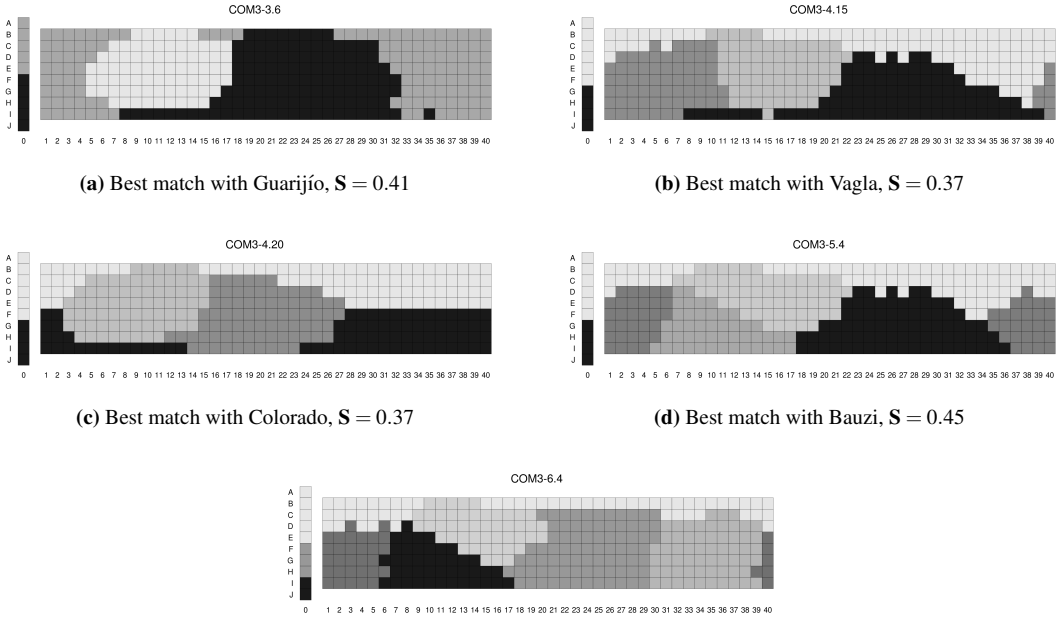
4.2 Proposal 3: Environment matters

In a realistic reflection of the phenomenon of color perception and categorization, it not only matters which colors we are able to see, but also which colors we actually see and how often. The model introduced in Section 3.2 is actually already prepared to accommodate this information in the prior probability distribution $\Pr \in \Delta(T)$, which encodes the likelihood of each point of the state space actually being observed. This influences the expected utility of the receiver strategy, which in turn plays a role in the evolutionary dynamics, thus having a potential impact on the development of the system as a whole.

All we need is to estimate the prior probability distribution of the state space. For this purpose, we use images of natural scenes available in the McGill Calibrated Colour Image Database (Olmos and Kingdom 2004), taking all images from each of the available categories. We randomly sample a number of individual pixels from these images, associate each of the CIELAB coordinates of the color represented by these pixels with the closest point in our state space, and build a frequency distribution of the latter. This of course slightly distorts the underlying continuous probability and can in practice never cover each and every point, so in order to smooth it out we perform a convolution of the frequency distribution with a negatively decaying distribution based on the normalized distance function. Mathematically, the implementation can be expressed as follows, where F is the frequency distribution, $\overline{\mathrm{dist}}$ the distance function normalized for the state space to range from 0 to 1, and λ a smoothing factor:

$$\Pr(x) = \sum_{x' \in T} F(x) \times (1 - \overline{\mathrm{dist}}(x,x'))^{\lambda}$$

Intuitively, what happens is that the prior probability of each point in the state space is influenced by the frequency distribution of all points (including itself), inversely weighted by their distance to it: the closer another point is, the more its frequency will influence the first point's prior probability. We explored several possible values for λ and pragmatically settled on $\lambda = 10$, since it seemed to provide a reasonable amount of smoothing without distorting the frequency distribution too much. Naturally, an infinite amount of other values could be used and we have no principled reason for this choice, but a full-fledged exploration of the



(a) Best match with Guarijío, $\mathbf{S} = 0.41$

(b) Best match with Vagla, $\mathbf{S} = 0.37$

(c) Best match with Colorado, $\mathbf{S} = 0.37$

(d) Best match with Bauzi, $\mathbf{S} = 0.45$

(e) Best match with Ocaina, $\mathbf{S} = 0.42$

Figure 7: Simulation results of COM3 with the highest similarity value against WCS languages per number of messages. We include both COM3-4.15 and COM3-4.20 because they achieve the same similarity value.

impact of this parameter on the results of the model would deviate us too much from the main questions at hand. An illustration of the outcome is presented in Figure 4c.

