Color terms: Native language semantic structure and artificial language structure formation in a large-scale online smartphone application

Thomas F. Müller[a,b]*, James Winters[a,c], Tiffany Morisseau[d,e], Ira Noveck[f], & Olivier Morin[a,g]

[a]Minds and Traditions Research Group, Max Planck Institute for the Science of Human History, Jena, Germany

[b]Center for Humans and Machines, Max Planck Institute for Human Development, Berlin, Germany

[c]Collective Computation Group, School of Collective Intelligence, UM6P, Ben Guerir, Morocco

[d]Université de Paris, LAPEA, Boulogne-Billancourt, France

[e]LAPEA, Université Gustave Eiffel, IFSTTAR, Versailles, France

[f]Laboratoire de Linguistique Formelle, UMR 7110, University of Paris, CNRS, Paris, France

[g]Institut Jean Nicod, ENS, EHESS, PSL University, CNRS, Paris, France

*mueller@mpib-berlin.mpg.de

Abstract

Artificial language games give researchers the opportunity to investigate the emergence and evolution of semantic structure, i.e. the organization of meaning spaces into discrete categories. A possible issue for this approach is that categories might simply carry over from participants' native languages, a potential bias that has mostly been ignored. We investigate this in a referential communication game by comparing color terms from three different languages to those of an artificial language. Here, we assess the similarity of the semantic structures, and test the influence of the semantic structure on artificial language communication. We compare the in-game communication to a separate online naming task providing us with the native language structure. Our results show that native and artificial language structure overlap at least moderately. Furthermore, communicative behavior and performance were influenced by the shared semantic structure, but only for English-speaking pairs. These results imply a cognitive link between participants' semantic structures and artificial language structure formation.

Keywords: semantic structure, artificial language, language evolution, smartphone application, color terms, categorical facilitation

1. Introduction

One striking feature of human language is that it exhibits structure on a variety of levels (Everaert et al., 2015). For instance, a limited number of phonological units that are meaningless by themselves are combined into a much higher number of meaningful words (duality of patterning: Hockett, 1960); morphemes (single units of meaning) combine to form more complex phrases; and the semantic space is organized into discrete categories that allow us to structure and successfully communicate an otherwise intractable and infinite number of meanings (Lakoff, 1987). This is what we call the *semantic structure* of a language: It refers to the way a language divides a meaning space into linguistic categories (Youn et al., 2016; "categorical structure" in Carr et al., 2017; Malt et al., 2003). For example, different objects that can be designated by the term "furniture" can be distinguished in English using words such as "chair", "table", "sofa", or "bed". This is based on the respective features of the objects, like their physical properties and usage (e.g. a chair is used for sitting, while a table is typically used to place objects rather than humans on it; meanwhile, there can still be overlap in properties like having four legs, being made of wood, etc.). Another example would be the domain of color: Here, discrete color terms like "red" or "green", but also "crimson" or "steel-blue", structure the entire space of colors perceivable by humans to make them communicable to others.

In this study, we investigate the evolution of the semantic structure of color terms in an artificial language game, namely an online smartphone application called the "Color Game". In particular, we link this artificial language, which emerged through repeated interactions between individuals, to the semantic structure found in three natural languages. This is important because almost none of the past studies that focused on the evolution of semantic structure using artificial language games have been concerned with a possible *bias for native language structure*. We also draw on previous literature on color terms and *categorical facilitation* to ask whether artificial language communication is influenced by the semantic structure inferred from the native language.

Artificial language games, semantic structure, and possible biases

Artificial language games are an appropriate method to study the evolution of linguistic structure in a controlled environment (for an overview, see Galantucci, 2009; Galantucci et al., 2012; Galantucci & Garrod, 2011; Scott-Phillips & Kirby, 2010; Tamariz, 2017). These tasks typically request participants to communicate without a pre-established set of conventional signs. Here, the challenge is to map novel and unusual signals onto a space of meanings. As such, the respective *signal space* and *meaning space* are important features of the task. To circumvent the use of natural language,

previous experiments have, for example, used non-words (Kirby et al., 2008), spontaneous gesturing (Nölle et al., 2018), or even the movement patterns of a virtual agent (Scott-Phillips et al., 2009). Likewise, meaning spaces in these experiments ranged from moving shapes of different colors (Kirby et al., 2008) to cartoon characters of different professions (Nölle et al., 2018) and differently colored locations within the game (Scott-Phillips et al., 2009).

How can artificial language games ensure that evolving conventions are novel? Two key features that can help here are the *avoidance of prior meanings* and the *reliance on unusual signals* – i.e., "genuinely 'alien' form spaces" (Cuskley, 2019, p. 3). One example is the groundbreaking study by Galantucci (2005): Here, pairs of participants were tasked to coordinate to move to the same room in a virtual environment with only a single movement possible for each player. To do so above chance, they had to communicate by making use of a novel graphical communication device. Crucially, the device prevented the use of conventional letters or numbers by only enabling control over the drawings' horizontal (but not vertical) trace, as well as rapidly fading potential signals as time passed. Still, participants managed to achieve successful communication, which is evidence for the evolution of novel conventions. A similar case is the experiment by Scott-Phillips et al. (2009), who tasked participants to coordinate in a comparable virtual environment. In their case, pairs of participants could not even make use of a pre-defined communication channel, but only observe the partner's movements in the virtual space instead. Many pairs still manage to establish successful communication through their movements on the screen; in doing so, they have to recognize the meaning of the partner's movements, and that the partner is attempting to communicate in the first place. Two other examples that made use of highly unusual signal spaces are the studies by Verhoef et al. (2014), who had participants learn and transmit sounds by using a slide whistle, and Cuskley (2019), who used graphemes originally created with ferrofluid ink; however, neither of these studies included a meaning space that the signals had to be mapped onto.

One key artificial language game that we heavily build on is the study by Müller et al. (2019), which investigated whether the amount of visual context shared between the interlocutors in a communicative task would influence the participants' performance with emergent conventions. In the study, participant pairs had to make use of a selection of black-and-white symbols to communicate the correct color out of an array of four colors. The amount of visual context shared between the participants was manipulated by granting access either to all four colors or the correct color only for the participant tasked with communicating. Crucially, symbols had been chosen beforehand to exhibit substantial ambiguity with regard to which colors they could be associated with (making the space, in this context, rather unusual), and participants received neither prior training nor external feedback for the task. In spite of this, the study could show that participant

pairs established conventional meanings, peculiar to their respective dyad, for the abstract symbols and that access to the visual context improved pairs' performance in the task.

We build and expand on this basic design of using colors and black-and-white symbols to study the emergence of semantic structure, which the previous study was not concerned with. Because language exhibits structure on so many different levels, a useful distinction that can be made here is between a *structuring of the signals* and a *structuring of the meanings* (Carr et al., 2017). For the most part, previous studies have focused on the former: This normally takes the form of an unstructured space of signals, which, through repeated interaction and/or transmission to new learners, acquires systematic and conventional rules about their combination and mapping to the meaning space (e.g. Christensen et al., 2016; Kirby et al. 2008, 2015; Nölle et al., 2018; Selten & Warglien, 2007; Winters et al., 2015, 2018; Winters & Morin, 2019). These rules refer to the "grammar" of the artificial language (in a general sense). This first line of experiments can provide us with valuable insights into how linguistic features such as compositional or combinatorial structure can evolve. One example study employing a highly unusual signal space has been conducted by Little et al. (2017), who had participants use an infrared sensor that recorded hand movements and translated them into sounds. Their experiments showed that the dimensionality of the meaning space (manipulated through images varying continuously in size, shade, and/or color) can affect the structure of the emerging signals, in particular when there is a match or mismatch between the two.

Less attention has been devoted to structuring of the meanings, whereby a continuous (possibly even open-ended: Carr et al., 2017) meaning space is discretized into categories via the formation of conventional signals. This refers to the semantic structure and only a few previous experiments (Carr et al., 2017; Perfors & Navarro, 2014; Silvey et al., 2019; Xu et al., 2013) have focused on this precisely. Perfors and Navarro (2014) let participants learn and transmit typed signals for a meaning space of squares that continuously varied in size and darkness. By manipulating this continuous space to show one extreme, abrupt change in either size or darkness, they were able to show that the semantic structure encoded by the signals tended to reflect the structure of the meaning space. A major influence for our study is the research by Carr et al. (2017), who investigated the effects of both transmission and communication on category systems. They presented participants with a vast space of continuously created triangles that participants had to label. This represents a particularly uncategorized meaning space, ensuring that meanings had to become structured from scratch within the task. Their experiments show that communication and transmission combined, but not transmission by itself, will prompt participants to structure the meaning space as well as the signals themselves. Lastly, Silvey et al. (2019) used a similar (but not open-ended) space of continuously morphed pentagons to also investigate the roles of communication and transmission, separately and

combined. They find that communication only increased category structure and alignment when it was also combined with transmission, and in fact that transmission alone eventually would lead to similar benefits.

Experiments conducted in this fashion circumvent one important issue: The resulting structure simply might be a mirror image of the meaning space built in by the experimenter. This can be intended, if the purpose of the task is to demonstrate a dependence on the stimulus set and its arrangement (Little et al., 2017; Nölle et al., 2018; Perfors & Navarro, 2014; Silvey et al., 2015; Winters et al., 2015, 2018), but becomes a hindrance whenever the evolving structure is meant to be interpreted outside of that view. If a meaning space varies on a set of dimensions with clear-cut unique stimuli that need to be distinguished for successful communication (e.g. black cats vs. white cats vs. black dogs vs. white dogs), participants will overwhelmingly encode the same distinction in the structure (one morpheme for black/white and one for cat/dog). By employing continuous meaning spaces, the few studies focusing on the semantic structure (Carr et al., 2017; Perfors & Navarro, 2014; Silvey et al., 2019; Xu et al., 2013) allowed participants to structure the meanings outside of a forced distinction along clear-cut dimensions, thus circumventing this issue.

