

# **The Emergence and Organization of Communicative Signals Through Interaction**

**Dissertation**

zur Erlangung des akademischen Grades

doctor philosophiae (Dr. phil.)

vorgelegt dem Rat für Sozial- und Verhaltenswissenschaften

der Friedrich-Schiller-Universität Jena

von M.Sc. Thomas F. Müller

geboren am 08.06.1991 in Lichtenfels

Gutachter:

1. Dr. Olivier Morin (Max-Planck-Institut für Menschheitsgeschichte, Jena)
2. Prof. Dr. Stefan Schweinberger (Friedrich-Schiller-Universität Jena)
3. Prof. Dr. Kristian Tylén (Aarhus Universitet)

Tag der mündlichen Prüfung: 07.01.2021

## Acknowledgments

This thesis would not have been possible without the support of many people. First, I would like to thank my supervisors, Olivier Morin and Stefan Schweinberger, for giving me the opportunity to write this dissertation in between the Max Planck Institute for the Science of Human History and the Friedrich Schiller University. I thank Olivier for all his hard work and scientific guidance regarding studies, manuscripts, presentations, and more, and Stefan for the scientific supervision as well as quick administrative help whenever needed. I would also like to thank Ira Noveck for his support and supervision as a part of my thesis advisory committee.

A big thanks goes to my colleagues (past and present) in the Mint, my scientific home during the PhD: To Oleg Sobchuk, Piers Kelly, Yoolim Kim, and Helena Miton (practically also a Mintie); a very special thanks to Barbara Pavlek, who had to suffer through sharing the office with me and survived, and who has always been a great support at work, and as a friend; and an equally big thank you to James Winters for his support at all times both as a scientific guide and as a friend (and also for that one time you pushed me out of the way when a pizza delivery motorcycle tried to run me over!). I also want to thank all the student assistants in our group that came and went during my PhD, especially Lidiia Romanova, Lisa Jeschke, Noro Schlorke, Moritz Dörfler, and Julia Bepamyatnykh.

I would like to thank my co-authors during the PhD not mentioned yet: Tiffany Morisseau, Juliane Bräuer, and Melanie Henschel. Thanks also to Joseph Watts, who kindly offered to share his office when ours was uninhabitable, and who organized the best board game evenings.

Thanks to all my friends in Jena who provided me with an active social life outside of the PhD, in particular Sophie, Tabitha, Alex, Caro, Celina, Sarah, Christian, Tina, and Fabi.

I want to thank my parents for all the care, support, and trust they provided without question, even though they never fully understood what my work was about. I thank my sister, brother-in-law, and grandmother for making me feel at home whenever I came back.

Finally, I thank Steffi for standing this through with me, patiently at my side and always supporting me, especially during the last months of write-up, pandemic, and home office. You have been a continuous motivation and I hope to be the same for you when you finish your dissertation.

## Table of Contents

|  |            |
|--|------------|
| Acknowledgments .....  | 3          |
| Abstract.....  | 6          |
| Deutsche Zusammenfassung .....   | 8          |
| <b>1. Introduction.....</b>  | <b>10</b>  |
| <b>1.1 Two Models of Communication.....</b>  | <b>15</b>  |
| <b>1.2 The Evolution of Communication Systems at Micro- and Macro-Level .....</b>  | <b>21</b>  |
| <b>1.3 Methodological Approach of the Thesis.....</b>  | <b>27</b>  |
| <b>1.4 Thesis Overview.....</b>  | <b>30</b>  |
| <b>2. The Influence of Shared Visual Context on the Successful Emergence of Conventions in a Referential Communication Task .....</b>                        | <b>34</b>  |
| <b>2.1 Abstract .....</b>  | <b>34</b>  |
| <b>2.2 Introduction .....</b>  | <b>35</b>  |
| <b>2.3 Experiment 1.....</b>   | <b>42</b>  |
| <b>2.4 Experiment 2.....</b>   | <b>51</b>  |
| <b>2.5 General Discussion .....</b>  | <b>61</b>  |
| <b>2.6 Conclusion.....</b>   | <b>65</b>  |
| <b>2.7 Appendix .....</b>  | <b>66</b>  |
| <b>3. Compression in Cultural Evolution: Homogeneity and Structure in the Emergence and Evolution of a Large-Scale Online Collaborative Art Project.....</b> | <b>68</b>  |
| <b>3.1 Abstract.....</b>   | <b>68</b>  |
| <b>3.2 Introduction .....</b>  | <b>69</b>  |
| <b>3.3 Materials and Methods.....</b>  | <b>75</b>  |
| <b>3.4 Results .....</b>   | <b>82</b>  |
| <b>3.5 Discussion.....</b>   | <b>91</b>  |
| <b>3.6 Conclusion.....</b>   | <b>95</b>  |
| <b>4. Color Terms: Natural Language Semantic Structure and Artificial Language Structure Formation in a Large-Scale Online Smartphone Application.....</b>   | <b>97</b>  |
| <b>4.1 Abstract .....</b>  | <b>97</b>  |
| <b>4.2 Introduction .....</b>  | <b>98</b>  |
| <b>4.3 Method.....</b>   | <b>104</b> |

|  |            |
|--|------------|
| <b>4.4 Results .....</b>               | <b>111</b> |
| <b>4.5 Discussion.....</b>             | <b>120</b> |
| <b>4.6 Conclusion.....</b>             | <b>124</b> |
| <b>5. Summary and Conclusion .....</b> | <b>126</b> |
| <b>6. References .....</b>             | <b>132</b> |
| <b>Ehrenwörtliche Erklärung.....</b>   | <b>159</b> |

## Abstract

Social interaction is a key feature of our daily lives; humans simply cannot help but interact with one another. This interaction is special with regard to its quantity, but it also shows distinct qualities such as a special propensity to read each other's intentions. One specific kind of interaction that humans engage in frequently and that exemplifies this particularly well is communication. By producing and interpreting signals in their specific context, interlocutors are able to communicate successfully, even about concepts for which they do not yet share conventional signs. Over repeated interaction, these novel signals can conventionalize, and eventually be culturally transmitted to new individuals. Through repeated episodes of transmission, entire communicative systems, such as languages, can emerge and evolve.

In this thesis, I build on the framework outlined above to study how human communicative signals can emerge and become organized via interaction. To this end, I present the results of three empirical studies each concerned with one specific question under this account. The first study represents two artificial language experiments investigating the role of context for the successful emergence of novel communicative conventions. This relates to the broad question of how human communicative signals can emerge in the first place; although the necessity of communication in context has been argued for in many pragmatic frameworks, little empirical work about it has been done at the stage of emergence. The results demonstrate that access to the shared context, i.e. the amount of information two interlocutors have in common, improves communicative success with novel conventions. Furthermore, conventions that were created under the condition of shared context could be generalized more successfully to novel contexts as well.

The second study takes a step back in perspective and focuses on the evolution of population-level cultural patterns. A known result from artificial language experiments is that evolving communication systems can become compressed as a response to simplifying pressures, such as the need to memorize the signals' meanings. Building on this idea, the study investigates whether similar compression effects can be found in the evolution of a large-scale visual art collaboration by analyzing a massive dataset of online behavior. Here, a rising pressure for simplification was hypothesized to occur due to increasing competition between the participants. The main results of the study show that compressible patterns did develop, could be predicted through time from the increasing competition, and occurred due to the evolution of systematic structure, not mere

increasing homogeneity. This relates to the broad question of how cultural traits in general (and communicative signals in particular) can become organized via interaction – cooperative and competitive.

The third study combines the perspectives of the first two studies and aims to relate existing communicative conventions about color terms in natural languages to novel conventions created within the environment of a custom smartphone application, designed to study language evolution. This relates to both the emergence and organization of communicative signals, as the focus is on the emerging semantic structure applied for color categories. The first main finding is that for native speakers of English, German, and French, there was a good to moderate correspondence between natural language and artificial language semantic structure. Secondly, for native speakers of English, communicative performance and the number of sent signals could be predicted by the shared semantic structure. The study can provide important insight into potential biases towards natural language structure in artificial language experiments, and opens up the possibility to investigate the effect of semantic structure on communicative performance without actually making use of the natural language of participants.

The three studies show the usefulness of combining different methodological approaches – experimental laboratory studies, large-scale online experiments, and massive data sets of online behavior – to address questions at different levels of granularity. Taken together, the studies place individual interactions firmly at the base of both the emergence and organization of communicative signals. As a result of these interactions, entire systems of communication can emerge.

## **Deutsche Zusammenfassung**

Soziale Interaktion ist ein wesentlicher Teil unseres täglichen Lebens; als Menschen können wir nicht anders, als miteinander zu interagieren. Diese Interaktion ist besonders in Hinsicht auf ihre Häufigkeit, aber auch ihre Qualität, wie z.B. der speziellen menschlichen Neigung dazu, die Absichten seines Gegenüber zu erkennen. Eine spezifische Art der Interaktion, an der Menschen häufig teilnehmen und die diese Qualität besonders gut demonstriert, ist Kommunikation. Indem sie Signale innerhalb eines spezifischen Kontexts produzieren und interpretieren, sind Gesprächspartner dazu in der Lage, erfolgreich zu kommunizieren, sogar über Konzepte, für die sie sich noch keine konventionellen Signale teilen. Mit wiederholter Interaktion können diese neuartigen Signale zur Konvention werden und schließlich kulturell an neue Individuen übertragen werden. Durch wiederholte Übertragung können ganze Kommunikationssysteme, wie z.B. Sprachen, entstehen und sich weiterentwickeln.