We thus derive a quasi-realistic probability of each color being picked by "Nature" in any particular round of the sim-max game. The attribute "quasi-realistic" is meant to indicate that we are in no way aiming at full-fledged realism in this respect: for that, one would need a faithful representation of the various natural environments concerned, consider various lighting conditions, etc. Our point rather is that we increase the realism of a particular modeling component, even if quite imperfectly, and see the impact of this change upon the model's predictive power. Note that the default, uniform probabilities, which have been used in the previous stages, were themselves an arbitrary choice.

4.3 Results

As was done for COM1, the first variant of the model proposed in Section 3, we ran 20 simulations for 3, 4, 5, and 6 messages for each of the other two variants. The full data is available online (Correia 2015d,e). In order to provide some concrete illustration, in Figure 7 we present the simulation results that achieve the highest similarity values for each COM3-*n*. Although we place more emphasis on quantitative evaluation than on visual impressions, some of these schemes have issues when it comes to dividing the spectrum in the ways that are familiar from existing languages. For instance, COM3-3.6 does not display the categories of black/dark and white/light, and COM3-4.20 has no dedicated category around red. The similarity values should not make us expect much more. Especially when put into context and compared with some similarities of languages within the WCS (see Table 3 again), they are not very high. More than in these particular results, we are interested in general trends and the relative impact of each modification to the model.

Figure 8 plots the distributions of maximal similarities achieved by the 20 runs of each variant of the model. The plots make clear that COM1 is the model that achieves the highest similarity values against the WCS languages, especially for 3 and 4 terms. As was mentioned in Section 4.1, COM2 achieves a generally lower performance. However, the simple addition of the quasi-realistic probabilities to the

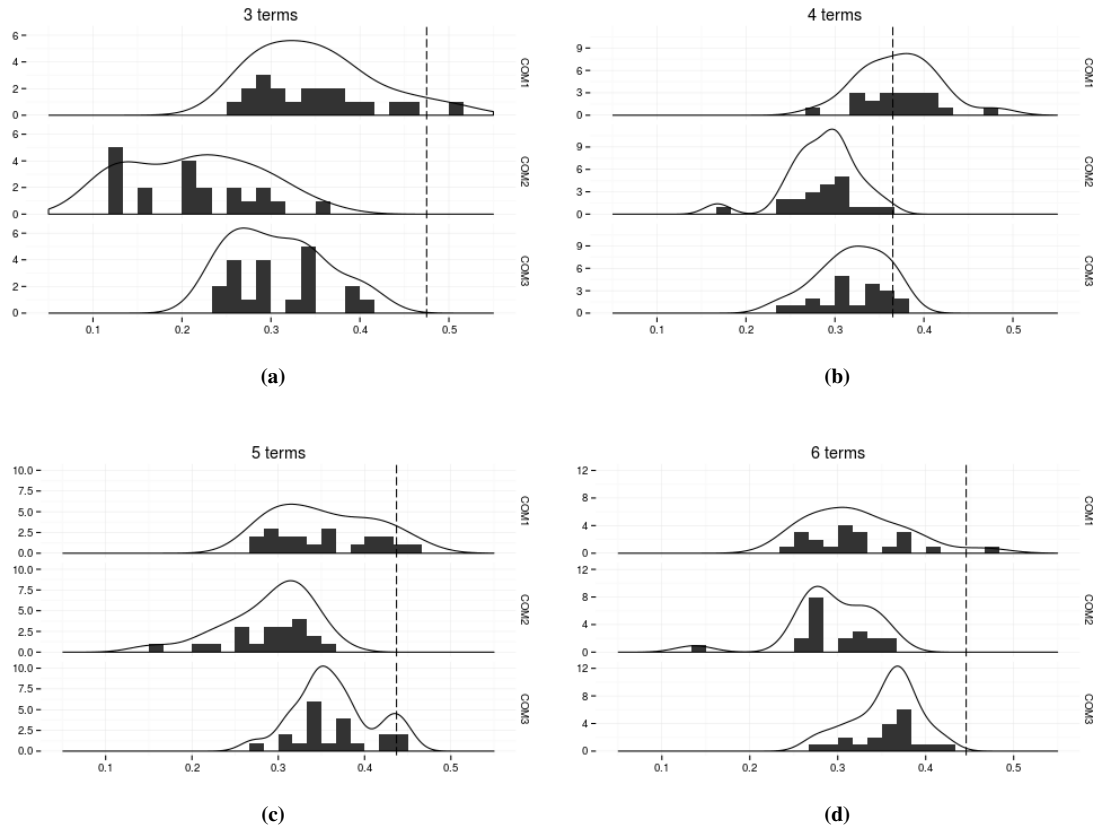


Figure 8: Histogram of maximal similarities for each color-naming scheme of each variant of the model. Solid line corresponds to a 1d kernel density estimate as produced with default parameters by the 'ggplot2' R package (Wickham 2009). Dashed line indicates the value obtained by RKK-*n*, for each *n*.