One issue not currently addressed is how the native language of participants influences the evolution of semantic structure in these experiments. Participants are already native speakers of one or more languages at the start of the experiments, and it remains unclear whether a natural language bias influences the outcomes of the task. Although the issue has been recognized early on (Kirby et al., 2008) and past studies have looked into a native language bias regarding a preference for suffixes over prefixes (Martin & Culbertson, 2020) and noun phrase word order (Martin et al., 2019), to our knowledge no study has systematically set out to address this question regarding semantic structure. The study by Xu et al. (2013) is particularly relevant here: In their artificial language task, participants repeatedly learned and transmitted initially random partitions of color spaces, i.e. subdivisions of a color space which are named with a single color term. Participants were limited to pre-set artificial terms to label the colors, the number of which was fixed within a transmission chain and varied between chains to reflect the number of terms in real-life languages. The authors then compared the partitions at the end of the transmission chains to color term systems found in the World Color Survey (Kay et al., 2009), representing data on over 100 unwritten real-world languages. Quantifying the difference between the two data sets, the results showed that artificial partitions evolved to become close to color term systems found in the World Color Survey. Since all the participants in this experiment were native speakers of English and Xu et al. (2013) wanted to rule out this potential native language bias, they compared the results to a control where an independent sample of participants was explicitly instructed to perform the same task, but to apply the English color term

structure. They found that participants' systems under the instruction to use the English structure were more similar to one another than to the systems created without this instruction. From this, they concluded "that participants did not simply apply English colour categories when classifying colours" (Xu et al., 2013, p.7). While we do not contest this statement, this does not exclude any potential bias towards English structure either; especially in light of the result that the experimental color systems outside of the control condition also moved closer towards English color term structure over time. Are artificial language semantic structures biased towards the ones found in the native language of the studies' participants? This is our first research question: (1) How similar are the semantic structures in native languages and an artificial language?

Color terms and categorical facilitation

The domain of color forms a continuous meaning space, which allows for minimal physical differences between colors, to the extent that they are indistinguishable for the human eye. It is subject to discrete structure in natural language, as the continuous space is carved up by color terms such as "red". Since colors are perceptual phenomena linked to language through color terms (Witzel, 2018), color terms have been the most prominent test case for studies on linguistic categorization, dating back to at least the seminal studies by Brown and Lenneberg (1954) and Berlin and Kay (1969). While neither the debates on linguistic universalism and relativism (e.g. Kay & Regier, 2006; Kay & Kempton, 1984; Terry Regier & Kay, 2009) nor the hierarchy and number of color terms in the world (Berlin & Kay, 1969; Kay & Regier, 2003; Kay et al., 2009) are our concern here, color terms are nevertheless a useful framework for our purposes, i.e. testing native language interference in the emergence of artificial language semantic structure.

One particular phenomenon observed by the research on color terms is that of boundary effects. We can speak of a boundary effect occurring when continuous differences are treated differently across a category boundary as opposed to within the category. This is also known as *categorical perception* (Bornstein, 1987; Harnad, 1987). Studies over the years have observed boundary effects on performance in naming and memory tasks (Roberson et al., 2000, 2005), brain activity as measured by event-related potentials (Thierry et al., 2009), reaction times (Gilbert et al., 2006; Roberson et al., 2008; Winawer et al., 2007; Zhou et al., 2010), and verbal interference (Gilbert et al., 2006; Roberson & Davidoff, 2000). However, the evidence is mixed, with other studies claiming null effects or opposite effects (Brown et al., 2011; Davidoff et al., 2012; Witzel & Gegenfurtner, 2011, 2013; Wright et al., 2015). Witzel (2018) attributed these mixed findings to poor stimulus control in some experiments, and to different levels of processing: Color perception will always involve basic sensory

processing (such as excitation of the cones in the retina), but might, depending on the task, also involve more or less high-level cognitive processes (such as attention or subjective evaluation). Robust effects seem to occur mostly in tasks affording high-level cognitive and directly linguistic processing, such as those involving verbal interference or explicit deliberation on the linguistic categories. This led Witzel (2018) to coin the term *categorical facilitation*, which we adopt in the current study.

One special task that might engage this high-level processing of the color terms that has not seen much attention is referential communication. In particular, intentional communication that involves meta-cognitive processes, such as posited in many frameworks describing human communication (Clark, 1996; Frank & Goodman, 2012; Garrod & Pickering, 2004; Grice, 1989; Scott-Phillips, 2015; Sperber & Wilson, 1996; Tomasello, 2010), is a good candidate for involving the high-level processes mentioned above. For example, given Grice's (1989) maxim of quantity, interlocutors should take into account that they and their partner provide as much information as needed, but not more. Testing the referential communication of colors is difficult when using natural language, since participants already possess color terms, making the task trivial; instead, we require an artificial language game. Artificial language games do not necessarily engage the high-level processes mentioned above, but can be reasonably expected to if they involve interaction between participants and little to no feedback (Müller et al., 2019). Hence our second research question: (2) Does the semantic structure inferred from the native language influence communication with an artificial language?

Lastly, sharing a common language obviously makes communication easier. Relying on shared conventions, interlocutors profit in their interaction, since they are closer to mutual understanding already (Lewis, 1969). Transferring this to signaling systems in artificial language games, it applies to the individual signals (e.g. the meaning of a non-word as "red"), but also in a more general sense to the underlying representations: in our case, the semantic structure (e.g. the information of where "red" ends and "orange" begins). Thus, that if two interlocutors have different underlying semantic structures (e.g. considering a borderline color as "red" that the partner classifies as "orange", even though the general meaning of the terms is mutually understood), they should have a hard time understanding each other. Combining this with native language structure could mean that native speakers of different languages, in which the semantic structures differ, perform worse than pairs that share the same native language when they communicate in an artificial language. This could be the case even though the use of their native languages is blocked. This is what we want to address with our third and last research question: (3) Do mixed-language pairs that show a different semantic structure in their native languages experience more problems in communication?

2. Method

The Color Game

We address these questions in an online smartphone application, the "Color Game", designed to evolve an artificial language through communication between its players. The Color Game was freely available on the Google Play Store and Apple App Store for a runtime of roughly one year. All hypotheses, the exclusion criteria and analysis plan were preregistered before conducting any of the analyses and can be inspected here: https://osf.io/c8nme/. The processed data files and code of the current study can be accessed on the Open Science Framework (https://osf.io/a8bge/). Importantly, the project presented here was part of a larger registration that involved six projects related to the results of the Color Game in total. The application was created with all six projects in mind, and when controls or manipulations concern other projects we will outline them as such. The registration documents and results of the other five projects can be viewed here, as well as an in-detail presentation of the application as a whole, and the app's source code: https://osf.io/9pdzk/files/. The full raw data will be made available after a period of embargo.

Participants

During the runtime of the game, anyone could download it for free and, after a short tutorial, play with another player in one of several game modes. Thus, our approach to the data acquisition for the different projects was not to be limited to a fixed number of participants, but adhere to predefined and preregistered exclusion criteria and thresholds, specific to the relevant project (see the Results section for the resulting breakdown after exclusions). Participants agreed to have their data collected in anonymous format and for research purposes only in a consent form approved at the start of the game. The form and the app itself were approved by the Max Planck Society's ethical committee.

To make the game easily accessible to a wide audience, we offered the choice of 8 different languages for the instructions and menus in-game: English, German, French, Spanish, Portuguese, Russian, Chinese, and Japanese. Note that this did not mean that players with other native languages were prevented from playing the game; the referential communication task worked without requiring any specific native language, since it was based on the symbols and colors only. In fact, a lot of players chose to play the game in English or some other language, even though their native language was different. This native language was the only personal information players were asked to provide for the research, along with their country of origin. In this project, we ended up focusing

on native speakers of English, German, and French only, since the sample sizes for these three languages allowed for robust analyses (see Results section).

Materials

Designed for the color perception aspects of this project in particular, the meaning space was a basic constant of the game, used in every mode and not manipulated. Using the CIE2000 color space (Luo et al., 2001), we constructed a circular selection of 32 colors (Fig. 1). We chose this space because it provides a metric for distance between color hues ("Delta E") that was built to reflect perceptual distance, as opposed to merely physical quantities. The colors are equal in physical luminance and saturation, but show a constant perceptual distance to their two neighbors (Delta E = 7.8). 32 color arrays were formed from this set of colors by picking every fourth color, until a four-colors array was formed, using each of the colors as a starting point once. This way, all colors occur in exactly four arrays. These 32 arrays constituted the communicative contexts for the entire game.

Fig. 1. The game's color space. Each color is given its associated Hex code (as used by the app). Each of the game's 32 colors is drawn from the CIE2000 color space (see Luo et al., 2001), in constant perceptual distance to its left and right neighbors. This includes the first and last color, meaning that the space is circular and can be represented as on the right. Figure adapted with permission from Morin et al. (2020).

The signal space of the game overall comprised 35 black-and-white symbols (Fig. 2). Symbols were initially hand-drawn, then digitized and image processed to make them look cleaner and to standardize them all to the same squared format, taking up equal space on the app display. The symbols were chosen such that they had ambiguous associations with regard to the colors they could be used for, but at the same time allowed the players to solve the communication task above chance

level. To achieve this, we drew on our experience with previous similar setups and in particular a previously published study by three of the authors that we also reused several symbols from (Müller et al., 2019). Müller et al. (2019) had shown that associations for the symbols vary sufficiently, leading to stable conventions that still differ between participant pairs. There were an additional five symbols (bottom row of Fig. 2) which were not used in the actual game, but only for advertising the application and for the in-game tutorial. This was to avoid prior biasing of the meanings for any of the symbols actually used within the game. Players initially only started with ten random symbols from the signal space that they could use as a sender, but eventually unlocked all of the symbols as they progressed in the game (for details, see Procedure).