In dieser Arbeit baue ich auf diese Grundlagen auf, um zu erforschen, wie menschliche Kommunikationssignale entstehen und durch Interaktion organisiert werden können. Dazu stelle ich die Ergebnisse dreier empirischer Studien vor, die sich jeweils mit einer spezifischen Frage in diesem Rahmen beschäftigen. Die erste Studie besteht aus zwei Experimenten mit einer künstlichen Sprache, die die Rolle des Kontexts für die erfolgreiche Entstehung neuer kommunikativer Konventionen untersuchen. Dies bezieht sich auf die übergeordnete Frage, wie menschliche Kommunikationssignale überhaupt entstehen können; obwohl viele pragmatische Modelle die Notwendigkeit des Kontexts für Kommunikation annehmen, ist bisher wenig empirische Forschung zum Zeitpunkt in Entstehung befindlicher Kommunikation durchgeführt worden. Die Ergebnisse zeigen, dass Zugang zu geteiltem Kontext, also die Menge an Information, die zwei Gesprächspartner gemeinsam haben, den kommunikativen Erfolg mittels neuer Konventionen verbessert. Weiterhin konnten Konventionen, die in der Bedingung mit geteiltem Kontext entstanden sind, besser für neue Kontexte generalisiert werden.

Die zweite Studie vergrößert diese Perspektive und beschäftigt sich mit der Evolution kultureller Strukturen auf Populationsebene. Ein bekanntes Ergebnis aus Experimenten mit künstlichen Sprachen ist, dass sich entwickelnde Kommunikationssysteme komprimiert werden können, wenn Druck zur Vereinfachung auf sie ausgeübt wird, wie z.B. durch die Notwendigkeit, die Bedeutung von Signalen im Gedächtnis zu behalten. Die Studie baut auf diese Idee auf und



prüft, ob ähnliche Kompressionseffekte in der Evolution eines kollaborativen visuellen Kunstprojektes entdeckt werden können, indem ein großer Datensatz zu Online-Verhalten analysiert wird. Ein steigender Druck zur Vereinfachung wird angenommen, weil der Wettbewerb zwischen den Teilnehmern über die Zeit zunimmt. Die zentralen Ergebnisse der Studie zeigen, dass sich komprimierbare Strukturen herausbildeten, die durch den mit der Zeit zunehmenden Wettbewerb vorhergesagt werden konnten und wegen der Evolution von systematischer Ordnung und nicht nur steigender Homogenität entstanden. Dies bezieht sich auf die übergeordnete Frage, wie kulturelle Merkmale im allgemeinen (und Kommunikationssignale im besonderen) durch Interaktion – kooperativ und konkurrierend – organisiert werden können.

Die dritte Studie vereint die Perspektiven der ersten beiden Studien und vergleicht bestehende kommunikative Konventionen zu Farbwörtern in natürlichen Sprachen mit neuen Konventionen, die innerhalb einer eigens erstellten Smartphone-App zur Erforschung von Sprachevolution entstehen. Dies bezieht sich sowohl auf die Entstehung als auch die Organisation von Kommunikationssignalen, weil der Fokus auf der entstehenden semantischen Struktur liegt, die für Farbkategorien angewendet wird. Das erste zentrale Ergebnis ist, dass englische, deutsche und französische Muttersprachler eine gute bis moderate Übereinstimmung zwischen der semantischen Struktur in natürlicher Sprache und künstlicher Sprache zeigten. Zweitens konnte für englische Muttersprachler auch der kommunikative Erfolg und die Anzahl der benutzten Signale durch diese gemeinsame semantische Struktur vorhergesagt werden. Die Studie hat wichtige Implikationen hinsichtlich potentieller Verzerrungen in Richtung natürlicher Sprachstruktur bei Experimenten mit künstlichen Sprachen und eröffnet die Möglichkeit, den Effekt der semantischen Struktur auf kommunikativen Erfolg zu erforschen, ohne tatsächlich die natürliche Sprache von Teilnehmern zu benutzen.

Die drei Studien zeigen die Vorteile davon auf, verschiedene methodische Herangehensweisen – Laborexperimente, groß angelegte Online-Experimente, und große Datensätze zu Online-Verhalten – zu kombinieren, um Fragestellungen in unterschiedlicher Auflösung zu betrachten. Insgesamt demonstrieren die Studien, dass individuelle Interaktionen die Grundlage für sowohl die Entstehung als auch Organisation von Kommunikationssignalen sind. Als Ergebnis dieser Interaktionen können ganze Kommunikationssysteme entstehen.

# 1. Introduction

Human interaction is ubiquitous: Whether we buy a ticket from the bus driver to get to work, have a lengthy conversation with our colleague at the office, or cook as a team by sharing the work of cutting and frying vegetables, humans cannot help but interact with one another. As Schegloff put it, interaction can be considered “the primordial scene of social life” (1996, p. 54). Human interaction is special with regard to the intensity and duration that humans engage in it, for example when compared to primates (Dunbar, 1998; Enfield & Levinson, 2006). It is, however, not only the quantity of interactions, but especially the quality of interactions that makes humans stand out: One particular human feature that has been emphasized frequently is the role of *intention-reading* in interaction (e.g. Atlas, 2005; Clark, 1996; Grice, 1989; Horn, 2004; Levinson, 2000). Levinson (2006) adds the special *turn-taking structure* underlying human interaction to this as another remarkable feature, and the way humans interact in a very *cooperative* way.

Regarding the ability to read others’ intentions, a common proposition is that humans possess a *Theory of Mind* (see e.g. Baron-Cohen et al., 1985; Premack & Woodruff, 1978). This means that we can form beliefs about other people’s beliefs, and they can in turn form beliefs about our beliefs about their beliefs, and so on – in short, metarepresentations. While the exact nature of these mental representations and the degree to which they are necessary for successful interaction is debatable (e.g. Breheny, 2006; Sperber & Wilson, 2002), a common view is that cognitive abilities of this sort underlie human interaction in some way (Scott-Phillips, 2015; Sperber, 2000; Tomasello, 2010). What follows is that humans typically will reply to the intentions of their interaction partner, not to their visible behaviors only. In fact, the same behavior could either warrant a reply or no reaction at all, depending on whether the partner displays their intention to communicate along with it (Scott-Phillips, 2015). In section 1.1, I will outline the role of intentionality and intention-reading in communication more concretely. Communication is the main domain that many concepts within the scientific study of interaction have originated in.

With regard to the special structure of human interactions, the focus has been particularly on *turn-taking* and the sequential flow of interaction (Sacks et al., 1974; Schegloff, 2006). The term refers to an interaction structure characterized by alternating activity of the interaction partners, rather than simultaneous one. Instead of, for example, talking over each other, humans attempt to alternate in conversation by rapidly chaining action sequences after one another. This happens in

a semantically coherent manner, which means that, for example, a request is usually followed by rejection or by granting it (Schegloff, 1996). A simple, minimal sequence such as this one can be expanded by chaining other sequences before or after it, or even by embedding additional turns (such as a request for clarification) into the sequence itself (e.g. Schegloff, 2006). The structure is also typically not planned in advance and laid out before the conversation, but rather an orderliness that emerges through live interaction (Clark, 1996). The turn-taking structure may seem less efficient due to alternating rather than double, simultaneous activity (which humans are capable of, for example, in simultaneous translation; Levinson, 2006). However, it allows for constant and step-wise testing whether the intended result has been achieved with regard to the interaction partner: We work with other people's intentions rather than only their visible behavior and those cannot be directly observed. On top of this, there can potentially be noise in the transmission of this behavior. Consequently, we need to constantly make sure we understand one another, which has been termed *repair* (e.g. Schegloff et al., 1977). For example, a misunderstanding can be followed up by a simple "Huh?" or by asking "What do you mean?". These human repair mechanisms are an essential part of the structure of human interaction and have been found to be culturally universal (Dingemans et al., 2013, 2015), like the turn-taking and sequence organization of communication itself (Kendrick et al., 2020; Stivers et al., 2009).

Lastly, when stated above that human interaction displays *cooperation*, we can distinguish at least two different ways in which this can be understood: Hurford (2007) differentiates between material cooperation and communicative cooperation. Material cooperation refers to the intent of the interaction being either pro- or anti-social. For instance, deception or other competitive behavior would be seen as materially un-cooperative. Humans show outstanding material cooperation, even if other species are known to collaborate as well (e.g. Boyd & Richerson, 2006; Tomasello, 2010). Communicative cooperation, in contrast, concerns the question of whether an interaction partner is aiming to be comprehensible in their actions, meaning they "play by the same rules". In day-to-day conversation, this can, for example, simply mean using the same language. One can be materially un-cooperative but communicatively cooperative; for instance, Hurford (2007) mentions the example of a tennis match, in which both players will be communicatively cooperative but inherently materially un-cooperative due to the competition between them. If they were communicatively un-cooperative, they could, for example, not show up for the match at all. Day-to-day interaction between humans is mostly communicatively cooperative (Hurford, 2007);

as a species, we are particularly motivated to cooperate communicatively, from an early age on (Tomasello et al., 2005).

The three features outlined here (intention-reading, sequential structure, and cooperation) are, however, not disconnected, but feed into each other (Levinson, 2006): Our reliance on reading others' intentions might be part of what makes testing for misunderstandings necessary, especially since we need to form coherent replies on the spot; likewise, inferring an interaction partner's perspective allows us to realize what it would take to be cooperative (both materially and communicatively) in a given situation in the first place. Similarly, revealing our intentions to each other makes particular sense because our interactions tend to be mostly cooperative; if they are aiming to be competitive, we should try to conceal our intentions instead. Being motivated to cooperate encourages our tendencies to reassure mutual understanding and to apply our repair mechanisms.

Interaction thus can be seen as a complex dynamical system (Dale et al., 2013), with a multitude of interdependent mechanisms self-organizing to create a functional interactive performance, depending on the situation at hand. The discussed features and the profound importance of human interaction have even led Levinson (2006, 2019) to postulate an innate "interaction engine", specific to humans only. The engine encompasses the necessary cognitive abilities and behavioral dispositions displayed in the features described in more detail above. It does not prevent cross-cultural differences, but instead serves as a building block for diversity in social interaction: By starting out with the same cognitive abilities, different cultures are able to form their own traditions.