state space seems to have a very positive impact on the model's predictive performance, as seen by the distributions of the COM3 results. A statistical analysis (see Appendix A.2 for the details) supports these subjective observations.

In conclusion, replacing the Munsell array with a more realistic state space hurt the model's predictive power, but adding an environmental ingredient in the form of more realistic priors improved its performance again. This, we believe, is evidence that environment also matters and that we have to take into account the contribution of various factors if we are looking for a more complete explanation of the patterns of color categorization observed in real-world languages. We are, however, not there yet, as witnessed by the absolute similarity values as well as our very visual inspection of the simulations results.

5 Discussion

In this section, we discuss the significance of our model and its results, as well as a number of factors that we have abstracted away from, but which should arguably be taken into account in a more comprehensive approach to the color categorization phenomenon. Furthermore, we review the pragmatic choices that had to be made in various stages of our project, making the model more open to possible reconsideration.

5.1 Significance of the results

As has been stated already, a more general ambition of the present paper is to elaborate on the very notion of satisfactory explanation in the domain of color categorization. This includes giving importance to the requirement of independent motivation for any principles involved in a color categorization model. Furthermore, devising a transparent method of evaluation allows us to better assess the predictive power of such models. With these proposals, we are putting an emphasis on objectivity: if we can build a concrete model to embody whatever theories we might have on color categorization, we should also strive for analyzing its results objectively in order to increase the confidence in whatever claims we make. In a similar spirit, we presented and examined our proposals in an incremental fashion so that we could better assess the separate impact of each. Given that the approach of Regier, Kay, and Khetarpal (2007) has produced reasonable results (as scrutinized in Section 2.5), we used it as a reference for analyses of each of our proposals. The confrontation of the results of COM3 with those of Regier, Kay, and Khetarpal, as well as COM1 (where we mostly adopt their assumptions), has shown that our final model achieves comparable predictive performance, although not across-the-board. However, it is important to see our model not as a finished product, but more as a baseline for future models with similar amount of realism in mind. If COM3 is, as we claim, superior in terms of motivation for the principles used and achieves comparable results, it is important for future models to consider the relevance of these principles as explanatory factors.

In absolute terms, it is revealing to confront the results of COM3 with the sample of WCS inter-similarities presented in Table 3. Based on these comparisons, there is no denying that we are still quite far from providing an explanation for the empirical color naming patterns, or at least one that would be truly satisfactory or exhaustive. This is the more true given that we have primarily focused on the precision part of the modeling problem, while leaving aside the recall aspect. This conclusion is in sharp contrast with the strong (and, in our opinion, hasty) explanatory claims raised by Regier, Kay, and Khetarpal (2007), Regier, Kemp, and Kay (2015), Baronchelli et al. (2010), and Loreto, Mukherjee, and Tria (2012). In particular, our results do not confirm the hypothesis by Jameson and D'Andrade (1997); they rather suggest (contra Regier, Kay, and Khetarpal (2007) and Regier, Kemp, and Kay (2015)) that the shape of the color space on its own is by far not a sufficient *explanans* for the empirical patterns of color categorization. It is, however, an important and promising result that with the imposition of an additional (quasi-)realistic environmental constraint on our basic scenario, the performance of the model has markedly increased (the comparison between COM2 and COM3). Apparently environment also matters.

That leads us to the question of what further factors are involved in the extremely complex phenomenon of color categorization by the languages of the world, and which of them could possibly be incorporated in a model of the present type.

5.2 The multifarious phenomenon of color

Our choice of the occurrence probability of colors in the natural environment as the factor to be included was primarily motivated by the relative ease of implementation. There is a number of other factors that would be very much worth examining in the present context, but which had to be left for future research.

For instance, we might want to take into account *color blindness*, *i.e.* the existence in human populations of observers whose perceived similarities between objectively specified colors variously differ from the standard ones which are accounted for by perceptual color spaces such as CIELAB (see Baylor 1995; Mollon 1995; Fairchild 2005). Jameson and Komarova (2009a,b) were the first to explore the impact of a proportion of such observers upon the emergent color categorization system. We could try to incorporate that into our model and test their results from a different point of view.