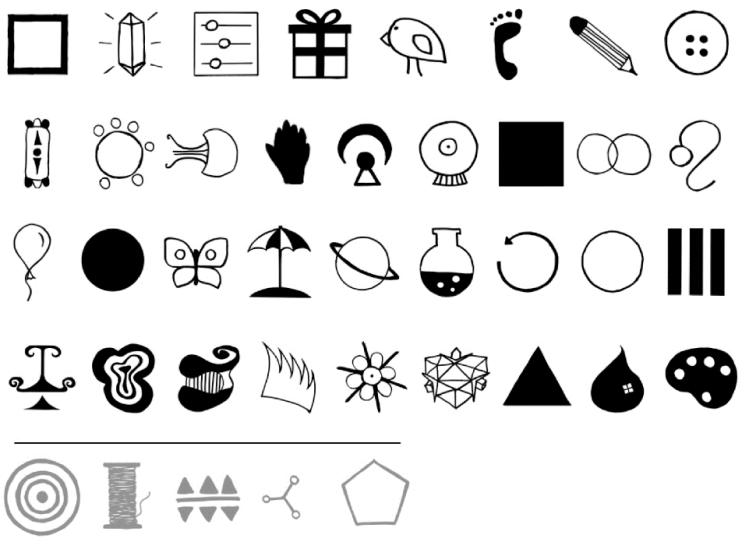

Fig. 2. The 35 symbols used in the game (first four rows). Bottom row, in grey: the five symbols used for the tutorial and for advertising the game. Players were given a random set of 10 symbols at the start and could unlock the full set of 35 symbols by successfully playing the game. Figure adapted with permission from Morin et al. (2020).

Procedure

Regardless of the game mode, the game always brought together two participants to play a referential communication task: One of the players (the sender) was tasked to communicate a color, using the black-and-white symbols, to the other player (the receiver), who then had the pick this color out of an array of four colors. Fig. 3 shows an example trial and the view from both sides. On every trial, participants were shown a randomly chosen array from the space of 32 communicative contexts, and the four colors were arranged in random positions in a 2x2 grid for each participant (see Fig. 3). On the sender's side, one of the colors was randomly chosen and marked as a "target color" by a transparent dot. This left three remaining colors of the array in the role of *distractors* for

the current trial, since they were incorrect responses for the receiver. Here, a manipulation was implemented, which was relevant for a different project (https://osf.io/qz597/): Senders did not always have access to all four colors like the receiver, but randomly saw one (the target) to four of them (the full array). We control for this randomized variable in our models, where necessary.

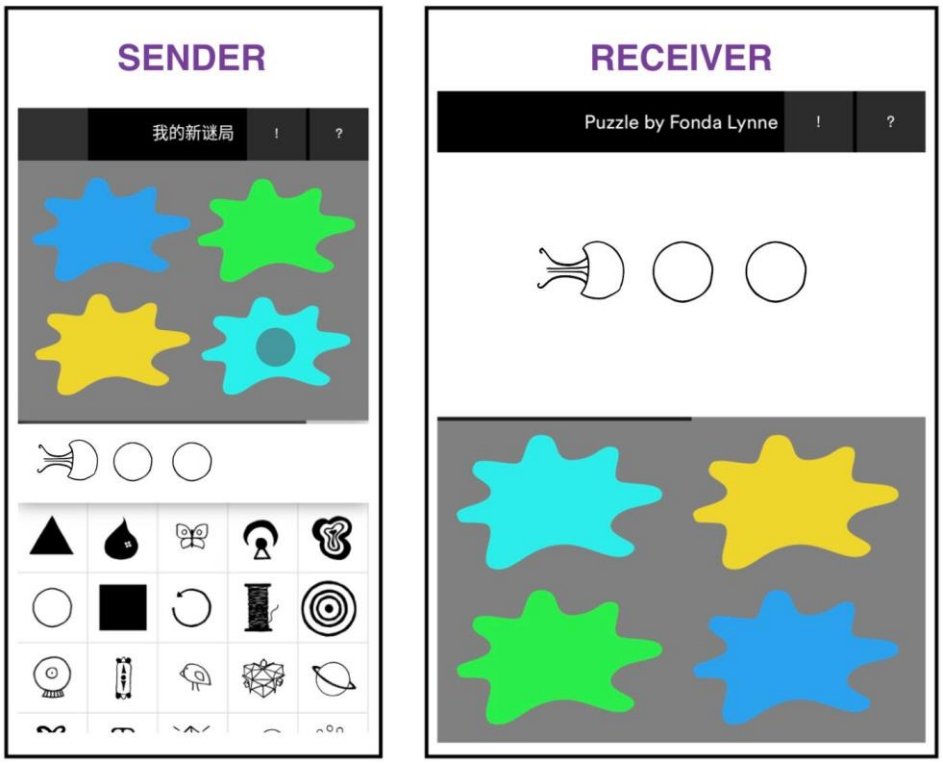

Fig. 3. An example trial in synchronous mode. The sender in the current trial communicates with a receiver to help find the target color (here, the brighter shade of blue), marked for the sender by a dot. Figure adapted with permission from Morin et al. (2020).

Participants always played the game in sets of ten trials. In each trial, the sender could choose up to ten symbols (including reduplications) from their current symbol repository to guide the receiver towards choosing the correct color from the array. When playing as a sender, symbols were displayed at the bottom of the screen. Tapping a symbol, like on a keyboard, meant it would be displayed in the white row (between the set of symbols and the set of colors). Senders could also remove previously tapped symbols from this row by tapping on them. Receivers were sent the row of symbols, in the original order, and could choose a color from the array by simply tapping it. From our experience in past laboratory experiments, players typically solve this task by forming meaning conventions on single symbols that stand for a single basic color term or modifiers such as “dark” or “light” (Müller et al., 2019).

The game never provided feedback about player performance, apart from a general statement announcing how many trials out of the total of ten trials the pair solved correctly (but not which ones), which was displayed at the end of each set of 10 trials. Our reason to avoid trial-by-trial feedback is that it would let receivers know instantly which symbol their sender associates with which color, allowing them to learn a sender's code by mere association. After completion of a set, both players were awarded points, depending on their performance.

Progress in the game for the players occurred in the following way: Upon entering the game for the first time, players had to complete a short tutorial explaining the basics of the game. This involved acting on one trial each as a sender and as a receiver, using the tutorial symbols only, and with the same predefined trial for every player. Having completed the tutorial (and every time they entered the game after that), players entered the main lobby screen. Players had the free decision to enter a game with another player of their choice, and also to do so repeatedly and whenever they wanted, a core difference between our study and classical experiments that makes for more realistic evolutionary dynamics (Morin et al., 2018). When players earned points, they unlocked new symbols and game modes. For the symbols, participants started with a basic set of ten randomly chosen symbols to use as a sender and gradually unlocked new ones every time they upgraded their rank. This was done to create a less taxing start than with a full selection of 35 symbols, to preserve more diversity in symbol-color mappings, and to motivate the players. For a recurring player, which is the only player base we can make use of in the scientific projects, these symbols could be unlocked rapidly.

All players started with access to basic "synchronous" and "asynchronous" play. Asynchronous play meant that the player could choose to start a set of trials at any time, upon which they played the 10 subsequent color arrays as a sender. These trials were created by the sender in isolation and solved later by an interested receiver. Senders could decide to send out these asynchronous sets of trials to a receiver of their choice, or to release them publicly. If they did the latter, a number would appear next to their pseudonym for other players indicating the number of sets that the player had created, and anyone could open these sets and try to solve them as a receiver. However, they would have to pay a small amount of points to do so, only increasing their score if they were reasonably successful. Meanwhile, creating sets as a sender was always free. Once a set of trials had been played by any receiver, it was removed from the public lobby for everyone. In contrast, synchronous interactions could happen by inviting another player directly into a set of trials, as the game also showed who was currently available. Synchronous invitations never cost any points. In this mode, players communicated in live interaction and the receiver could see changes to the sender's communication channel in real time. Additionally, both players could also interact by sending the signs "!" and "?" to

one another, the meaning of which was intentionally left open. This was to make the task more interactive and make players relate to the partner, to separate this live communication from the asynchronous situation.

When players had collected a large amount of points in the game, they also unlocked a new "speed mode" to play either asynchronous or synchronous trials in. Here, the rules were the same as before, but time was severely limited. This mode was included to present a constant challenge also for very experienced players and rewarded many points when successful, but also cost a lot of points to enter. The different game modes were included to give more variability to players other than simply the basic task, to assess the importance of live interaction, and to allow for content that players could access anytime, regardless of who else is online. As such, we are not interested in the differences between these modes, but game mode is controlled for in our analyses.

Online survey

We need baseline data to find the semantic structure that speakers of the three languages use for our (unique) set of colors, using these languages' basic color terms. For this purpose, we set up an additional online study, somewhat similar to the method of the World Color Survey (Kay et al., 2009). This online survey was separate and independent from the Color Game and created for the needs of the present study.

Participants

The survey was set up online on Prolific, a platform enabling researchers to recruit participants all over the world in exchange for payment. This study and the Color Game thus used a different sample of subjects. For each of the three languages that showed robust sample sizes in the Color Game data (English, German, and French), we got survey data for 50 individuals. Only native speakers of the specific language were able to access the respective online survey. Participants were paid £1 in exchange for their help and time (about 5 minutes), and gave their informed consent before starting the survey.