One example of such a cultural tradition are communicative signals, organized into culture-specific communicative systems, such as languages. It is this organization of communicative signals that I aim to study in this thesis. Communication is a particularly good case for an interaction displaying the special features above: All of the three broad features (intention-reading, sequential structure, and cooperation) apply very well to communication, and conversational interaction features centrally in the organization of human social life (Schegloff, 1996, 2006). Most of the special features of human interaction described above could also be described as "pragmatic competence" (e.g. Scott-Phillips, 2015; Woensdregt & Smith, 2017). When humans interact face-to-face, they cannot help but communicate (Clark, 1996). In fact, it is hard to imagine an

interactional setting that involves intention-reading, a sequential structure, and cooperation, but no communication at all. It is of no surprise then that many of the ideas about the special role of human interaction came from insights related to language in the first place.

An important distinction to make here is that the object of interest in the current thesis is not restricted to language (or better, *languages*; see section 1.2), but to human communicative signals more generally. *Signals*, in the most general sense (i.e. also applying to animal communication), are defined as actions designed to cause a response in an addressee (Maynard-Smith & Harper, 2003); this response must also have been designed to react to the signal, in an interdependent relationship. Many animal warning calls are good examples of signals under this definition: The interdependence between signal and response is visible in that if either the signalers stop signaling *or* the addressees stop responding, the signal would cease to exist. Communicative signals thus differ from *coercion* because the latter does not require the reaction to be designed for the action: In this case, the action simply forces a response, and it does not carry information for the addressee. They also differ from *cues* because the latter do not require the action to be designed for the respective reaction: In this case, the user of the cue responds to a piece of information that was displayed unwillingly or unknowingly.

As an animal behavior, signals can develop from cues or coercion by processes of natural selection, or ontogenetically. As I will describe in section 1.1, in humans novel signals can also be created intentionally (all three options can thus lead to the “design” of interdependence mentioned earlier). Human *communication systems* then represent a collection of communicative signals that has become organized. This organization usually involves rules about the interrelations between the individual signals, and about how they can be combined. The most obvious example is found in human languages (both speech and signed languages) and their underlying semantic and grammatical rules, but other non-linguistic communication systems exist as well, such as traffic signs or emojis, for example.

In this thesis, I aim to investigate how social interaction creates and organizes human communicative signals. This entails two sub-questions, namely 1) how communicative signals can emerge in the first place and 2) how the signals get organized into systems after communication has been established. To this end, two research areas have to be introduced in more detail: First, I will go deeper into how human communication can be described on the interactional level. Thus,

section 1.1 will introduce two influential models of communication, the *code model* and the *ostensive-inferential model*, and discuss concepts relevant to them such as the notion of *context* and *convention*. This section will explain how humans are able to create novel signals even if they have no shared conventions yet. Second, in section 1.2, I will go into more detail about how individual interactions can create and transmit entire systems of communication. This is a question that concerns *cultural evolution* and the problem of linkage between processes of *micro-* and *macro-evolution*. Having introduced these concepts, I will briefly outline the *evolution of languages* under the present account, mostly from an experimental point of view. I then summarize the key *methodological approaches* of the thesis in section 1.3, before providing a *thesis overview* about the three empirical studies that form the main body of the thesis, in section 1.4. Chapters 2, 3, and 4 represent the three *empirical studies*. At the end, Chapter 5 contains a *conclusion* summarizing the results of the thesis and their implications, also looking into potential future work.

## 1.1 Two Models of Communication

Communication is the form of human interaction that I will focus on in the thesis. Although not every interaction is necessarily communicative, most human interactions involve communication in some way. In its very simplest form, communication can be defined as occurring when information gets transmitted from one individual to another by first being encoded in some form on the side of the sender, and then decoded on the receiver's side. This view forms the so-called *code model* of communication (Shannon & Weaver, 1949). Typically, some sort of noise is included in the model to distort the signal during transmission through the channel. Communication under this definition covers not only human interactions, but a variety of animal behaviors as well as automatically transmitted signals too (such as chemical communication via hormones in the human body). Signals are still defined as described in the previous section, i.e. as interdependent sets of an action and the designed reaction. However, the code model emphasizes that they are pieces of encoded information that are transmitted. They simply have to be translated into their meanings through usage of the code, knowledge of which is presupposed. As such, communication of this kind is dependent on the associations between the signals and their meanings (Scott-Phillips, 2015).

A pure code model is very useful to systematically represent and break down a communicative interaction into analyzable parts. However, it runs into problems as a model of human communication once we consider that a typical utterance will involve a large amount of ambiguity (e.g. Clark, 1996): If I say to you "This movie is great" while standing in front of a rack of different DVDs, simply decoding the intended referent will be impossible for you from the linguistic signal alone. As such, a central problem of human communication is that it is underdetermined from the view of a pure code-to-meaning translation in communication (Atlas, 2005; Grice, 1989). This underdeterminacy could be accommodated for by suggesting more complex forms of associations within the framework of the code model. For instance, we could imagine a word with ambiguous but distinct meanings (i.e. polysemy) such as "bank" to be disambiguated based on a code that says "if close to a river, this means 'river bank'". In this way, associations could be formulated probabilistically rather than deterministically. For demonstratives like "this" in the example above, however, this is hardly going to be the case; as will be important when describing the ostensive-inferential model of communication, the possible interpretations of an utterance are basically

unlimited, and thus the possible associations required to account for all these cases would be intractable as well (Clark, 1996).

Millikan (1984, 2005) proposed a more recent model of communication that is based on a similar premise as the code model, even though she did not phrase it like that herself. In her view, communicative signals that are systematically associated with a conventional meaning (the signals' "direct proper function") are still the key concept that makes human communication possible. However, every time they use a signal in a specific instance, interlocutors endow this signal with a "derived proper function" from the meaning; that is, they may deviate from the conventional meaning or stick very close to its core. As such, knowledge about the core meaning of a signal is translated into tokens of its use every time the signal is produced or comprehended, and the token may or may not be very close to that fundamental core meaning. This acknowledges massive ambiguity in every usage context of an utterance (Origg & Sperber, 2000). Crucially, this ambiguity is assumed to resolve because all possible contextual meanings of an utterance are associated with the signal for the interlocutors. The task for the receiver of an utterance thus consists of recognizing which of all the possible meanings is the appropriate one; meanwhile, inference of the intentions or other mental states of the partner is claimed not to play a role in this disambiguation process. Like in the basic code model above, it remains unclear how disambiguation from infinite possible meanings of an utterance can be plausible psychologically. In any case, disambiguation in this way should be incredibly hard and inefficient if it has to rely only on the rest of the utterance.

Thus, the important factor that is missing from both models presented so far is the inference of intentions by the interaction partner; especially after the introduction detailing the special features that make up human interaction, the shortcomings of code models are apparent. Since the code has to be mutually known (or misunderstandings would arise), the models do not explain how the arbitrary associations between signal and meaning were created in the first place, or indeed how it is possible for humans to spontaneously communicate a meaning that has never been referred to before. This is both a feature and a problem, since it makes the model more general and has led to its application to, for example, animal communication (Rendall et al., 2009) or information technology (Shannon & Weaver, 1949), but at the same time falls short of fully describing what happens in human communication. For instance, a candidate explanation for the origin of many



animal signals is that they developed from repeated cues or coercive behaviors (that evolved for reasons independent of communication), which then lead the second half of the signal because the signaler/addressee benefits from an interdependence, too (Maynard-Smith & Harper, 2003). This can be observed to be the case both phylogenetically and ontogenetically. However, these kinds of signals cannot develop spontaneously but only over many iterations, and they are necessarily constrained to take only the form of pre-existing cues and coercive behaviors.

The second framework that I rely on throughout the thesis (this one applying only to human communication) is the ostensive-inferential model of communication (Sperber & Wilson, 1996). Here, the role of inference and intentionality in communication is emphasized, following the influential work of Grice (1989). As the name of the model implies, it includes the processes at both sides of a communicative act: The sender makes manifest a communicative intention and at the same time produces a signal. The action of producing the signal while manifesting the intention to communicate through this signal is called *ostension*. The receiver of the signal takes the signal and interprets it, which constitutes *inference*. An inferential process is characterized by logical reasoning based on a set of premises (Sperber & Wilson, 1996). To recognize the content of the sender's informative intention, the receiver reasons about it, taking into account the meaning of the signal and contextual factors such as the immediate environment. As such, the communicative process now is not restricted to de-coding anymore, but is flexible with regard to the sender's intention behind a signal. This is the difference between so-called signal meaning and speaker meaning (Grice, 1989). The contrast to the code models described above is that for ostensive-inferential communication, no reliance on associations between signals and meanings is necessary to achieve successful communication, although there can still be codes nevertheless. Through ostension and inference, interlocutors are also able to communicate about meanings they never communicated about before, and in fact about meanings they had no prior signals for before: The model thus also delivers an explanation as to how human communicative signals can emerge directly and spontaneously in human interaction.

As an example, consider you are on your way to a concert with a friend and already running late when you randomly meet a mutual acquaintance of yours that starts talking to your friend and holds you up. Getting impatient, you make eye contact with your friend and point to your left wrist in an exaggerated manner, even though you are not wearing a watch today. Still, this ostensibly

signals “We are running out of time” to your friend, who manages to politely cut the ongoing conversation down and say goodbye to your acquaintance. We can now dissect this communicative interaction in the terms of the ostensive-inferential model: To express a communicative intention to inform the friend that “We do not have time right now”, you drew their attention by making eye contact. You then signaled this information by producing the gesture. Your friend realized you wanted to communicate with them and, perceiving the gesture, inferred the intended meaning: Knowing that people commonly wear watches on their left wrist and seeing the urgent nature of the situation, the most relevant interpretation of your intention for them is that you want to urge them to move on. This works even though you might have never used the signal in an interaction with this friend before, and even though you did not explicitly agree to use this gesture in this situation beforehand. In fact, the very same gesture could be interpreted very differently in different contexts, such as if you pointed to your empty wrist after someone had asked “Whose watch did I just find?”.