Besides potential variation in the perceptual apparatus of each individual, inter-individual variation in linguistic color categorization can also be connected to the phenomenon of semantic *vagueness*. Different ways of accounting for vagueness within the framework of signaling games have been brought forward by Franke, Jäger, and van Rooij (2011), O'Connor (2014), and Franke and Correia (2018). Moreover, in some of this work it has been argued that incorporating vagueness in our models of linguistic behavior is not only a necessity to be faithful to natural language use, but it can also actually have beneficial effects such as enabling faster convergence to more homogeneous outcomes. This could present a potential improvement to the models presented in this paper.

We could also try to incorporate the possible factor of uneven environmental significance of colors across the perceptual space: for survival and prosperity in a particular environment, some colors are probably more important to distinguish than others. Provided we had an independent specification of this empirical constraint, it could be reflected in the utility function U of our sim-max signaling interaction. However important this factor might be, we do not envisage the necessary data being accurately produced in the near future, given the difficulty of its empirical collection.

Similarly, the phenomenon usually referred to as categorical perception of color in infants and prelinguistic toddlers seems to constitute a potentially relevant constraint on the evolution of categorical systems (Ocelák 2016).

5.3 Pragmatic choices

In order to implement what we believe are new and promising ideas in a computational model, we had to make some pragmatic choices. First, we do not claim that our use of CIELAB as the approximation of the perceptual space of color is the only, or the best choice available. The same could be said about the estimate used to represent the spectrum of visible colors in COM2 and COM3. One could probably argue for using other models of color perception, or perhaps for some of their particular variants based on specific viewing conditions (lighting, background etc.). We should also note that the very possibility of an adequate three-dimensional, Euclidean representation of color similarities has been questioned, particularly for large color distances (see Kuehni 2002).

A number of pragmatic choices had to be made regarding certain parameters. These concern the aggregation over the informants of particular WCS languages, the imputation process for our similarity measure, the utility function used in the game-theoretic interaction, stopping criteria for simulations, the smoothing of probabilities calculated from natural scenes, and maybe more we are not even aware of. We tried to select sensible values for these parameters, since searching the space of possibilities would be impractical and would deviate us from the main points at hand. Also, models of the kind proposed in this paper are more akin to thought experiments than to hyper-realistic representations, making a search for the perfect parameter settings less imperative or even desirable. However, we cannot exclude the possibility that certain parameters could have a significant impact on the results obtained. In the future, studying other settings of these parameters could be considered, preferably based on independent motivations.

Another pragmatic choice was to consider fixed sets of 3 to 6 messages in all variants of our model, one that provided for clearer comparison with the work of Regier, Kay, and Khetarpal (2007). A more realistic option would arguably be not to determine the set of available messages from the start, but to let agents invent new messages in the progress of communication, as in the work of Alexander, Skyrms, and Zabell (2012). We see extending the model to incorporate this as important future work.

5.4 Modeling approach

The use of signaling games is a point in which we diverge from the loosely game-theoretic approach to color categorization of the agent-based models mentioned in Section 1. We do so for principled reasons. The models of Steels and Belpaeme (2005), and those by Baronchelli et al. (2010) and Loreto, Mukherjee, and Tria (2012), operate with two levels of categorization above a continuous color space: the perceptual and the linguistic. We find this distinction problematic since it seems to subscribe to the notion of pre-linguistic concepts as mental representations. As we share some of the philosophical skepticism about this notion, we prefer to try to do without it.

The approach of Komarova, Jameson, and Narens (2007) (and more recently Park et al. 2018) depends on the notion of a similarity range (k-similarity) that is used to define the success (or lack thereof) of a categorization interaction between two agents. We find the concept of a fixed range of similarity, upon which a binary definition of success is construed, cumbersome and difficult to motivate. We prefer to think of success as gradual, and similarity as continuous, as captured in the utility function of sim-max games.

Dowman (2007) defines a Bayesian inference model that is motivated by a picture of language as purely a form of expression. As such, the learning process of agents is not influenced by any notion of communicative success. They learn by inducing models of color categorization systems based on examples provided by either the experimenter or other agents. Even though admitting that these are relevant dimensions of our linguistic practices, we find this picture of language limited and believe that the game-theoretic approach, where adaptation of strategies is driven by communicative success, better captures what underlies the evolution of color categorization systems.