Materials

The task made use of the same color space as the Color Game. We operationalize the native language semantic structure in the form of basic color terms much like Berlin and Kay (1969), an approach that has proven useful for assessing the naming patterns in different languages worldwide (Kay et al., 2009). The goal here is to identify linguistic categories that allow us to study the semantic structure of colors (e.g. for English, "red", "green", "blue", and so on). With a set of basic criteria, we can use a finite number of language-specific terms that allow for a full description of our color space. Moreover, by minimizing the number of color terms and avoiding overly specific descriptors (like "crimson" or "steel-blue"), we can more readily find commonalities in the underlying structure of the color space of participants while opening up the possibility of making robust cross-linguistic comparisons. Examples of other approaches that, like us, have gone beyond observation of the distribution of naming patterns and successfully built on the concept of basic color terms include agent-based simulations of the emergence of the patterns (Baronchelli et al., 2010; Steels & Belpaeme, 2005) and experiments with human participants to recreate the known real-world patterns (Boster, 1986; Xu et al., 2013).

The basic color terms that were used for the three languages had been determined in advance by piloting with a small group of native speakers of the respective languages. This was done by freely eliciting the first color terms that came to mind, and then letting participants map the terms they named onto our space of 32 colors to rule out terms that were not applicable (like "black" or "brown"). The most frequently named color terms were compared and confirmed to correspond to well-established basic color terms for English and their respective equivalences in German and French, with one additional term for German, and adopted as displayed in Figure 4. English and French terms turned out to be very similar, both ending up with seven basic terms that applied to our space (note that achromatic terms, i.e. "black", "white", and "grey", do not apply because we only vary hue in the space; and "brown" does not apply because lightness is too high in the space). In contrast, for German our piloting revealed that an eighth "türkis" term should be added, specifically referring to colors in the blue-green spectrum.

	Color Terms							
English	Blue	Red	Yellow	Green	Purple	Pink	Orange	
German	Blau	Rot	Gelb	Grün	Lila	Pink	Orange	Türkis
French	Bleu	Rouge	Jaune	Vert	Violet	Rose	Orange	

Figure 4. Color terms used for the online survey in the three languages.

Procedure

After participants had given consent, they read the instructions to the task before proceeding. The instructions and the entire task were presented to participants in their respective native language. In the task, the 32 colors of the Color Game's color space were presented to the participant, all at the same time, organized in a circular pattern to avoid effects of position and start/end points (see Fig. 5). The participants must then provide a label for every color by associating it with at last one basic color term. This was done by presenting the respective terms, one by one and in random order, and asking participants to click on all colors in the circle that they associated with that term. Participants' selected colors appeared at the top of the screen and could be removed from the selection by clicking on them again. When participants were finished with selecting all associated colors for a specific term, they could click "next" and would be presented with the next randomly chosen color term from their language. Participants continued with labeling colors until all colors had been named; thus, if after a complete cycle of all terms some colors were not named yet, these colors were presented for all terms in that language again. Colors could also be named with more than one term, if they were independently chosen for different terms within one cycle of the naming procedure; in fact, this was crucial to account for boundary cases and to mirror the potential structure of the communication in the Color Game, since different symbols could get used for the same colors.

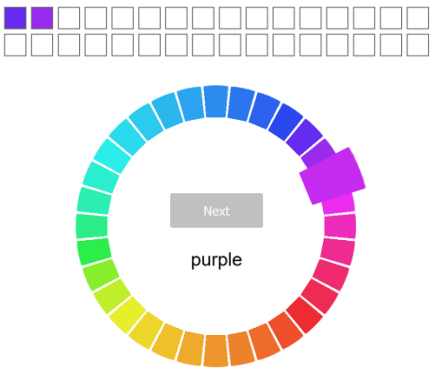

Fig. 5. The color wheel used in the online survey to gather baseline data for the semantic structure in native language color terms. The example shows a participant in the English sample tasked with selecting the colors associated with the term "purple". Colors were highlighted when moused over (like the one on the right side) and appeared at the top when clicked on, and their position on the screen was randomized while the circular order was maintained between participants.

Predictions

Bringing together the data on the native language semantic structure and the artificial language communication in the game, we can address our research questions outlined in the beginning. We do this by applying exploratory factor analysis, a method used to summarize data by reducing its variation to a smaller set of factors that reveal the underlying structure, to the data from the online survey. After that, we try to confirm the structure found in the exploratory factor analysis on the data from the artificial language game by applying a confirmatory factor analysis whose parameters are set to the structure resulting from our baseline. For clarity, from here on we will use the term "color categories" or simply "categories" to refer to these statistical factors. Based on the research questions outlined in the introduction, we made the following predictions:

Research question 1: How similar are the semantic structures in native language and an artificial language?

Prediction 1. We predicted that the categorical structure (assessed by exploratory factor analysis) of the native language should fit the structure of the artificial language. At the very least, a confirmatory model that imposes the native language structure on the artificial language should not get rejected. This would indicate that the artificial language structure reflects the native language structure, to some degree.

Research question 2: Does the semantic structure inferred from the native language influence communication with an artificial language?

Prediction 2.1. If categorical facilitation is at play for communication in the game, we would expect color arrays that cross more boundaries between categories to be easier to solve for participant pairs (of the same native language) as compared to color arrays that cross fewer of these boundaries. This is because, if the color array covers more natural color categories, each of the colors within the array should be more nameable.

Prediction 2.2. Taking the target color of a current trial into account, we also predicted that pairs would send more symbols when a distractor was present that was part of the same category in the native language structure: Presumably, they would realize that a simple symbol representing a meaning such as "blue" would not suffice in this case, and add modifiers, e.g. "dark blue". This measure complements the simple frequency of boundaries in Prediction 2.1 by focusing directly on the relevant pragmatic contrast that needs to be expressed by the sender in the communicative situation.

Prediction 2.3. Regarding the effects of these same-category distractors on communicative performance, we put forward and preregistered three alternative predictions, supported by different researchers in the project. The first is that player pairs should be more likely to succeed when the target color is accompanied by one or more same-category distractors, because the use of modifiers could help to identify the target color more precisely by making a pragmatic inference (if the sender uses a modifier, e.g. "darker", then the target must be amongst the same-category colors, and be the darker one). The alternative prediction to this is that player pairs should be less likely to succeed under the same circumstances, because colors within the same category should be harder to distinguish than colors across boundaries, and more symbols mean misunderstandings could arise more easily. A third possibility is that the effect of same-category distractors could be dependent on the experience of a pair (i.e. an interaction): With an increasing number of trials between the participants, we could observe a change in the effect for same-category distractors from less success to more success; thus, participants' performance would first suffer due to colors being harder to distinguish and more misunderstandings because of the higher number of symbols, but profit from the pragmatic specificity later on, leading to higher success.

Research question 3: Do mixed-language pairs that show a different semantic structure in their native languages experience more problems in communication?

Prediction 3. Here, the prediction is that the coordination problems arising from different native language structures would lead to worse performance in pairs not sharing a semantic structure, compared to those that do, but only for items for which their languages do not align.

3. Results

The following data resulted from the Color Game's runtime from May 2018 to April 2019: Overall, a total number of 4,277 users accessed the game, providing us with 435,842 trials of raw data. After applying the general exclusion criteria for the data (common to all projects on the Color Game for reasons like bugs, empty trials, or users that did not provide their mother tongue; see "CleanUp" in the online data), we are left with 347,606 trials by 2,615 users. Most relevant for our main analyses in this project is the number of senders and same-language pairs sharing a specific mother tongue, as per the preregistration. In principle, we preregistered that we only wanted to analyze data from senders and pairs that were sufficiently involved in the game, since symbol-color mappings of infrequent players might be too noisy, inaccurate and not as exhaustive with regard to the coverage of the symbol and color spaces. As such, we set and kept to the fixed cutoff of at least 100 trials

played in the game for individual senders, and at least 50 trials played together for individual pairs (regardless of the distribution of role in the pair as sender or receiver). It is important to note that these cutoffs apply to different analyses (concerning prediction 1 and predictions 2 to 3, respectively). The number of users and trials in the specific subsets of relevant languages (from the preregistration) can be seen in Table 1.

Language	senders > 100 trials		pairs > 50 trials	
	n	Trial count	n	Trial count
English	85	57,086	101	13,234
German	88	122,116	116	37,070
French	53	27,226	44	3,981
Spanish	12	4,437	0	0
Chinese	0	0	0	0

Table 1. Number of senders and same-language pairs that reached the preregistered thresholds of trials for each of the 5 languages we intended to use.

Although we did not reach the cutoffs preregistered for an early analysis in most cases, there is still enough data to perform the planned analyses fully for English, German, and French. Spanish and Chinese had to be dropped due to the lack of data of players with a high number of trials; since it was impossible to fully anticipate the amount of users that the app would attract, sampling issues for some languages were expected, however.

Since we need it for the baseline on the color term structure in the respective languages, we start by summarizing the results of the online survey. The resulting data was restructured to represent one row for each participant-term pair, i.e. each unique combination of participants and color terms (e.g. participant 1/red, participant 2/red, participant 1/green, etc.; compare also Jäger, 2012). We can think of this structure as a "profile" for each participant, recording the number of times that each of the 32 colors were associated with a specific term in their language. Then the exploratory factor analysis (*EFA* from here on) was applied using R version 3.5.1 (R Core Team, 2018). EFA is an exploratory, i.e. data-driven, clustering technique that can be used to reduce datasets that show high dimensionality to a pre-set lower number of dimensions, called "factors", while describing the data as well as possible. By structuring the data in the way described above, we can summarize the categorization inherent in individual color terms and participants, all at once. The number of factors

that should be extracted in an EFA can be hard to decide on if there are no pre-defined conceptions of the dimensions; however, in our case we could simply set it to the number of basic color terms determined in the piloting study (see Method section) and used throughout the survey. The analysis also reveals the factor loadings, i.e. the strength of association between each variable and the common factors; in our case, these correspond to the association between colors and their respective terms. Ideally, colors should show high loadings on their common term, but low cross-loadings to other terms. We consider loadings higher than |0.3| to be meaningful. A common approach with EFA is to rotate the factors to reach a simplified orientation to facilitate interpretation; here, we chose promax rotation, as the factors were expected to correlate. The goal of these analyses was to see whether the data would reflect our assumptions concerning the structure, resulting in clean categories that represent the native language terms and show low cross-loadings, with clear boundaries which then could be used for the further analyses.