Interlocutors thus arrive at a common understanding of a communicative action by making inferences as to the most reasonable intention of the interaction partner, which only happens in *context*. In fact, sometimes the context can carry so much information that the display of the content itself is unnecessary, and all that has to be communicated is the communicative intention of the sender (Scott-Phillips, 2015; Tomasello, 2010). Returning to the example above, if you are confident that your friend is equally aware of the time pressure, simply raising your eyebrows, intently staring at them, or slightly tapping their shoulder should be enough to get the informative intention across; the exact form that the signal takes in this situation does not matter, as long as the communicative intention is recognized. Context in this notion refers to the shared knowledge, i.e. the “mutual cognitive environment” of the two interlocutors (Sperber & Wilson, 1996). This can encompass representations of the shared physical environment of the interlocutors, their shared conversation history, common knowledge related to the social environment and customs within society, and more (Clark, 1996). In the example above, the friend is drawing on shared knowledge about the situation being one in which time is an urgent matter and in which the acquaintance is an obstacle to this urgency. They are also drawing on shared knowledge about the world, for instance the fact that people often wear watches on their wrists. The important part for Sperber and Wilson (1996) is that the environment itself is not the basis of inference, but the cognitive representation thereof is. They are particularly interested in “relevance”: A piece of information is

cognitively relevant if the informational benefits derived from processing it are high, subtracting its processing costs. Communicative interpretations are relevant if their cognitive relevance is high compared to other possible interpretations. The range of interpretations that an utterance may receive is potentially infinite, and the search for possible interpretations would thus be endless if we could not rely on context (Clark, 1996). I will turn to studying the effect of context in Chapter 2 of the thesis.

One subtle consequence of viewing communication through this lens is that we have qualitatively separated the codes used in ostensive-inferential communication from the idea of codes in the code model: Scott-Phillips (2015) speaks of *conventional codes* and *natural codes*; the former have to be commonly agreed upon in a population, whereas the latter automatically signal an associated meaning. Examples would be the “Thumbs up” gesture for a conventional code (or, for brevity, simply *convention*), and the genetic code for a natural code. Illustrating the former, a traveler from Germany to Iran might produce the “Thumbs up” gesture in the intent to signal positivity or support, when in fact the local meaning is extremely offensive (Archer, 1997). This highlights the cross-cultural variation of conventions.

Humans can coordinate even without explicit signals by recognizing focal points (Schelling, 1960), which will lead to conventions by repeated trial. For instance, if I tell my colleague that we will meet after work to go for a drink together at 6pm tonight, but do not specify the meeting point, it is reasonable to expect them at the entrance or reception of the office building at that time. If we regularly meet and go for drinks in this way on Fridays, after a few weeks it is very likely that the meeting does not have to be communicated at all; colleagues will just meet at the entrance at 6pm conventionally, and whoever wants to join will be there. Conventions are thus arbitrary solutions to repeated coordination problems, which are jointly created (Lewis, 1969). They are not the only possible way to coordinate meaningful interactions, since they, by definition, only come about through repeatedly coordinating in some other way: Clark (1996) also mentions explicit agreement, precedence, or perceptual salience as potential coordinators.

Our special human capacity to interact offers the tools needed for the *emergence* of ostensive-inferential communication, which is a direct route from no communication to functional signals. Conventionalization is then part of the *evolution* of communication systems. Put this way, linguistic communication is a special case of ostensive-inferential communication, in which the

inferences are made on the contextual cues of the conventional linguistic signals. This is then also the main difference between linguistic communication and other types of communication that humans are capable of: The former makes use of an exceptionally powerful conventional communication system, but is still relying on the same human abilities that underlie all of our intentional communication (Scott-Phillips, 2015). In this view, language can be seen as a powerful tool mediating between minds, thus facilitating and expanding interaction (Fusaroli & Tylén, 2012; Tylén et al., 2010).

## 1.2 The Evolution of Communication Systems at Micro- and Macro-Level

Importantly, human interaction (communicative or not) does not happen in a vacuum, but, as I try to demonstrate in this chapter, is the behavioral link between the minds of individuals and their cultural traits. From introducing how humans can communicate ostensive-inferentially even without shared conventions, I now want to focus on the cultural evolution of communication systems through the usage of signals. The continued evolution of these systems constitutes the organization of the communicative signals. *Cultural evolution* is the change of cultural traits over time. Cultural behaviors are characterized by being learned socially, as opposed to learning individually or being genetically predisposed for them (Shennan, 2002). One central idea here is to compare its dynamics to the ones found in biological evolution (Dawkins, 1979), and to model them accordingly (Boyd & Richerson, 1988; Cavalli-Sforza & Feldman, 1981), while it should also be acknowledged that not all processes in cultural evolution are akin to biological evolution.

A working model of cultural evolution for the purposes of the thesis can be described as follows: There are three key elements that make up the cultural evolutionary process (also see Ferdinand, 2015; Höfler, 2009); at the base, there need to be *cultural traits* as the element that cultural evolution is enacted on. These are ideas or practices that are acquired through either innovation or transmission. They are not physical objects, but cognitive cultural representations. As such, they are the unit of cultural evolution, which can be manifested as visible behavior or artefacts. Because of that, they can only be observed “*in relation to* something else, including not only things of similar kinds, but also the social norms and intentions associated with items and the contexts in which they appear” (Enfield, 2014, pp. 53–54; emphasis in the original). *Innovation* (Barnett, 1953) is the process that creates a new cultural trait, or alternatively modifies an existing cultural trait (Charbonneau, 2015). It is only in its context of usage that innovation on a trait can be observed and made sense of, not in isolation (Tomasello, 1999): Since traits are directly visible only in the tokens of their usage, it is unclear what an innovation is innovating on until it is put into behavior. This is not unlike the way ostensive-inferential communication can only be made sense of in context. Depending on the context, the interpretation of a cultural behavior (and thus its underlying cultural trait) can change drastically (Höfler, 2009). Lastly, *transmission* refers to the social learning process from one individual to another, transmitting the cultural trait. Through transmission, traits can also be innovated on, either deliberately or by unfaithful transmission, i.e.

errors. There has been some debate on whether transmission in general has to be particularly faithful (e.g. Boyd & Richerson, 1996; Tomasello, 1999) or simply transformative (e.g. Morin, 2015; Sperber, 1996) for the ongoing evolution of particular human culture, but the discussion is not of most urgent importance for the purposes here.

The transmission process within the model of cultural evolution can be summarized as follows: An individual produces a cultural behavior at time step  $t$ . This behavior is observed by a different individual, passing through the new individual's cognitive system. The new individual can now form a cognitive cultural representation themselves, meaning the cultural trait has been transmitted by inference from the observed cultural behavior. From then on, they can produce the behavior themselves, potentially transmitting the trait again to yet another individual at time step  $t+1$ , and so on: this is what has been called a *transmission chain* (Bartlett, 1932). One way to think of these real-life processes that has been used to describe language evolution in particular is a design that has been introduced as *iterated learning*, whereby an individual learns a behavior via this inductive process from another individual that has acquired the behavior earlier in the same way (Kirby et al., 2014). Of course, experimental or simulation work simplifies the real process in many ways, for example by making transmission not voluntary or even deliberate (for an overview of different transmission designs in experiments, see Mesoudi & Whiten, 2008). One resulting issue is that these designs often result in transmission chains that are closed rather than open (Morin, 2015): Instead of the cultural traits spreading freely in a population, typically there is only a straight line of transmission, one to one. I will demonstrate a design making use of an open transmission chain in the study presented in Chapter 4. Another major difference between experimental iterated learning and the concepts of "iterated practice" (Enfield, 2014) or a "cognitive causal chain" (Sperber, 2006) is that the former usually involves transmission of an entire set of behaviors at once (such as a communication system), from one individual to another, whereas in the real world, behaviors are transmitted one interaction at a time, while the overarching system is not expressed entirely.

This working model thus simplifies the real picture a lot. One important detail for the thesis that it does not address is that the model involves processes at different levels of granularity. While a multitude of different levels can be distinguished conceptually and be relevant frames for specific research questions (Enfield, 2014, for example, distinguishes six "causal frames"), a sufficient

distinction for the current thesis is the difference between what has been termed *micro-* and *macro-evolution* (see e.g. Höfler, 2009). Micro-evolution refers to the perspective of an individual innovating, to a fine granularity of single behaviors. Here, the usage context of the innovation can be observed. Note that for communicative signals, this inherently means at least dyadic interaction, however; for Enfield (2014), this perspective would entail both the psychological “microgenetic” frame and the social-interactional “enchronic” frame, which I choose to collapse into micro-evolutionary processes as a whole for the purposes of the thesis. Ultimately, studies at the micro-evolutionary level can inform us about mechanisms involved in ostensive-inferential communication, about how individual dyads create shared conventions, i.e. innovations and their context, and about the mechanisms of transmission.

Macro-evolution refers to the cultural spread of innovations, viewed on a broader scale. This generally means numerous behaviors at once, which govern the transmission and distribution of cultural traits within a larger population (a “diachronic” frame; Enfield, 2014). Here, individual contexts and innovations cannot be observed clearly any more, but large-scale population-level patterns can. As such, macro-evolution is more concerned with group-level behaviors and their dynamics; studies at this level can inform us about the spread of a convention within a population, and the distributional patterns of (potentially multiple) communication systems. One important question is trying to link the individual- (or dyadic-) level processes to the population-level patterns, since typically they cannot be looked at simultaneously (Kandler et al., 2017; Laland & Brown, 2011). In language evolution, this question has been coined as the problem of *linkage* (Kirby, 1999). Cultural evolution via iterated learning is an attempt to bridge this gap by trying to demonstrate how individual behaviors get amplified over repeated interaction to shape the resulting communicative systems at population level (Griffiths et al., 2008; Kalish et al., 2007; Kirby et al., 2014).