We chose to develop our model along the paradigm of signaling games because we believe it avoids some of what we see as limitations of these other approaches in the literature. Moreover, we believe our model is simpler, more general (via the use of abstract population dynamics), computationally more tractable, and has the added advantage of being designed to allow for direct testing against the empirical data. One important aspect in which it is, however, less realistic than the agent-based models discussed here, is that the equilibria it achieves can be characterized as homogeneous populations (the interpretation being that every agent uses the same categorization system). This is not in line with the empirical data. Given its generality, it is possible to interpret the model in terms of individual learning dynamics rather than in terms of evolutionary processes, but that would require a reformulation of some of the motivation given in this paper.

We see the model simply as a tool to test general assumptions about the evolution of linguistic color categorization. As such, we are not committed to any form of metaphysical realism regarding its elements. Because of that, and because we believe that there are more similarities than differences between all of these approaches, we would have no qualms about having this particular model re-implemented in terms of a different paradigm, provided that the implementation avoids the issues pointed out here.

6 Conclusion

We proposed a new color categorization model that roughly follows other game-theoretic approaches in implementing linguistic interaction on top of the level of individual color perception. Furthermore, we attempted to provide improvements in the motivation of the perceptual principles used in its architecture, and provided a more naturalistic explication for the emergence of color categories from linguistic interaction. In addition, we have developed a transparent evaluation methodology which enables us to assess the predictive power of our model, as well as other models, with respect to the empirical patterns of color categorization (color naming) in the languages of the world.

The results indicate that the performance of our model is roughly comparable to that of a previous model by Regier, Kay, and Khetarpal (2007). Given the added realism, we believe that its explanatory relevance with respect to the phenomenon in question is higher. In particular, our results do not confirm the hypothesis by Jameson and D'Andrade (1997) that these patterns are due solely to the irregular shape of the color space. However, imposing an additional (quasi-)realistic environmental constraint on the model had a clear positive effect on the model's predictive power. This suggests a very promising direction for future research.

Further predictive success in the development of the present model could have strong implications for how we conceive the nature of color categories in human language and thinking. It adds support to the view that at least some of our linguistic categorization practices do not originate in a single source, but are results of a rich interplay between factors innate, environmental, as well as interactional. In absolute terms, we are still far from a full-fledged explanation of the cross-linguistic tendencies of color naming. We hope, however, that the proposal advanced here can serve as a step towards more realistic modeling of the phenomenon of linguistic color categorization.

Acknowledgments

We would like to greatly thank Jan Mašek, who conducted the curation of the WCS languages and also contributed to other parts of Section 2. The paper has additionally benefited from the input of Michael Franke and Elliott Wagner, who supervised the project that led to this research and provided additional feedback, Gernot Hoffmann, who pointed us to research on estimation of the visible spectrum, and Kenichiro Masaoka, who provided us with valuable data. Helpful comments were furthermore provided by Ondřej Beran, Henk Zeevat and Reinhard Blutner. All potential mistakes in the paper are naturally the authors' responsibility. We furthermore acknowledge support from the research project GA UK No. 330214 'Color and Meaning' at Charles University. Finally, during the end stage of this work, José Pedro Correia benefited from a PhD studentship by Fundação para a Ciência e a Tecnologia (grant number SFRH/BD/100437/2014).

A Additional material

A.1 Additional results by Regier, Kemp, and Kay

In Table 4 we present similarity values for the results of Regier, Kemp, and Kay (2015), in an analogous way as was done in Table 2. We refer to their results as RKeK-*n*, with *n* corresponding to the number of color categories. The full data is available online (Correia 2015f).

	RKeK-3			RKeK-4			RKeK-5			RKeK-6		
	Language	**S**	ARI	Language	**S**	ARI	Language	**S**	ARI	Language	**S**	ARI
Top 3	*Bété*	**0.38**	0.51	Wobé	**0.38**	0.40	Bauzi	**0.50**	0.56	Bauzi	**0.46**	0.53
	Wobé	**0.37**	0.38	Colorado	**0.38**	0.50	*Colorado*	**0.45**	0.60	Colorado	**0.42**	0.56
	Culina	**0.36**	0.56	Bauzi	**0.36**	0.41	Múra Pirahá	**0.42**	0.47	*Cofán*	**0.38**	0.48
Additional	*Ejagam*	**0.31**	0.32	*Culina*	**0.34**	0.54	*Iduna*	**0.34**	0.53	*Buglere*	**0.36**	0.54
							Cayapa	**0.34**	0.59			

Table 4: Similarities between RKeK results and some languages in the WCS.