The categorical structure of the data from the naming task for the three languages is visualized in Fig. 6. In English and German, it reflects the native language terms that were given to participants, with low cross-loadings and clear-cut (i.e. overlap only on single border colors in all but one case) contiguous categories. It is important to emphasize that emerging categories are not pre-defined in the analysis, but easily identifiable as the respective color terms, confirming our approach. Another positive result is that we observe the highest loadings on colors that are in the center of categories rather than at the boundaries. The differences between these two languages mostly boil down to the light blue area in the color space, which is statistically explained by the "türkis" term in German but subsumed into the general "blue" term in English, with one specific color in this spectrum even not loading highly on any factor in English. This reflection of our assumptions in the results is not trivial, as can be seen in the French data: Here, we find that the 7-factor EFA results in an unexpected lack of a "red" category in favor of a "light blue" category (such as in German). Still, cross-loadings are also low and boundaries clear-cut. The "red" category does appear in the structure when an 8-factor solution is proposed instead. The results for this 8-factor structure only differ meaningfully from the data-driven suggestion with the 7 factors for one single model, which we flag in the analyses below; otherwise, the results for the 7-factor structure are reported.

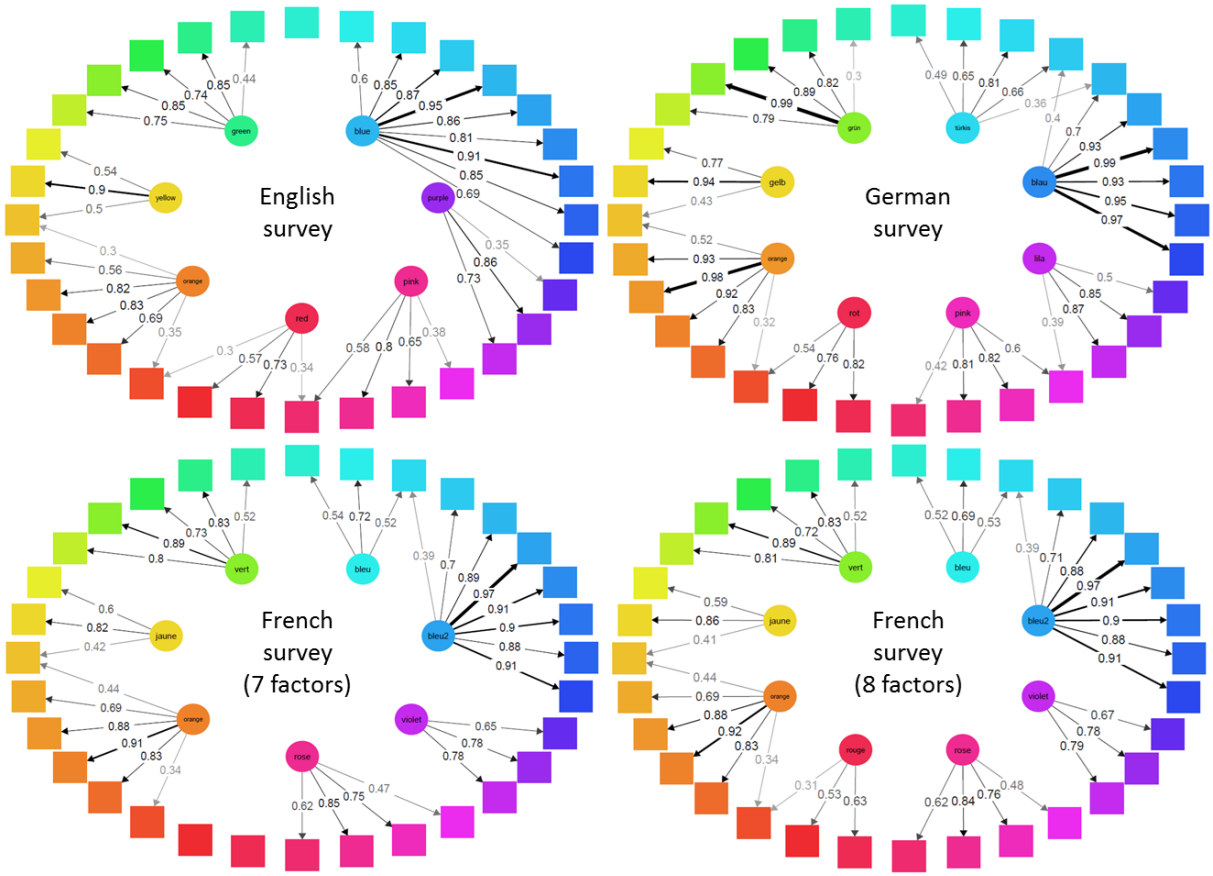

Figure 6. Visualizations of the categorical structure resulting from the EFA on the data of the online survey. Boxes represent the 32 colors used in our study. Circles represent categories, named with the term that was most frequently associated with the colors loading on them; these categories are colored in the hue that has the highest loading. Arrows are drawn for all factor loadings (= the numbers on the arrows) in the EFA that are .3 or higher. **Top left:** English data. **Top right:** German data. **Bottom left:** French data, 7-factor-structure. **Bottom right:** French data, 8-factor-structure.

Prediction 1

We then proceeded with confirmatory factor analyses (*CFA* from here), trying to replicate the structures found in the EFA by directly fitting it on the communicative data from the Color Game. CFA is the complementary approach to EFA, used to confirm a pre-defined structure by fitting it in a structural equation model. As such, it is able to provide a test of our native language structures' fit on the artificial language communication data. We subset the cleaned-up Color Game data to all trials from senders of the three languages that had played at least 100 trials in that role. This data was arranged so that each row represented a unique pairing of a symbol and sender, recording, for each color and each symbol, whether the symbol was used to indicate the color (thus mirroring the structure used for the EFA). We then fit the CFA once for each language using *lavaan* (Rosseel, 2012).

The structures specified for the models came from the results of the EFA, represented by loading each of the colors that were paired together onto a common factor (only considering loadings greater than .3). Model fit was assessed both by robust CFI and RMSEA estimates, two common measures of fit. As can be seen in Table 2, the CFA was at least a moderate fit for all languages (see the guidelines in Hooper et al., 2008), with values that are moderate to good for German and English. Descriptively, colors with high loadings in the CFA also correspond to the colors with high loadings in the EFA (i.e. typically the central colors of a category), and boundaries from the EFA are not contradicted by the loadings of the CFA.

Language	n	CFI	RMSEA (95% CI)	Model Fit
English	2056	.88	.063 (.061-.065)	good to moderate
German	2096	.92	.054 (.052-.056)	good to moderate
French	1123	.86	.077 (.074-.080)	moderate

Table 2. Goodness of fit measures (robust estimates) for the CFA on the three languages.

Prediction 2.1

Next, we tested whether colors could be communicated more accurately when there were more boundaries present in a given array for a given language. Again, the results of the EFA were taken as a baseline, and by assigning each color to the category it had the highest loading on, each color array in the game could be described in terms of how many boundaries were present for each of the three languages (from the view of the full array, regardless of how many colors were shown to the sender). In case of single colors not loading highly on any factor (only relevant for one color in English and three in French), a full transition between the two categories bordering these colors was needed within the arrays to count the array as exhibiting an additional boundary. Coding the 32 color arrays in this way for each of the languages revealed that German exhibited much less variation in the number of boundaries than English due to its 8 categories, being limited to arrays that crossed either two or three boundaries (resulting mean number of boundaries in German: $M = 2.63$; standard deviation $SD = 0.49$; English: $M = 2.25$; $SD = 0.76$). In French, this depended on whether the 7-factor or the 8-factor structure was used (7-factor: $M = 2.41$; $SD = 0.67$; 8-factor: $M = 2.44$; $SD = 0.62$), but was overall still more varied than for German.

The analyses were performed on subsets of the data limited to pairs of the same native language that had played at least 50 trials together. We used separate logistic mixed-effects models (lme4

package in R; Bates et al., 2015) for each language to test the effect of the number of boundaries on accuracy on a trial-by-trial level, controlling for the fixed effects of the number of trials a pair had played together and the number of colors in the senders' array. Random intercepts were added for participant pairs, color arrays, and the game mode. Random slopes were added for the number of boundaries and then reduced in a stepwise approach until a model could converge with acceptable correlations in the random-effects structure. We then compared the AIC (as an indicator of model quality) of these final models to the AIC of a simpler model that was identical but had the fixed effect of the number of boundaries removed, respectively. This simpler model should show an increased AIC value to support our hypothesis. As a guideline, we consider a ΔAIC of 2 or greater to be meaningful evidence of a better performing model (Burnham & Anderson, 2004); at the same time, we also report the results of likelihood-ratio tests to see how robust this analytic strategy is. This procedure will be repeated in a similar way for all upcoming analyses. The results for these models can be seen in Table 3. For the English data, the number of boundaries in the arrays had a positive effect on performance. This means that English pairs were better when the colors in a given trial loaded on more different categories (according to English color terms). For German and French, no such effect could be detected.