A well-established result of transmission chains is that humans are biased to reproduce conventional forms of cultural behaviors through repeated transmission (e.g. Bartlett, 1932). For example, in a transmission chain study tasking participants with copying an initially unconventional drawing by hand, it was found that chains would gradually converge to conventional drawings, such as letters or numbers (Tamariz & Kirby, 2015). Crucially, this only occurred when participants had to memorize the drawings and could not merely copy them from a

permanently available source. This is an example of how a bottleneck in learning (such as memory) can amplify weak biases in the cultural transmission process, guiding the result into a certain direction. This can be simulated even in the absence of actual human behavior, as is the case in the results of agent-based models (e.g. Brighton, 2002; Brighton, Smith, et al., 2005; Kirby, 2002).

Generally, cultural evolutionary studies have successfully shown how different constraints can affect the evolution of communication systems over time (Kirby et al., 2007, 2015; Thompson et al., 2016). Most important for my purposes here is a bias for *simplicity* (e.g. Chater & Vitányi, 2003; A. Clark, 2015; Culbertson & Kirby, 2016; Kemp & Regier, 2012; Zipf, 1949): Under specific pressures, such as the need to keep information stored in memory, it has been found that the complexity of a communication system will get reduced by making it more *compressible*, i.e. describable by a rule that is shorter than simply listing each of the individual traits (Kirby et al., 2015). Similarly, pressure can also lead to the compression of the signals themselves, leading to a simplification of their form (Tamariz & Kirby, 2015; Zipf, 1949). I will apply the general idea, in a different context, in the study presented in Chapter 3.

Since pressures other than the ones for simplicity shape the real-life evolution of communication systems too, compression is not the only possible outcome. This is, in fact, crucial, because mere pressures for simplicity in experiments typically result in systems that have been described as “degenerate”, as they collapse into a single signal for every possible meaning, which is maximally ambiguous and maximally compressible (Kirby et al., 2008, 2015). These systems are not functional for communication. Thus, a common explanation has been that in the actual evolution of communication systems, the pressure for simplification is counteracted by a pressure for expressivity, i.e. the need to keep signals useful for communication by avoiding total ambiguity (Kemp & Regier, 2012; Kirby et al., 2008, 2015). This leads to communication systems that are compressible but do not collapse into a small number of nonfunctional signals; they remain expressive but get compressed in that they also divert from “holistic” signals, which are signals that are used for a single meaning only. The need for expressivity is but one example of how other potential pressures can shape the form of communicative signals towards increasing complexity rather than compression. Kelly et al. (in press) identify a possible countervailing force in their analysis of the historic development of the shape of the Liberian Vai script in a potential pressure for distinctiveness: While the general trend of the script’s graphemes is to move towards increased



compressibility, a minimum complexity is upheld, most likely by the need to keep the graphemes distinct from one another. A counterexample to this general development towards increasing compression in many domains is shown by Miton and Morin (2019), who find that complexity is maintained rather than decreased for frequent European heraldic emblems, most likely due to the need to maintain the iconicity of motifs.

Having summarized the basic mechanisms of the working model of cultural evolution, I want to briefly describe the evolution of *languages*, in plural, as one specific example of the evolution of a communication system under this experimental account and in view of human ostensive-inferential communication taken earlier (compare Scott-Phillips, 2015). This is opposed to the evolution of *language*, i.e. the history of the biological capacity to communicate linguistically (which is not the concern of this thesis). Early ostensive-inferential communication, once this capacity had been established, likely relied on spontaneous iconic signals to a good degree, i.e. signals that have some physical resemblance to the meaning they refer to (Scott-Phillips, 2015). These iconic signs can over time become conventional: As was established earlier, conventions are solutions to repeated coordination problems such as communication (Lewis, 1969). During the process of conventionalization, the signals also tend to become more and more arbitrary, a recurrent finding in artificial language experiments (Garrod et al., 2007, 2010; Healey et al., 2007; see section 1.3 for a summary of the artificial language method and its relation to real-life languages). Here, in particular the interactive use of the signals and the ability to repair miscommunications has been found to be important, a result similarly demonstrated in transmission chains (Tan & Fay, 2011). In other words, if signals are simply created by individuals and their meanings are not used and negotiated in direct dyadic interaction, the drift to the arbitrary does not take place. This is in line with the special turn-taking structure of human interaction outlined in the introduction, and makes sense especially in light of the setting of ostensive-inferential communication.

Once a linguistic convention has been established, it can spread to naïve learners (for an experimental demonstration, see Caldwell & Smith, 2012); this will contribute to the arbitrariness of the convention, since the original context of usage might get lost (Tomasello, 2010). Likewise, since transmission enacts a pressure for simplicity to make the signals more learnable (see above), this can lead to the compression of the signal too (i.e. a loss in the complexity of the signal; Kirby

et al., 2015; Tamariz & Kirby, 2015). This is not unidirectional, however, since the need for successful communication enacts a counter-pressure to keep communicative signals expressive. Efficient conventions do not have to be created by dyads and then spread to a community in a later step either, but can also emerge directly at group level, if interaction partners swap regularly (Fay et al., 2008, 2010). The arbitrary conventions can also gain additional expressiveness through the evolution of grammatical rules, organizing the potential combinations in a language, their order, and other things; combinatorial communication in humans means that we can express more than the sum of the individual signals by chaining conventional signals together (Scott-Phillips, 2015).

It has already been implied but not made explicit that the evolution of real-life languages does not stop at this point, but in fact is permanently ongoing. Linguistic conventions can still shift in meaning, either through metaphor, i.e. the usage of a convention for a novel meaning that is inferable by analogy, or through reanalysis, i.e. the novel interpretation of a convention on the side of the receiver (Hopper & Traugott, 2003). If a convention is used in a particular metaphoric sense frequently or re-analyzed in the same way frequently, it is likely is it that the new meaning will become part of the conventional meaning as well. Just like with the spontaneous creation of novel signals, for this novel usage or interpretation of conventions to be possible, a view such as the one the ostensive-inferential model of communication takes is necessary. Relying on the literal, coded meanings of conventions without the expression and recognition of intentions would mean that semantic change in these conventions would simply make them unreliable, rather than massively expressive (through the use of metaphor, for example). Thus, the inferential abilities that make human communication possible in the first place are also responsible for the subsequent ongoing evolution of languages (Höfler & Smith, 2009; Smith, 2008; Smith & Höfler, 2015).

### 1.3 Methodological Approach of the Thesis

In the thesis, I fully commit to empirical quantitative analyses to approach the questions outlined in section 1.4. The credo here is that in order to understand human communicative interaction, we have to observe and analyze the behaviors in question. In general, the data is provided by participants interacting within the limits of constrained artificial tasks (in the case of two of the empirical studies, these represent *artificial language games*). This reduces the number of possible factors that might impact the interactions and thus makes for data that is easier to interpret than observations of real-life interactions, which are hard to contain. Over the course of the three studies in the thesis, I make use of different methodological approaches that, while committing to the same empirical foundation, differ in their level of granularity and aptness to illuminate specific aspects of the emergence and organization of communicative signals: Classical laboratory experiments to study dyadic interactions and fine-grained micro-evolutionary mechanisms, a large-scale observational online study to investigate unrestricted mass interaction and macro-evolutionary patterns, and a large-scale online smartphone application, combining useful features of the first two approaches. The multimethod approach pursued in this thesis makes use of the distinct advantages of the different methods for studying the different levels of the emergence and organization of communicative signals.

Generally speaking, a laboratory setting (such as in Chapter 2) is ideal when the focus is on the mechanisms underpinning the communication in a task, because it enables full experimental control. This means that many alternative mechanisms that could have produced similar results can be ruled out. There is, however, a limitation to the scope of questions that can be asked and approached via this method, in at least two ways: Interactions in the tasks are forced and artificial and might be quite far from natural interaction, and sample sizes cannot exceed a reasonable number of participants since testing occurs manually, with each pair of participants being supervised individually. One way to overcome these methodological limitations in scale and restrictiveness of experimental designs is to turn to massive data sets of online behavior (Chapter 3). In fact, it has been argued that the people with the best access to data about human behavior are no longer psychologists, but computer scientists (Griffiths, 2015). In exchange for making use of this data, however, a lot of control over the interactions in the task has to be given up on, and the room for interpretations regarding the mechanisms behind the reported results is limited.

One approach that has the potential to combine the useful features of both worlds methodologically – experimental control on the one hand, and less rigid designs with potentially large sample sizes on the other – is that of a large-scale online smartphone application (Chapter 4). Smartphones are widely distributed, powerful devices capable of recording high-quality psychological data (Miller, 2012). A smartphone application is a useful compromise between the two approaches outlined earlier: It allows for very reasonable sample sizes of participants, beyond those of any laboratory experiment, and more diverse than those in laboratory experiments (Dufau et al., 2011). Simultaneously, experimental control is relaxed as compared to the laboratory, but still much tighter than in an unconstrained online environment (Morin et al., 2018).

The thesis makes extensive use of artificial language games to study communication (for reviews of this method, see: Galantucci, 2009; Galantucci et al., 2012; Galantucci & Garrod, 2011; Scott-Phillips & Kirby, 2010; Tamariz, 2017). This concerns Chapters 2 and 4 especially. In general, artificial language games are characterized by requesting from participants to communicate without any pre-established, conventional signals. By experimental manipulation, the studies can answer questions about why certain aspects of communication evolve to be the way they are (Galantucci, 2009). They are not attempting to recreate the origin of language in the laboratory, but rather to investigate how languages might have evolved once they have appeared (Scott-Phillips & Kirby, 2010). To prevent the use of natural language, the tasks challenge participants to associate the signals of a novel *signal space* with the meanings of an artificial *meaning space*. Some examples for the unusual signals are pseudo-words (e.g. Kirby et al., 2008; Winters et al., 2015), spontaneous gesturing (Nölle et al., 2018), graphical signs that prohibit writing (Galantucci, 2005; Healey et al., 2007), and even the movement patterns of a virtual agent (Scott-Phillips et al., 2009). Meaning spaces used in these tasks are usually a set of visual stimuli of some form, varying on color, shape, or movement, but sometimes also simply abstract concepts (Fay et al., 2010; Garrod et al., 2007), information about a spatial position in a virtual environment (Galantucci, 2005; Scott-Phillips et al., 2009), or even music pieces (Healey et al., 2007).