A.2 Statistical comparison of COM1, COM2, and COM3

In order to compare the distributions of maximal similarity for particular variants of the model, we can use the Mann–Whitney–Wilcoxon (MWW) *U* test, a non-parametric statistical test of the null hypothesis that two sets of values are drawn from the same distribution. Not being able to reject the null hypothesis based on the test would mean that the values obtained by the two variants are not sufficient to distinguish them, that is, that we cannot reject the possibility that the models produce equivalent results, at least regarding how well they match real-world languages. Table 5 contains the results of the test.

Comparing COM1 and COM2, for 3, 4, and 5 messages the two are significantly different. The positive location shift[13] indicates COM1 has a tendency to produce results that are more similar to the WCS languages than COM2. For 6 messages, however, the two are not distinguishable. COM3 seems to be a clear improvement from COM2 producing significantly better results across the board. When comparing COM3

[13]Location shift δ is an estimator for the median difference between a sample from the first distribution and a sample from the second, not an estimator of the difference of the medians. See the documentation of the implementation for more details (R Core Team 2014).

	median			COM1 vs. COM2			COM2 vs. COM3			COM1 vs. COM3		
	COM1	COM2	COM3	U	δ	p-value	U	δ	p-value	U	δ	p-value
3	0.338	0.216	0.296	369	0.134	**5.15E-06	66	-0.100	**3.04E-04	270	0.037	6.01E-02
4	0.376	0.293	0.319	376	0.083	**2.06E-06	104	-0.033	**9.76E-03	331	0.049	**4.15E-04
5	0.347	0.300	0.355	306	0.059	**4.31E-03	45	-0.065	**2.92E-05	172	-0.016	4.49E-01
6	0.316	0.281	0.364	241	0.021	2.79E-01	54	-0.059	**7.82E-05	119	-0.041	*2.84E-02

Table 5: Results of the MWW test between each pair of variants of the model. Calculations performed in R (R Core Team 2014) using the 'wilcox.test' function. We report median similarities for each variant of the model, the value of the U statistic, the location shift δ between the two distributions, and p-values. Significant differences are marked in the p-values, with ** for a 99% confidence level and * for 95%.

with COM1 the story is not so linear. Similarities for 3 and 5 messages are indistinguishable between the two variants at both 99% and 95% confidence levels. For 4 messages, COM1 seems to produce consistently better results than COM3, which is also patent in the plots in Figure 8. For 6 messages, however, it seems to be the other way around, albeit only at a 95% confidence level.

References

Alexander, J. McKenzie, Brian Skyrms, and Sandy L. Zabell (2012). "Inventing new signals". In: *Dynamic Games and Applications* 2.1, pp. 129–145.

Baronchelli, A. et al. (2010). "Modeling the emergence of universality in color naming patterns". In: *Proceedings of the National Academy of Sciences* 107, pp. 2403–2407.

Baylor, D. (1995). "Colour Mechanisms of the Eye". In: *Colour: Art and Science*. Ed. by T. Lamb and J. Bourriau. Press Syndicate of the University of Cambridge, pp. 103–126.

Belpaeme, Tony (2014). "Color Category Learning in Naming-Game Simulations". In: *Encyclopedia of Color Science and Technology*, pp. 1–5.

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

Cook, R., P. Kay, and T. Regier. *WCS Data Archives*. URL: `http://www.icsi.berkeley.edu/wcs/data.html` (visited on 01/2013).

Correia, J.P. (2015a). "COM model and data analysis scripts". In: URL: `http://dx.doi.org/10.6084/m9.figshare.1428652`.

— (2015b). "COM state spaces". In: URL: `http://dx.doi.org/10.6084/m9.figshare.1428659`.

— (2015c). "COM1 results". In: URL: `http://dx.doi.org/10.6084/m9.figshare.1411267`.

— (2015d). "COM2 results". In: URL: `http://dx.doi.org/10.6084/m9.figshare.1411268`.

— (2015e). "COM3 results". In: URL: `http://dx.doi.org/10.6084/m9.figshare.1411269`.

— (2015f). "RKeK results". In: URL: `http://dx.doi.org/10.6084/m9.figshare.1411252`.

— (2015g). "RKK results". In: URL: `http://dx.doi.org/10.6084/m9.figshare.1411250`.

— (2015h). "WCS majority maps". In: URL: `http://dx.doi.org/10.6084/m9.figshare.1411240`.

Dedrick, D. (1998). *Naming the Rainbow: Colour Language, Colour Science, and Culture*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

— (2006). "Explanation(s) and the Patterning of Basic Colour Words Across Languages and Speakers". In: *Progress in Colour Studies, Vol. II: Psychological Aspects*. Ed. by N. J. Pitchford and C. P. Biggam. Amsterdam/Philadelphia: John Benjamins, pp. 1–11.