Language	n	AIC simple model	AIC model including #boundaries	ΔAIC	Likelihood-ratio tests: p-value
English	13234	15174.9	15171.6	3.3	.022*
German	37070	43074.3	43075.3	negative	.326
French	3981	5193.2	5193.9	negative	.250

Table 3. AIC values and p-values of the likelihood-ratio tests computed from the models testing the effect of the number of boundaries on performance. "Negative" in the ΔAIC column implies the simpler model minimized the AIC and was favored over the model including the number of boundaries.

Prediction 2.2

We then turned to the analyses investigating whether same-category distractors would impact performance, given the native language structure. For this, we had to restrict the data from the previous models only to trials that showed all 4 colors to the sender (roughly 25% of the data in the analyses for prediction 2.1), since otherwise the sender would not necessarily be aware of the presence of same-category colors. We then coded each trial for whether the color array presented the sender with a distractor that was part of the same category as the target, given their native

language structure (the method for this coding being similar to how the previous analyses were handled): the "distractor" variable. We predicted the number of symbols sent in the given trial by this variable, ignoring reduplications of the same symbol. We did this in separate linear mixed-effects models for each language that again controlled for the fixed effect of the number of trials a pair had played together and the random intercepts of participant pairs, color arrays, and the game mode played in. Random slopes were added and reduced similarly to the previous analyses.

The results for these models can be seen in Table 4. We again found the expected results for the English data, as suggested by the difference in the AIC between the two models. As such, English senders sent more symbols when a same-category distractor was present in a given trial. For German and French, no such difference could be detected; additionally, the direction of the estimate of the effect pointed into the opposite direction. This analysis also is the only case in which the French structure with 8 factors differed meaningfully from the 7-factor structure: Here, the difference between the models became significant, meaning that, only with 8 factors in the structure, French senders sent less symbols when a same-category distractor was present in a given trial.

Language	n	AIC simple model	AIC model including "distractor" variable	ΔAIC	Likelihood-ratio tests: p-value	Direction of effect
English	3323	8503.8	8499.5	4.3	.012*	positive
German	9356	25717.8	25717.8	0	.154	negative
French	995	2750.4	2749.3	1.1	.082	negative
French, 8 factors	995	2742.3	2739.9	2.4	.035*	negative

Table 4. AIC values and p-values of the likelihood-ratio tests computed from the models testing the effect of the presence of a same-category distractor on the number of symbols sent in the given trial.

Prediction 2.3

After that, we also predicted performance by the presence of same-category distractors in three separate models fit to the data used for prediction 2.2. These were logistic mixed-effects models with the same controls as before. Additionally, we included the interaction between the presence of same-category distractors and the trial experience of pairs in another set of models and tested these

against the models including the main effects only. Again, there was a difference between the simplest model and the model including the distractor variable for English, but it was close to non-significance (ΔAIC of 1.9 and p-value of .049; see Table 5). This means that English pairs also tended to perform worse when a same-category distractor was present in a trial, but this result is less robust. This supports option 2 from our different predictions, implying that English pairs potentially found colors that belonged to the same native language category harder to communicate. No such effect could be found for German or French, nor for the interaction effect in any of the three languages.

Language	n	AIC simple model	AIC model including "distractor" variable	AIC model including interaction	ΔAIC	Likelihood-ratio tests: p-value
English	3323	3881.8	3879.9	3881.4	1.9; negative	.049*; .458
German	9356	10905.9	10907.6	10909.6	negative; negative	.602; .985
French	995	1307.5	1309.4	1310.7	negative; negative	.725; .408

Table 5. AIC values and p-values of the likelihood-ratio tests computed from the models testing the effect of the presence of a same-category distractor on performance in a given trial. Note the additional column for the AIC including the interaction effect between the distractor variable and the number of trials a pair had played together, and thus the two values for the ΔAIC and likelihood-ratio test columns: The first number indicates the value for the comparison between the simplest model and the model including the main effect, and the second the value for the comparison between the model including the main effect and the model including the interaction effect.

Prediction 3

Lastly, we investigated the performance of mixed-language pairs of speakers of German and English more closely. We focus on these two languages specifically because they have shown an interesting contrast in their exploratory structure for four colors in the blue-green spectrum, and because their structures and analyses have shown clearer results than French so far. Hence, we created a variable coding whether a trial in the data was conducted by a same-language pair or by a mixed-language pair (no matter who was sender or receiver). We tested for an effect of this variable in a model comparison including random effects of participant pairs, color arrays, and game mode. There was

no meaningful difference between the model including the same/different-language variable and the one without (Table 6). After this, we subset the data for an additional analysis concerning the same effect for the four colors mentioned above only. Similar to the results of the first model, there was no difference for the same/different-language variable.

Data subset	n	AIC simple model	AIC model including effect	ΔAIC	Likelihood-ratio tests
Same vs. different language	75252	89126.1	89126.2	negative	.166
4 colors in the blue-green spectrum only	9435	10724.3	10726.2	negative	.782

Table 6. AIC values and p-values of the likelihood-ratio tests computed from the models testing the effect of same-language vs. different-language pairs on performance in the given trial.

4. Discussion

By combining the results of the online survey with data collected in the artificial language game, this study was able to compare the Color Game's referential conventions to the respective semantic structures of English, German, and French, and found a good to moderate correspondence. Our findings provide evidence in favor of similarities between the semantic structures present in the native language of participants and their evolved structures in the artificial language game. One important point to note is that the samples of the online survey and the application were independent. As a consequence, we do not follow individuals' tendencies to apply their personal semantic structure to artificial language structure, but generalize to the average behavior of a native language community instead.

Over the three languages, our EFA of the data gathered in the online survey revealed structures consistent with our expectations based on the concept of basic color terms and on our piloting before the study. First, category boundaries were clear-cut, which implies that participants divided the space by applying mutually exclusive color terms. This is very much in line with the idea of the basicness of color terms proposed by Berlin and Kay (1969) and their criterion that the basic categories should not be included in any other color category. Second, the color categories resulting from the EFA were also maximally contiguous, with no interruptions within the arrangement of the space. This confirms the validity of our approach to create the color space in a circle of hue while

keeping lightness and saturation constant. Third, colors that were located more centrally within a category showed very high loadings overall, whereas peripheral colors showed the lowest loadings. This is in agreement with central colors being prototypes within their category (Berlin & Kay, 1969).

Between the three languages, there was one peculiar case with mixed results that came to attention during the EFA, and it concerned the naming of the colors in the blue-green spectrum. English speakers tended to name one particular color on this boundary as neither "green" nor "blue". German speakers applied their eighth term, "türkis", exclusively in this area, and thus filled the gap that could be seen in the English data. This supports our decision to work with eight German terms after the piloting, and chimes in with research suggesting the growing basicness of the term "türkis" in German (e.g. Zimmer, 1982; Zollinger, 1984). The results for the French speakers were weaker and unexpected, with a category for the blue-green colors instead of a "rouge" category, even though they had not been offered a term for blue-green. While the addition of an eighth factor to the EFA remedied this issue, it is still puzzling, but also shows that this data-driven approach to the semantic structure of the languages was by no means a guaranteed way to arrive at the results we had expected. Overall, we believe that the reason for the peculiarities surrounding this exact part of the space lies in the large number of colors that could be classified, in English terms, as either "blue" or "green". Even though we created the color space such that neighboring colors were equidistant, this turned out to be one characteristic feature. French speakers, then, ostensibly preferred distinguishing between light and dark blue first (rather than orange and red) in the online survey, which is understandable given the high number of colors there (in contrast to the low number of red colors). This touches upon a related point regarding what an "optimal" categorical system to organize our color space should theoretically look like (see Regier et al., 2007): Given the imbalance between the "red" and "blue" regions, a theoretically optimal system would have to choose to include more colors into the "red" category and less colors into the "blue" category than we find in English and German, such that communicating any color can be achieved with comparable efficiency. Instead, we find that the participants in the task fall back on, in this sense, "suboptimal" categories close to their native language semantic structure, a central result of the current study.

An important point is that while our results on the CFA speak for similarities between native and artificial language semantic structure, they do not necessarily imply a direct causal link. It might be tempting to argue that native language structure should have caused participants to apply similar structuring in creating the artificial language; however, an alternative explanation could be that a common factor is underlying and causing the structure both in native and artificial languages. This is one reason for the emphasis in the outline of the paper that we are neither providing support for theories claiming relativity nor those arguing for universalism among color terms. Instead, we argue

for a general cognitive link between the native and artificial language structures, but do not make claims as to where it might come from. A methodological caveat for artificial language studies, then, is to keep this potential confound in mind: When dealing with colors (but possibly with other meaning spaces as well), participants might not create novel structure spontaneously in the task, but rather recreate ones they know from their native language (coming from a relativist stance) or have a general preference for (coming from a universalist stance). Future work would be in a good position to dive deeper into this question, and could in particular concentrate more on the contrast between two specific languages that differ substantially in their native language structures, a sample that our smartphone application could not aim specifically for: We were limited to working with three closely related Indo-European languages that mostly overlapped in their semantic structure. If it turns out that participants in such a sample create artificial language structures that fit well on their own native language, but not on the contrasting language, this would imply that potential biases are language-specific.