Importantly, we can distinguish different types of experimental designs regarding the evolutionary aspect of the tasks. One way to run the experiments is repeated communication between individuals in the artificial language game (e.g. Galantucci, 2005; Garrod et al., 2007; Nölle et al., 2018; Scott-Phillips et al., 2009; Selten & Warglien, 2007; Winters et al., 2018). This

presents us with conventions between a pair of participants, allowing the researcher to gain insights about the micro-evolutionary mechanisms and patterns developing through repeated interaction alone. Another way to design the experiments is to have no real communicative interaction, but transmit communicative systems to a new generation of learners instead: the classical transmission chain (e.g. Kirby et al., 2008; Perfors & Navarro, 2014; Silvey et al., 2015; Tinitis et al., 2017; Xu et al., 2013). This prevents participants from creating a two-sided, common “language”, but instead focuses on the effects of the repeated learning and transmitting, and thus generational overturn. Other studies have combined the two approaches and involved both repeated interaction and intergenerational transmission, enabling them to study the effects of these processes themselves by adding them systematically (Carr et al., 2017; Fay et al., 2008, 2010; Healey et al., 2007; Winters et al., 2015) or contrasting them directly (Garrod et al., 2010; Kirby et al., 2015).

## 1.4 Thesis Overview

Having introduced the central theoretical concepts and methodological approaches, I now move on with the presentation of the three central questions that I aim to address in this thesis. Each question is associated with one empirical study, represented in Chapters 2, 3, and 4.

### **“How does the shared context affect the successful emergence of communicative conventions, at the micro-evolutionary scale?” – Chapter 2**

The theoretical importance of context for ostensive-inferential communication has been emphasized repeatedly (Clark, 1996; Sperber & Wilson, 1996; see section 1.1), but little empirical work has been done at a stage where conventions are still emerging, rather than at a stage of fully developed human language. Tasks that have studied the use of natural language can inform us about the use of conventions, but not so much about their emergence and the conditions under which communication can be successfully established in the first place. In this study, I present two laboratory experiments that investigate how the shared context, i.e. the amount of information two interlocutors have in common, affects the successful emergence of communicative conventions. In an artificial language game, pairs of participants were tasked with communicating the correct color out of an array of four colors to their partner, using only arbitrary black-and-white symbols. There was repeated interaction between the participants, but no transmission chain. Access to the shared visual context was manipulated for the sender by either showing all four colors or the correct color only to them. The results of both experiments demonstrate that access to the context improves participants' communicative performance. Furthermore, the second experiment shows that participants sharing the visual context adapt their conventions more successfully to novel contexts than those with no access to the context.

These results underline the importance of context in ostensive-inferential communication, empirically demonstrating its direct benefits for successfully establishing novel conventions. In the bigger picture, this chapter demonstrates how the ostensive-inferential model of communication presented earlier can inform the design of an artificial language paradigm to test some of the model's ideas. The study takes a micro-evolutionary perspective at the interactional

level to directly look at the mechanisms behind the emergence of communicative signals, and their conventionalization.

## **“How do cultural traits get organized through interaction, at the macro-evolutionary scale?”**

### **– Chapter 3**

As outlined in section 1.2, through cultural evolution communicative systems will increase in compression, which happens in reaction to simplifying pressures (Kirby et al., 2015; Tamariz & Kirby, 2015). In Chapter 3, I present a study inspired by this phenomenon of increasing compression, investigating it in the context of an online collaborative art project (i.e., not a communicative system in a strict sense). More precisely, I analyze a large-scale online data set recording the interactions of over 1 million individuals that collaborated and competed in shaping the evolving artworks on a 1000x1000 pixel canvas. To do so, participants could only place a single colored pixel at a time, and then had to wait for a fixed period until they could place another one. The main result of the study is that a predictable quadratic pattern of compressibility was present within the evolving visual collaboration, most probably due to a pressure for simplification through increasing competition. This competition is introduced due to the limited available space on the canvas. The trajectory of compression is subsequently shown not to be an outcome of mere increasing homogeneity, but of the evolution of structure throughout the overall canvas.

Overall, the art project shows not only the spread of compressible patterns as a signature of cultural evolution, but additionally highlights the importance of temporal and spatial dynamics in its ongoing development. As outlined in detail in Chapter 3, the online collaboration fulfils the minimal requirements of cultural evolution by undergoing a repeated process of production and transmission of cultural traits (Kirby et al., 2014). In this view, the colored pixels represent the visible behaviors resulting from these cultural traits. Although the art collaboration differs from other cultural phenomena in many ways, it can still rely on the same cultural evolutionary principles. In fact, the nature and scale of the task makes for a unique example that is rather complex, compared to standard cultural evolutionary experiments. The interactions between the participants demonstrate an extraordinary example of naturalistic cooperation and competition in the material sense, i.e. with a pro- or anti-social purpose, because competing directly means taking space from another participant.

The study relates to the organization of communicative signals in the following way: Although individual cultural traits do not always represent communicative signals per se, as there may be many other reasons why they are produced rather than to communicate, it is reasonable to state that the emerging cultural artefact, i.e. the pixel canvas, at least partially also serves a communicative function. There are plenty of examples where participants create artworks as symbolic representations of some common interest, or outright produce fully written language. The main theoretical focus of the study then should be taken to be on how the organization of cultural traits can happen through the massive interactions of participants, both cooperative and competitive. Some of these interactions are also communicative. The study takes a macro-evolutionary perspective, only considering the population-level outcomes of the participants' interactions. Since it can give us some insight into cultural evolutionary patterns overall (with at least some communicative context visible too), there is good reason to argue it can also inform us to some degree about the organization of communicative signals in a stricter sense. Furthermore, by distinguishing between patterns of homogeneity and structure, it sets a useful example that can similarly be applied by cultural evolutionary studies on communication systems.

**“How do existing communicative conventions compare to novel jointly created conventions and their usage in interaction, from a macro- to a micro-evolutionary scale?” – Chapter 4**

Previously, I outlined the role of conventional codes from an ostensive-inferential view of communication (section 1.1): They are solutions to repeated coordination problems (Lewis, 1969) that are not necessary for successful communication under this model, but have the potential to make communication vastly more powerful (Scott-Phillips, 2015). In Chapter 2, I show the importance of shared context for the successful emergence of these communicative conventions; in Chapter 4, I extend on this research by focusing on the semantic structure of artificial conventions and its similarity to natural language structure. One signature feature of language is that it exhibits structure on many levels (Everaert et al., 2015); semantic structure, characterized by the organization of continuous meaning spaces into discrete categories, is one of them.

In this chapter, I investigate the evolution of artificial language semantic structure for color terms, and try to assess its similarity to natural language color terms, as well as its usage with regard to the natural language structure. This is an important question because a potential bias



towards natural language in artificial language games has received little explicit attention so far, and because I can test the influence of color term structure on communicative performance while excluding the benefits that our powerful natural language offers. To do so, I present the results of an artificial language game, similar in design to the one presented in Chapter 2, but this time administered via the medium of a large-scale smartphone application. Importantly, the game allowed not only for repeated interaction between players, but also free partner choice and open transmission to other players, a more realistic feature than in artificial language games that combine repeated interaction with transmission by design (see section 1.3). Comparing the in-game communication of native speakers of English, German, and French to a separate naming task that was carried out online, I demonstrate that semantic structures in the artificial language fit natural language semantic structures to a moderate to good degree. Also, for English speakers, this shared semantic structure affected the performance and pragmatics within the artificial language, a result that could not be replicated for speakers of German or French.

These results link artificial language semantic structure and natural language semantic structure cognitively, however the precise nature of this link remains unclear. The study links the macro-evolutionary with a micro-evolutionary perspective by first observing the color term structure in a language group, then connecting it to the artificial language use in dyadic interaction. This is a demonstration of how distributional patterns of the natural language semantic structure could be connected to direct, communicative interaction. As such, the study ties together results from both earlier studies by investigating both the emergence and organizational structure of communicative signals through interaction.

## **2. The Influence of Shared Visual Context on the Successful Emergence of Conventions in a Referential Communication Task**

This chapter represents the following study published in *Cognitive Science*:

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>

Author contributions: The concept of the study was developed by all three authors, under main supervision by Olivier Morin. I programmed and led the conduction of the experiments. Data analysis and visualization was carried out by me, with help from James Winters. The manuscript was written by me and revised with the help of all authors.

### **2.1 Abstract**

Human communication is thoroughly context-bound. We present two experiments investigating the importance of the shared context, i.e. the amount of knowledge two interlocutors have in common, for the successful emergence and use of novel conventions. Using a referential communication task where black-and-white pictorial symbols are used to convey colors, pairs of participants build shared conventions peculiar to their dyad without experimenter feedback, relying purely on ostensive-inferential communication. Both experiments demonstrate that access to the visual context promotes more successful communication. Importantly, success improves cumulatively, supporting the view that pairs establish conventional ways of using the symbols to communicate. Furthermore, experiment 2 suggests that dyads with access to the visual context successfully adapt the conventions built for one color space to another color space, unlike dyads lacking it. In linking experimental pragmatics with language evolution, the study illustrates the benefits of exploring the emergence of linguistic conventions using an ostensive-inferential model of communication.