Dowman, Mike (2007). "Explaining color term typology with an evolutionary model". In: *Cognitive Science* 31.1, pp. 99–132.

Fairchild, M. D. (2005). *Color Appearance Models*. The Wiley-IS&T Series in Imaging Science and Technology. Chichester, England: Wiley.

Fraley, Ch., A. Raftery, and L. Scrucca (2014). *mclust: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*. R package version 4.3. URL: `http://CRAN.R-project.org/package=mclust`.

Franke, M., G. Jäger, and R. van Rooij (2011). "Vagueness, Signaling & Bounded Rationality". In: *JSAI-isAI 2010*. Ed. by T. Onoda, D. Bekki, and E. McCready. Springer, pp. 45–59.

Franke, Michael and José Pedro Correia (2018). "Vagueness and Imprecise Imitation in Signalling Games". In: *The British Journal for the Philosophy of Science* 69.4, pp. 1037–1067.

Franke, Michael and Elliott Wagner (2014). "Game Theory and the Evolution of Meaning". In: *Language and Linguistics Compass* 8.9, pp. 359–372.

Huttegger, Simon M. (2014). "How Much Rationality Do We Need to Explain Conventions?" en. In: *Philosophy Compass* 9.1, pp. 11–21.

Jäger, G. (2010). "Natural color categories are convex sets". In: *Logic, Language and Meaning - 17th Amsterdam Colloquium, Amsterdam, The Netherlands, December 16-18, 2009, Revised Selected Papers*. Ed. by M. Aloni et al., pp. 11–20.

Jäger, G. and R. van Rooij (2007). "Language structure: psychological and social constraints". In: *Synthese* 159, pp. 99–130.

Jameson, K. (2010). "Where in the World Color Survey is the support for the Hering primaries as the basis for color categorization?" In: *Color Ontology and Color Science*. Ed. by J. D. Cohena and M. Matthen. Cambridge, Massachusetts: MIT Press, pp. 179–202.

Jameson, K. and R. G. D'Andrade (1997). "It's not really red, green, yellow, blue: an inquiry into perceptual color space". In: *Color Categories in Thought and Language*. Ed. by C. L. Hardin and L. Maffi. Cambridge University Press, pp. 295–319.

Jameson, K. J. and N. L. Komarova (2009a). "Evolutionary models of color categorization. I. Population categorization systems based on normal and dichromat observers". In: *J. Opt. Soc. Am.* 26, pp. 1414–1423.

— (2009b). "Evolutionary models of color categorization. II. Realistic observer models and population heterogeneity". In: *J. Opt. Soc. Am.* 26, pp. 1424–1436.

Kallens, Pablo Andrés Contreras, Rick Dale, and Paul E. Smaldino (2018). "Cultural Evolution of Categorization". In: *arXiv preprint arXiv:1803.06588*.

Kay, P. (1975). "Synchronic variability and diachronic change in basic color terms". In: *Language in Society* 4, pp. 257–270.

Kay, P. and L. Maffi (1999). "Color appearance and the emergence and evolution of basic color lexicons". In: *American Anthropologist* 101, pp. 743–760.

Kay, P. and C. K. McDaniel (1978). "The linguistic significance of the meanings of basic color terms". In: *Language* 54, pp. 610–646.

Kay, P. and T. Regier (2003). "Resolving the question of color naming universals". In: *Proceedings of the National Academy of Sciences* 100, pp. 9085–9089.

Kay, P. et al. (2009). *The World Color Survey*. Stanford: Center for the Study of Language and Information.

Kay, Paul and Richard S. Cook (2016). "World Color Survey". In: *Encyclopedia of Color Science and Technology*. Ed. by Ming Ronnier Luo. New York, NY: Springer New York, pp. 1265–1271.

Komarova, Natalia L., Kimberly A. Jameson, and Louis Narens (2007). "Evolutionary models of color categorization based on discrimination". In: *Journal of Mathematical Psychology* 51.6, pp. 359–382.

Kuehni, R. G. (2002). "CIEDE2000: Milestone, or final answer?" In: *COLOR research and application* 27, pp. 126–127.

Lenneberg, E. H. and J. Roberts (1956). "The language of experience: A study in methodology". In: *Indiana University Publications in Anthropology and Linguistics, Memoir 13*. Baltimore: Waverly Press.

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

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

Lucy, J. A. (1997). "The linguistics of "color"". In: *Color Categories in Thought and Language*. Ed. by C. L. Hardin and L. Maffi. Cambridge University Press, pp. 320–346.