Answering research question 2, we also looked more closely at the importance of semantic structure for the communicative performance of pairs of the same language in the game. More precisely, we found that English participant pairs communicated more successfully when color arrays crossed more boundaries in their native language semantic structure. They also communicated less successfully and sent more symbols when an array contained a distractor that belonged to the same color category as the target. Regarding the alternative predictions we put forward to answer this question, our data suggests that participants did not profit from additional pragmatic information; nor does the more complex hypothesis involving pairs' experience with the game seem accurate. Instead, given the results, we favor the explanation that same-category distractors are harder for participants to delineate clearly in communication than distractors from a different color category. This also explains the need for a higher number of symbols in the relevant trials. Overall, the results are the first evidence for categorical facilitation within a communication task, and for artificial language performance and pragmatics being influenced by pre-existing semantic structure. This expands our study from merely observing similarities between native and artificial language structure (research question 1) to finding concrete behavioral impact of the shared semantic structure. As outlined in the introduction, we believe that communication as a rather involved and explicit task engages linguistic processing of the structure, which in turn facilitates communicating different-category colors in the game.

However, we were not able to observe the same effects that we found for English speakers for either German or French. We believe it is unlikely that the impact of the semantic structure would be specific to English speakers only, especially since we found overlap in the native and artificial

structures of German and French that was close to the one for English. Instead, our suggestion is that the most likely explanation for the null effects lies in the stimuli and their performance for German and French: For German, applying the eight basic color terms to the space meant that color arrays in the game never showed less than two boundaries for the language, limiting variation in the statistical analysis. For French, the EFA with seven color terms did not mirror our expectations, leading to an unplanned alternative version with eight terms that suffered from problems similar to the German analysis. For this reason, we also do not put any weight on the result that for the 8-factor structure, French senders sent less symbols when a same-category distractor was present; especially since it was unexpected. The conclusion, then, is that stimuli have to be carefully selected with regard to the structure that the respective languages are going to be tested on. Again, future studies with similar aims to ours could profit from explicitly focusing on specific contrasts between two or more languages, designing stimuli in a way that allows all tested languages to vary in the crucial conditions to a reasonable degree.

Regarding our last research question, we did not find differences in performance between mixed pairs of English and German speakers and same-language pairs, neither for the overall color space nor for the specific set of colors in the blue-green area that we were interested in. For the complete color space, overlap between the two languages was great, so we did not expect any difference. That native language did not make much of a difference for the four targeted colors was more surprising, however. Presumably, the distinction for an eighth category specifically describing these colors that was found for German was straightforward to incorporate for English speakers as well; again, given our distribution of color in the space with a heavy reliance on the blue area, players might have become well aware of the need to distinguish these colors more clearly as they continued playing. This is also demonstrated by the sample of French speakers in the survey that tended to create this light blue category over the expected red category. One positive conclusion we can draw from the results on mixed-language pairs is that performance in the artificial language task apparently was rather independent from the native language of one's partner, suggesting participants were able to adjust to their partners flexibly.

Concerning the choice of a smartphone application as a means to operate our artificial language game, there were several advantages and limitations. We cannot, for instance, be certain that all of our participants had normal color vision or that some smartphone screens had not been calibrated in ways that significantly bias color perception (although we did warn about that when players downloaded the game). Another issue stems from the size of the overall project, involving not only this study but six different preregistered projects overall; this led to some design choices (e.g. manipulating the senders' color arrays or the different game modes) that were not relevant for the

current study and had to be controlled for statistically. We cannot tell whether a direct experiment only dedicated towards addressing our questions could have revealed clearer results in the cases of German and French, and concerning research question 3.

Hopefully, some of these concerns were compensated for by the sheer scale of the project: It is rare to have the opportunity to analyze experimental data that, in the most extreme example, includes over 100,000 trials of native speakers of German. Even if the numbers on other languages and same-language pairs were lower than this and we could not obtain enough data on Spanish and Chinese, the separate data sets used in our final study included several thousand observations, respectively. Even if a conventional online study invested the time and effort to reach a similar scale in sheer numbers, there are still advantages to the approach we have taken here with the smartphone application. In particular, we believe the application allowed us to create more realistic interaction and transmission dynamics (Morin et al., 2018). For the interactions between participants, there is free partner choice: Instead of being forced to interact with the one same person over the course of an entire experiment, participants could decide to switch partners as much as they want, for instance if they could not reach a common convention. For the individual player, there was also the choice of when and for how long they wanted to access the game, as opposed to the fixed and rigid application of trials typical of regular artificial language experiments.

5. Conclusion

In this study, we investigated the semantic structure of an artificial language that participants evolved by communicating colors, and compared it to the respective native language structure of speakers of English, German, and French. To do so, we combined the results of a large-scale online smartphone application with a separate online survey, whilst building on previous work on color terms and structure in artificial language games. Our first result is that structures developing in the artificial language fit native language semantic structures to a moderate to good degree, confirming our expectations. This does not necessarily imply a causal effect of native language on artificial language formation, but at the very least demonstrates a cognitive link between the two, the exact nature of which remains unclear. Our second result showed that the semantic structure shared between the native and artificial language influenced the performance and pragmatics of the artificial language, however only for speakers of English. This is evidence for categorical facilitation in artificial language communication, and for a direct behavioral influence of the semantic structure shared by the artificial and native languages. Methodologically, we argue 1) that potential biases towards native language structures in artificial language games should be taken into consideration

more often, and 2) that meaning spaces used to study several different languages at once should be carefully tailored towards the respective structures within those languages.

Data availability statement

The data that support the findings of this study are openly available in the Open Science Framework at https://doi.org/10.17605/OSF.IO/A8BGE.

Acknowledgements

We would like to thank all people that contributed to the Color Game, either in its creation or as a participant. A summary of acknowledgements regarding these contributions can be found here: https://osf.io/nsxu4/.

References

Baronchelli, A., Gong, T., Puglisi, A., & Loreto, V. (2010). Modeling the emergence of universality in color naming patterns. *Proceedings of the National Academy of Sciences*, *107*(6), 2403–2407. https://doi.org/10.1073/pnas.0908533107

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. California UP.

Bornstein, M. H. (1987). Perceptual categories in vision and audition. In *Categorical perception: The groundwork of cognition.* (pp. 287–300). Cambridge University Press.

Boster, J. (1986). Can Individuals Recapitulate the Evolutionary Development of Color Lexicons? *Ethnology*, *25*(1), 61. https://doi.org/10.2307/3773722

Brown, A. M., Lindsey, D. T., & Guckes, K. M. (2011). Color names, color categories, and color-cued visual search: Sometimes, color perception is not categorical. *Journal of Vision*, *11*(12), 2–2. https://doi.org/10.1167/11.12.2

Brown, R. W., & Lenneberg, E. H. (1954). A study in language and cognition. *The Journal of Abnormal and Social Psychology*, *49*(3), 454–462. https://doi.org/10.1037/h0057814

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*(2), 261–304.

Carr, J. W., Smith, K., Cornish, H., & Kirby, S. (2017). The Cultural Evolution of Structured Languages in an Open-Ended, Continuous World. *Cognitive Science*, *41*(4), 892–923. https://doi.org/10.1111/cogs.12371

Christensen, P., Fusaroli, R., & Tylén, K. (2016). Environmental constraints shaping constituent order in emerging communication systems: Structural iconicity, interactive alignment and conventionalization. *Cognition*, *146*, 67–80. https://doi.org/10.1016/j.cognition.2015.09.004

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Cuskley, C. (2019). Alien forms for alien language: Investigating novel form spaces in cultural evolution. *Palgrave Communications*, *5*(1), 87. https://doi.org/10.1057/s41599-019-0299-5

Davidoff, J., Goldstein, J., Tharp, I., Wakui, E., & Fagot, J. (2012). Perceptual and categorical judgements of colour similarity. *Journal of Cognitive Psychology*, *24*(7), 871–892. https://doi.org/10.1080/20445911.2012.706603

Everaert, M. B. H., Huybregts, M. A. C., Chomsky, N., Berwick, R. C., & Bolhuis, J. J. (2015). Structures, Not Strings: Linguistics as Part of the Cognitive Sciences. *Trends in Cognitive Sciences*, *19*(12), 729–743. https://doi.org/10.1016/j.tics.2015.09.008

Frank, M. C., & Goodman, N. D. (2012). Predicting Pragmatic Reasoning in Language Games. *Science*, *336*(6084), 998–998. https://doi.org/10.1126/science.1218633

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science*, *29*(5), 737–767. https://doi.org/10.1207/s15516709cog0000_34

Galantucci, B. (2009). Experimental Semiotics: A New Approach for Studying Communication as a Form of Joint Action. *Topics in Cognitive Science*, *1*(2), 393–410. https://doi.org/10.1111/j.1756-8765.2009.01027.x

Galantucci, B., & Garrod, S. (2011). Experimental Semiotics: A Review. *Frontiers in Human Neuroscience*, *5*. https://doi.org/10.3389/fnhum.2011.00011

Galantucci, B., Garrod, S., & Roberts, G. (2012). Experimental Semiotics. *Language and Linguistics Compass*, *6*(8), 477–493. https://doi.org/10.1002/lnc3.351

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, *8*(1), 8–11. https://doi.org/10.1016/j.tics.2003.10.016

Gilbert, A. L., Regier, T., Kay, P., & Ivry, R. B. (2006). Whorf hypothesis is supported in the right visual field but not the left. *Proceedings of the National Academy of Sciences*, *103*(2), 489–494. https://doi.org/10.1073/pnas.0509868103

Grice, P. (1989). *Studies in the way of words*. Harvard Univ. Press.

Harnad, S. (1987). Psychophysical and cognitive aspects of categorical perception: A critical overview. In *Categorical perception: The groundwork of cognition* (pp. 1–52). Cambridge University Press.

Hockett, C. F. (1960). The origin of speech. *Scientific American*, *203*(3), 88–97.

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural Equation Modelling: Guidelines for Determining Model Fit. *Electronic Journal of Business Research Methods*, *6*(1), 53–60.