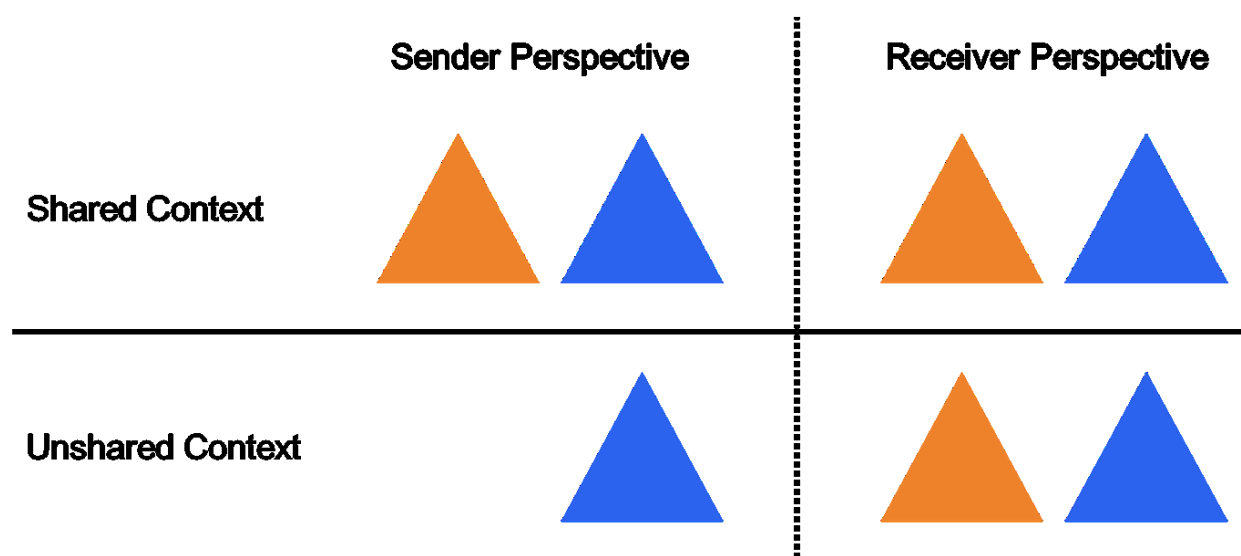
## 2.2 Introduction

An outstanding puzzle for language scholars is that of its *emergence*: How does language come about from pre-linguistic or non-linguistic states (Christiansen & Kirby, 2003; Höfler, 2009; Tomasello, 2010)? Since an important function of language is communication, we can arrive at insights on language emergence by studying human communication in general (Höfler, 2009). Linguistic communication, in this view, is a special case of communication enriched by a “structured collection of conventional codes” (Scott-Phillips, 2015, p. 20). Based on this, it has been argued that the human capacity for *ostensive-inferential communication* (Sperber & Wilson, 1996) is what allows complex languages to evolve (Scott-Phillips, 2015), with our pragmatic capacity being the cognitive foundation for all of semantics, morphology, syntax, and phonology (Scott-Phillips, 2017). In ostensive-inferential communication, the sender of a message provides evidence for an intended meaning, while the receiver interprets this evidence (Sperber & Wilson, 1996). Imagine, for instance, you are listening to music over your headphones, when your friend Barbara wants to tell you something. To signal this, she makes eye contact with you, puts her hands to her ears, and mimics taking off the headphones in a slow and stylized way. Even though this specific signal might have never been used before in your shared conversational history, inference may suffice to interpret and understand the underlying message.

This interpretation of signals cannot happen in a vacuum, however; the possible meanings would be practically unlimited (Berwick et al., 2011). Consider the example outlined above again: Had you not been wearing the headphones (or had you not been aware that you were still wearing them), Barbara’s gesture would have left you quite puzzled indeed. Alternatively, suppose that after taking off the headphones, she asks you “Are you listening to this album?”, whilst holding up a copy of “Led Zeppelin”. Here, the demonstrative “this” could have referred to virtually any album in existence, had the meaning not been clarified by the additional visual information. As proposed in classic theories of communication (Clark, 1996; Grice, 1989; Lewis, 1969; Sperber & Wilson, 1996), interlocutors have to create and interpret messages according to their *context*, a wide range of information that includes the time and place of a message’s utterance, the interlocutors’ previous conversational history, and more.

Context is a notoriously vague notion, which has been operationalized in various ways (see e.g. the differences between Clark’s “common ground” compared to Sperber and Wilson’s “mutual

cognitive environment”). In this study, we investigate the immediate *shared context* (cf. Fig. 2.1). Following Winters, Kirby, and Smith (2018), it is defined as the amount of relevant knowledge that the interlocutors have in common. Unlike “common ground” (Clark & Carlson, 1981), shared context does not require explicit mind reading of the “I am aware that she is aware that I am aware...” type. Unlike Sperber & Wilson’s mutual cognitive environment, it is restricted to information that is actually present to each interlocutor’s mind, as opposed to information that is merely accessible. Both are important dimensions of context that we choose to disregard for the purpose of this study. To further simplify the issue, this paper focuses on the shared knowledge that two interlocutors have of their environment, leaving aside other aspects of shared knowledge such as shared membership in a community (cf. Clark, Schreuder, & Buttrick, 1983) or shared discourse histories (Barr & Keysar, 2002; Clark & Wilkes-Gibbs, 1986).



**Fig. 2.1. Illustration of the shared context for the sender and receiver of a message in a referential communication situation.** The sender in the situation is tasked with communicating the blue triangle to the receiver. The situations differ only in the availability of contextual information to the sender. With access to the shared context, the sender might refer to the triangle as “the blue triangle”, whereas without it simply “the triangle” would be sufficient from their perspective.

In line with other studies, we focus specifically on shared visual information. For two interlocutors, having a piece of visual information in common impacts communication. On the sender’s side, shared visual information allows for audience design – the tailoring of a message to its receiver’s

state of knowledge (Brennan & Hanna, 2009; Galati & Brennan, 2010; Holler & Wilkin, 2009; Isaacs & Clark, 1987; Krauss & Fussell, 1991). This is a crucial aspect of the interactive alignments that characterize conversation (Garrod & Pickering, 2004). On the receiver's side, shared information is more likely than non-shared information to be taken into account when interpreting a message (Hanna et al., 2003).

These studies, and others like them studying the impact of shared information on communication, are based on natural language. This has two consequences. First, communicative success is usually at ceiling: Any participant can solve a simple referential communication task using words, regardless of the amount of shared information (e.g. Brennan, 2005). This makes it difficult to gauge the impact of shared information on communicative success (but see Clark & Krych, 2004; Schober & Clark, 1989; Sulik & Lupyan, 2018). Second, experiments conducted in natural language are appropriate to study the use of linguistic conventions, rather than their emergence. Conventions, whether they are linguistic or not, are solutions to repeated coordination problems (Lewis, 1969), such as referential communication (Millikan, 2005; Skyrms, 2010). They are at least partly arbitrary forms of behaviors that are sustained by the weight of precedent as opposed to any intrinsic aptness. This paper aims to investigate the impact of shared information upon the emergence of novel conventions.

### **Artificial Language Experiments and the Emergence Problem**

In the past, the emergence problem of language has been addressed in the laboratory using artificial language experiments. Here, the general idea is to study interactions occurring without the presence of established communicative conventions (for studies reviewing this field, see Galantucci et al., 2012; Galantucci & Garrod, 2011; Scott-Phillips & Kirby, 2010; Tamariz, 2017). Several studies have focused more closely on the form of the emerging conventions themselves, i.e. the shape that the signals take, and their evolution (e.g. Galantucci, 2005; Garrod, Fay, Lee, Oberlander, & MacLeod, 2007; Healey, Swoboda, Umata, & King, 2007; Scott-Phillips, Kirby, & Ritchie, 2009). For instance, Garrod et al. (2007) showed that drawings emerging *de novo* in their “Pictionary”-style experiments become simpler and more symbolic with repeated interaction, while Healey et al. (2007) focused on the extent to which drawings were abstract or iconic when

participants were tasked with drawing music for each other. In comparison, the present study investigates the emergence of conventions in relation to communicative success.

The contextual circumstances under which conventions arise have not been considered in the studies above. The exception to this is the “embodied communication game” by Scott-Phillips et al. (2009), who found that “the establishment of [a] default convention provides the common ground from which a signal may be created and inferred” (p. 233). In this experiment, participants had to communicate their position on a 2x2 grid. As there was no established communication system for completing the task, participants had use a repertoire of basic actions to signal their intention to communicate (and to then use this as a means to derive a conventional signaling system of conveying meaning). Our study differs in this respect as it focuses on the immediate shared context (as opposed to the ability of participants to leverage their shared discourse history). Additionally, Scott-Phillips et al. (2009) did not investigate the role of context directly by manipulating it experimentally. Nevertheless, the study opens up the interesting question of how conventions multiply in relation to the shared context, which is what we also try address in this study as a secondary question. Presumably, conventions form from repeated successful usage of symbols building on the shared context (Höfler & Smith, 2009), until they become sufficiently entrenched (Langacker, 1987) and can be used as a contextual basis for novel inferences themselves (as in the case of Scott-Phillips et al., 2009). As such, we should expect the shared context to facilitate the establishing of more conventional symbols, with more conventions leading to more successful communication in turn.

In contrast to this first line of research, previous experiments have investigated contextual effects systematically, but focused on the further development of conventions after they have been established (through a training phase in the task). Several studies have demonstrated that artificial languages will optimize to the semantic dimensions relevant in context. Through simulating *iterated learning* in experiments, defined as the “process in which an individual acquires a behavior by observing a similar behavior in another individual who acquired it in the same way” (Kirby et al., 2008, p. 10681), artificial languages have been shown to develop: i) underspecification with regard to irrelevant dimensions in a reference space (Silvey et al., 2015), ii) overspecification when relevant dimensions are difficult to discern (Tinits et al., 2017), and iii) either underspecified, holistic, or systematic linguistic structure depending on their contextual

niche (Winters et al., 2015). These studies took the task's immediate perceptual context into account, but did not manipulate the extent to which it was shared or not. Still, the general observation that specific types of context will bias artificial languages to develop a certain structure leads us to another secondary question. We want to investigate how flexibly conventions can adapt when a change in contextual environment occurs: their generalizability. As the shared context should allow interlocutors to be more successful, it might also lead to more generalizable conventions emerging from their conversation.