Lyons, J. (1995). "Colour in Language". In: *Colour: Art and Science*. Ed. by T. Lamb and J. Bourriau. Press Syndicate of the University of Cambridge, pp. 175–193.

Masaoka, K. et al. (2013). "Number of discernible object colors is a conundrum". In: *J. Opt. Soc. Am. A* 30.2, pp. 264–277.

Mollon, J. (1995). "Seeing Colour". In: *Colour: Art and Science*. Ed. by T. Lamb and J. Bourriau. Press Syndicate of the University of Cambridge, pp. 127–150.

Ocelák, R. (2013). "Carving up the rainbow: how to model linguistic categorization of color". Master of Logic (MoL) Thesis Series, ILLC, University of Amsterdam, The Netherlands. MA thesis.

Ocelák, R. (2015). “The myth of unique hues”. In: *Topoi* 34.2, pp. 513–522.
— (2016). ““Categorical perception” and linguistic categorization of color”. In: *Review of Philosophy and Psychology* 7.1, pp. 55–70.
O’Connor, Cailin (2014). “The evolution of vagueness”. In: *Erkenntnis* 79.4, pp. 707–727.
Olmos, A. and F. A. A. Kingdom (2004). “A biologically inspired algorithm for the recovery of shading and reflectance images”. In: *Perception* 33.12, pp. 1463–1473.
Park, Jungkyu et al. (2018). “The Evolution of Shared Concepts in Changing Populations”. In: *Review of Philosophy and Psychology*.
R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `http://www.R-project.org/`.
Rand, W. M. (1971). “Objective criteria for the evaluation of clustering methods”. In: *Journal of the American Statistical Association* 66.336, pp. 846–850.
Regier, T., P. Kay, and R. S. Cook (2005). “Focal colors are universal after all”. In: *Proceedings of the National Academy of Sciences* 102, pp. 8386–8391.
Regier, T., P. Kay, and N. Khetarpal (2007). “Color naming reflects optimal partitions of color space”. In: *Proceedings of the National Academy of Sciences* 104, pp. 1436–1441.
Regier, T., C. Kemp, and P. Kay (2015). “Word meanings across languages support efficient communication”. In: *The Handbook of Language Emergence*. Ed. by B. MacWhinney and W. O. Grady. Hoboken, NJ: Wiley.
Roberson, D., I. Davies, and J. Davidoff (2000). “Color categories are not universal: Replications and evidence form a stone-age culture”. In: *Journal of Experimental Psychology: General* 126, pp. 369–398.
Roberson, D. et al. (2005). “Color categories: evidence for the cultural relativity hypothesis”. In: *Cognitive Psychology* 50, pp. 378–411.
Saunders, B. (2000). “Revisiting *Basic color terms*”. In: *The Journal of the Royal Anthropological Institute* 6, pp. 81–99.
— (2007). “Towards a new topology of color”. In: *Anthropology of Colour: Interdisciplinary Multilevel Modeling*. Ed. by R. E. MacLaury, G. V. Paramei, and D. Dedrick. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 467–479.
Saunders, B. and J. van Brakel (1997a). “Are there nontrivial constraints on colour categorization?” In: *Behavioral and Brain Sciences* 20, pp. 167–228.
— (1997b). “Colour: An exosomatic organ?” In: *Behavioral and Brain Sciences* 20, pp. 212–220.
Skyrms, B. (2010). *Signals: Evolution, Learning, and Information*. Oxford University Press.
Skyrms, Brian (1996). *Evolution of the Social Contract*. Cambridge University Press. 162 pp.
Stalnaker, R. (2006). “Saying and meaning, cheap talk and credibility”. In: *Game Theory and Pragmatics*. Ed. by A. Benz, G. Jäger, and R. van Rooij. New York: Palgrave MacMillan, pp. 83–100.
Steels, L. and T. Belpaeme (2005). “Coordinating perceptually grounded categories through language: A case study for colour”. In: *Behavioral and Brain Sciences* 28, pp. 469–529.
Taylor, P. D. and L. B. Jonker (1978). “Evolutionarily Stable Strategies and Game Dynamics”. In: *Mathematical Biosciences*, pp. 145–156.
Vinh, N. X., J. Epps, and J. Bailey (2010). “Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance”. In: *Journal of Machine Learning Research* 11, pp. 2837–2854.
Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. New York: Springer.
Witzel, Christoph (2018). “Misconceptions about colour categories”. In: *Review of Philosophy and Psychology*, pp. 1–42.