Jäger, G. (2012). Using Statistics for Cross-linguistic Semantics: A Quantitative Investigation of the Typology of Colour Naming Systems. *Journal of Semantics*, *29*(4), 521–544. https://doi.org/10.1093/jos/ffs006

Kay, P., & Regier, T. (2003). Resolving the question of color naming universals. *Proceedings of the National Academy of Sciences*, *100*(15), 9085–9089. https://doi.org/10.1073/pnas.1532837100

Kay, P, & Regier, T. (2006). Language, thought and color: Recent developments. *Trends in Cognitive Sciences*, *10*(2), 51–54. https://doi.org/10.1016/j.tics.2005.12.007

Kay, Paul, Berlin, B., Maffi, L., Merrifield, W. R., & Cook, R. (2009). *The world color survey*. CSLI Publications Stanford, CA:

Kay, Paul, & Kempton, W. (1984). What Is the Sapir-Whorf Hypothesis? *American Anthropologist*, *86*(1), 65–79. https://doi.org/10.1525/aa.1984.86.1.02a00050

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686. https://doi.org/10.1073/pnas.0707835105

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102. https://doi.org/10.1016/j.cognition.2015.03.016

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago press.

Lewis, D. (1969). *Convention*. Harvard Univ. Press.

Little, H., Eryılmaz, K., & de Boer, B. (2017). Signal dimensionality and the emergence of combinatorial structure. *Cognition*, *168*, 1–15. https://doi.org/10.1016/j.cognition.2017.06.011

Luo, M. R., Cui, G., & Rigg, B. (2001). The development of the CIE 2000 colour-difference formula: CIEDE2000. *Color Research & Application*, *26*(5), 340–350. https://doi.org/10.1002/col.1049

Malt, B. C., Sloman, S. A., & Gennari, S. P. (2003). Universality and language specificity in object naming. *Journal of Memory and Language*, *49*(1), 20–42. https://doi.org/10.1016/S0749-596X(03)00021-4

Martin, A., & Culbertson, J. (2020). Revisiting the Suffixing Preference: Native-Language Affixation Patterns Influence Perception of Sequences. *Psychological Science*, *31*(9), 1107–1116. https://doi.org/10.1177/0956797620931108

Martin, A., Ratitamkul, T., Abels, K., Adger, D., & Culbertson, J. (2019). Cross-linguistic evidence for cognitive universals in the noun phrase. *Linguistics Vanguard*, *5*(1). https://doi.org/10.1515/lingvan-2018-0072

Morin, O., Winters, J., Müller, T. F., & Morisseau, T. (2020). *An overview of the "Color Game" App project*. SocArXiv. osf.io/preprints/socarxiv/cjaxw

Morin, O., Winters, J., Müller, T. F., Morisseau, T., Etter, C., & Greenhill, S. J. (2018). What smartphone apps may contribute to language evolution research. *Journal of Language Evolution*. https://doi.org/10.1093/jole/lzy005

Müller, T. F., Winters, J., & Morin, O. (2019). The Influence of Shared Visual Context on the Successful Emergence of Conventions in a Referential Communication Task. *Cognitive Science*, *43*(9), e12783. https://doi.org/10.1111/cogs.12783

Nölle, J., Staib, M., Fusaroli, R., & Tylén, K. (2018). The emergence of systematicity: How environmental and communicative factors shape a novel communication system. *Cognition*, *181*, 93–104. https://doi.org/10.1016/j.cognition.2018.08.014

Perfors, A., & Navarro, D. J. (2014). Language Evolution Can Be Shaped by the Structure of the World. *Cognitive Science*, *38*(4), 775–793. https://doi.org/10.1111/cogs.12102

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, *104*(4), 1436–1441. https://doi.org/10.1073/pnas.0610341104

Regier, Terry, & Kay, P. (2009). Language, thought, and color: Whorf was half right. *Trends in Cognitive Sciences*, *13*(10), 439–446. https://doi.org/10.1016/j.tics.2009.07.001

Roberson, D., & Davidoff, J. (2000). The categorical perception of colors and facial expressions: The effect of verbal interference. *Memory & Cognition*, *28*(6), 977–986. https://doi.org/10.3758/BF03209345

Roberson, D., Davidoff, J., Davies, I. R. L., & Shapiro, L. R. (2005). Color categories: Evidence for the cultural relativity hypothesis. *Cognitive Psychology*, *50*(4), 378–411. https://doi.org/10.1016/j.cogpsych.2004.10.001

Roberson, D., Davies, I., & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, *129*(3), 369–398. https://doi.org/10.1037/0096-3445.129.3.369

Roberson, D., Pak, H., & Hanley, J. R. (2008). Categorical perception of colour in the left and right visual field is verbally mediated: Evidence from Korean. *Cognition*, *107*(2), 752–762. https://doi.org/10.1016/j.cognition.2007.09.001

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Scott-Phillips, T. C. (2015). *Speaking our minds: Why human communication is different, and how language evolved to make it special*. Palgrave Macmillan.

Scott-Phillips, T. C., & Kirby, S. (2010). Language evolution in the laboratory. *Trends in Cognitive Sciences*, *14*(9), 411–417. https://doi.org/10.1016/j.tics.2010.06.006

Scott-Phillips, T. C., Kirby, S., & Ritchie, G. R. S. (2009). Signalling signalhood and the emergence of communication. *Cognition*, *113*(2), 226–233. https://doi.org/10.1016/j.cognition.2009.08.009

Selten, R., & Warglien, M. (2007). The emergence of simple languages in an experimental coordination game. *Proceedings of the National Academy of Sciences*, *104*(18), 7361–7366. https://doi.org/10.1073/pnas.0702077104

Silvey, C., Kirby, S., & Smith, K. (2015). Word Meanings Evolve to Selectively Preserve Distinctions on Salient Dimensions. *Cognitive Science*, *39*(1), 212–226. https://doi.org/10.1111/cogs.12150

Silvey, C., Kirby, S., & Smith, K. (2019). Communication increases category structure and alignment only when combined with cultural transmission. *Journal of Memory and Language*, *109*, 104051. https://doi.org/10.1016/j.jml.2019.104051

Sperber, D., & Wilson, D. (1996). *Relevance: Communication and cognition* (2nd ed). Blackwell Publishers.

Steels, L., & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, *28*(4), 469–489. https://doi.org/10.1017/S0140525X05000087

Tamariz, M. (2017). Experimental Studies on the Cultural Evolution of Language. *Annual Review of Linguistics*, *3*(1), 389–407. https://doi.org/10.1146/annurev-linguistics-011516-033807

Thierry, G., Athanasopoulos, P., Wiggett, A., Dering, B., & Kuipers, J.-R. (2009). Unconscious effects of language-specific terminology on preattentive color perception. *Proceedings of the National Academy of Sciences*, *106*(11), 4567–4570. https://doi.org/10.1073/pnas.0811155106

Tomasello, M. (2010). *Origins of Human Communication*. MIT Press.

Verhoef, T., Kirby, S., & de Boer, B. (2014). Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals. *Journal of Phonetics*, *43*, 57–68. https://doi.org/10.1016/j.wocn.2014.02.005

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, *104*(19), 7780–7785. https://doi.org/10.1073/pnas.0701644104

Winters, J., Kirby, S., & Smith, K. (2015). Languages adapt to their contextual niche. *Language and Cognition*, *7*(03), 415–449. https://doi.org/10.1017/langcog.2014.35

Winters, J., Kirby, S., & Smith, K. (2018). Contextual predictability shapes signal autonomy. *Cognition*, *176*, 15–30. https://doi.org/10.1016/j.cognition.2018.03.002

Winters, J., & Morin, O. (2019). From Context to Code: Information Transfer Constrains the Emergence of Graphic Codes. *Cognitive Science*, *43*(3), e12722. https://doi.org/10.1111/cogs.12722

Witzel, C. (2018). Misconceptions about colour categories. *Review of Philosophy and Psychology*, *10*, 499–540. https://doi.org/10.1007/s13164-018-0404-5

Witzel, C., & Gegenfurtner, K. R. (2011). Is there a lateralized category effect for color? *Journal of Vision*, *11*(12), 16–16. https://doi.org/10.1167/11.12.16

Witzel, C., & Gegenfurtner, K. R. (2013). Categorical sensitivity to color differences. *Journal of Vision*, *13*(7), 1–1. https://doi.org/10.1167/13.7.1

Wright, O., Davies, I. R. L., & Franklin, A. (2015). Whorfian effects on colour memory are not reliable. *Quarterly Journal of Experimental Psychology*, *68*(4), 745–758. https://doi.org/10.1080/17470218.2014.966123

Xu, J., Dowman, M., & Griffiths, T. L. (2013). Cultural transmission results in convergence towards colour term universals. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1758), 20123073–20123073. https://doi.org/10.1098/rspb.2012.3073

Youn, H., Sutton, L., Smith, E., Moore, C., Wilkins, J. F., Maddieson, I., Croft, W., & Bhattacharya, T. (2016). On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*, *113*(7), 1766–1771. https://doi.org/10.1073/pnas.1520752113

Zhou, K., Mo, L., Kay, P., Kwok, V. P. Y., Ip, T. N. M., & Tan, L. H. (2010). Newly trained lexical categories produce lateralized categorical perception of color. *Proceedings of the National Academy of Sciences*, *107*(22), 9974–9978. https://doi.org/10.1073/pnas.1005669107

Zimmer, A. C. (1982). What really is turquoise? A note on the evolution of color terms. *Psychological Research*, *44*(3), 213–230. https://doi.org/10.1007/BF00308421

Zollinger, H. (1984). Why just turquoise? Remarks on the evolution of color terms. *Psychological Research*, *46*(4), 403–409. https://doi.org/10.1007/BF00309072