Winters et al. (2018) specifically considered the shared immediate context in the artificial language paradigm. Using a referential communication game setup, participants were first trained in an "alien language" consisting of random syllables mapped onto a small set of referents, and then used this alien language to communicate about referents they learned as well as novel ones. Both the shared context and the generalizability of the immediate context to future contexts were manipulated. Crucially for our purpose, shared contexts fostered languages that required more contextual enrichment for interpretation than non-shared contexts. Of special interest to us are their performance results that indicate higher levels of communicative success in the shared context conditions. However, the interpretation of these results is limited, since the effect is driven by one condition that is at ceiling, while performance in the other shared context condition was as low as in the unshared conditions. Additionally, the study used a training regime to make participants learn the starting language in the experiment; a commonality it shares with all iterated learning studies mentioned above, and with another line of evidence which demonstrates the influence of shared context on the choice of referral expressions in a word-learning paradigm (Craycraft & Brown-Schmidt, 2018; Gorman et al., 2012; Heller et al., 2012; Wu & Keysar, 2007). In all such studies, participants are provided with pre-established mappings between the artificial language's symbols (e.g. the strings of syllables in Winters et al., 2018) and the corresponding referents. This makes it difficult to study the emergence of conventions. To resolve these issues, the present study removes the training from the procedure, and allows participants to freely associate and create mappings from the start.

## Referential Communication Tasks and Interaction

At the core of our task is a *referential communication* paradigm. These tasks have been traditionally used in experimental pragmatics, dating back at least to Krauss and Weinheimer (1964). The basic premise is that a “sender” (alternatively, “director” or “speaker”) has to communicate a target object to a “receiver” (also known as “matcher”, “listener”, etc.), using natural language. In our task, participants take on the role of either sender or receiver, with no role reversal (for a study experimentally investigating the effect of role reversal on conventionalization, see Moreno and Baggio, 2015). The task consists of using black-and-white symbols to convey and identify colors (used as referents). The domain of colors has been of particular interest to studies on language evolution ever since the classic work by Berlin and Kay (1969), and has been proven to be useful as a reference space in pragmatic experiments as early as Krauss and Weinheimer (1967). Early results using the referential communication paradigm include the fact that the descriptions become shorter with increasing conversational history (Krauss & Weinheimer, 1964; Krauss & Weinheimer, 1966; Clark & Wilkes-Gibbs, 1986; Schober & Clark, 1989; Wilkes-Gibbs & Clark, 1992), but longer when referents are more similar (Krauss & Weinheimer, 1967) or when describing the referents for someone else as opposed to oneself (Fussell & Krauss, 1989). In these early studies, the intent of the research designs was to study linguistic communication in live interactions.

Later on, the focus shifted from conversational history to perceptual context, especially in the visual modality. These studies have typically been using eye-tracking in a task that involves a director instructing participants how to move objects around in a grid. Crucially, the objects are not always perceivable by both participants: Relevant items may be omitted from the director’s view, or the two participants may be given access to partially different sets of items. Initial studies interpreted their findings as demonstrations of failures in matchers’ usage of the context (Horton & Keysar, 1996; Keysar, Barr, Balin, & Paek, 1998; Keysar, Barr, Balin, & Brauner, 2000), but this was contested by studies showing either an impact of the shared context on utterance comprehension, methodological problems, or both (Brown-Schmidt, 2009; Brown-Schmidt et al., 2008; Hanna et al., 2003; Heller et al., 2008; Nadig & Sedivy, 2002). Summarizing the debate, Brown-Schmidt (2012) states that it mainly revolves around the timing of reference resolution, which is not a central concern of our study. However, similar to this line of research, we will focus



on the perceptual aspect of the shared context. The term we will use is *shared visual context*: The context is shared because we make it clear that interlocutors get access to the same information (or have restricted information, in our other condition), and only limited to the visual modality.

One advantage of pragmatic paradigms using natural language is that participants are often allowed to interact more freely. In particular, they are able to *repair* misunderstandings that might arise in the conversation, and will do so until they arrive at an acceptable interpretation (Clark & Schaefer, 1987). What remains poorly understood is the role that interaction might play in the early stages of emerging communication, especially in the absence of external feedback (i.e., information on the success or failure of communication, given by an outsider to the conversation, and usually provided by the experimenter, or programmed by them into the protocol's program). Most of the studies that trained participants on the initial conventions in the task have also relied on feedback provided systematically by the experimental setup (e.g. Kirby et al., 2015; Winters et al., 2015, 2018) or did not involve interaction between participants at all (Kirby et al., 2008; Silvey et al., 2015; Tinitis et al., 2017), whereas studies specifically concerned with the form of emerging conventions typically also privilege repair (Garrod et al., 2007; Healey et al., 2007; Scott-Phillips et al., 2009). In a similar fashion, we included basic repair mechanisms into the experiments presented here (even though we acknowledge we cannot comprehensively cover the extent of repair strategies used in real-world communication) and refrain from providing other types of feedback to the participants.

## **The Current Study**

The goal of this study is to show that the shared visual context is important for the successful emergence and use of communicative conventions. To test this idea, we conducted two referential communication experiments. Dyads of participants were tasked to accurately communicate the correct color out of an array of four colors by using novel symbols. These symbols were limited to a predetermined set of black-and-white visual signals (and combinations thereof) and some pre-established repair signals. Participants received no training on symbol meanings and no feedback by the experimental setup at any time; they had to make inferences about the most likely correct answer given the evidence, whilst they could use the repair mechanism to clarify or request further

explanation. We manipulated the shared visual context between dyads by giving the sender access to the distractor colors in only one condition.

The main hypothesis is that, overall, the access to the visual context will influence performance in the task. Based on this, we predicted dyads in the shared visual context condition would outperform those in the unshared visual context condition (prediction 1). Furthermore, we expected pairs to make progress in performance over the time course of the experiments, as they jointly create novel conventions for communication (prediction 2). If the pairs in the shared visual context condition also subsequently profit more from building on these conventions, we should see even faster progress in that condition (prediction 3). These three predictions are tested in both experiments. In experiment 2, we also consider the secondary hypotheses outlined in the introduction. Specifically, we ask whether the shared context would also lead to more numerous conventions, and better generalization to different contexts (predictions 4 and 5).

### **Ethical Approval and Preregistration**

Both experiments received approval by the ethical committee at the FSU Jena before they were conducted. All our predictions and sample sizes were preregistered on the Open Science Framework in advance. For experiment 1, this happened before data collection was underway; for experiment 2, due to a technical malfunction, the registration occurred after 3 of the 48 pairs had been tested, but no changes were made to the preregistration document in the meantime. The registrations can be accessed at <https://osf.io/rbhk2/> (experiment 1) and <https://osf.io/tn6e8/> (experiment 2).

### **2.3 Experiment 1**

Experiment 1 sought to establish the paradigm and provide a first test for our main hypothesis.

## Method

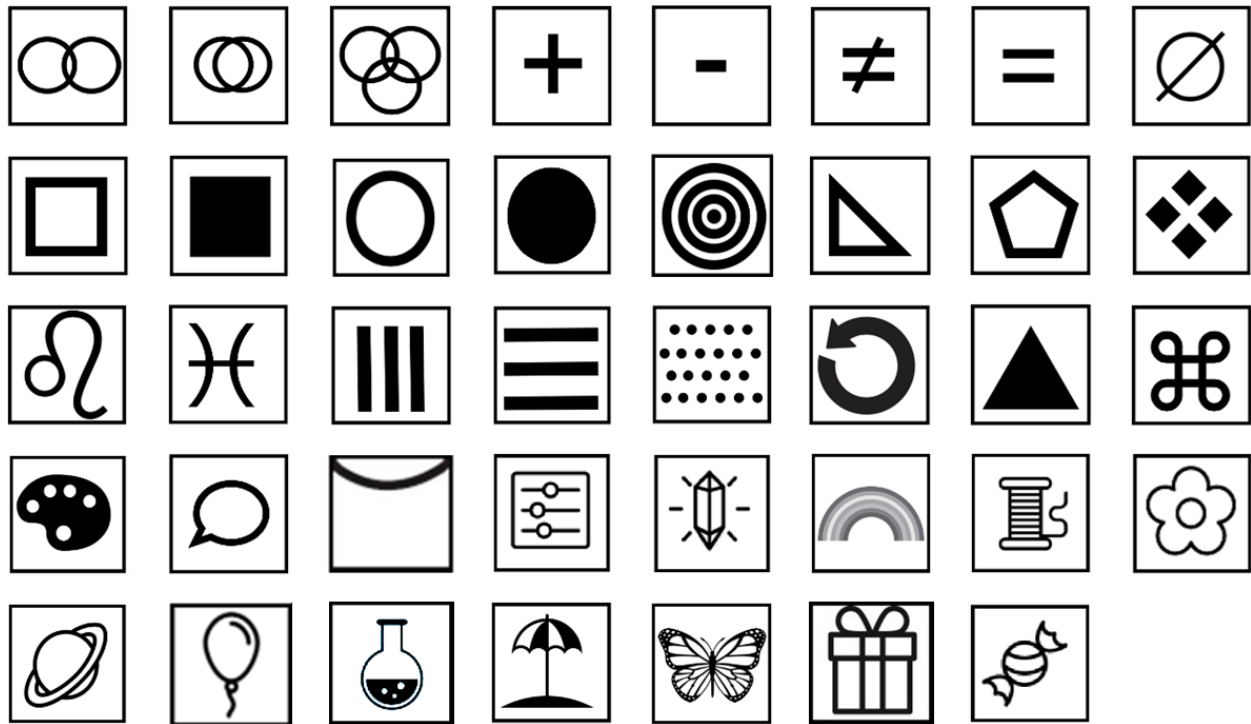
### Participants

52 participants (50 of which were students) were recruited and invited to play the “Color Game” in the laboratory. Their mean age was 23 ( $SD = 3.5$ ); 35 were female and 17 male. All participants were fluent speakers of German, and all but two participants reported German as their native language. Before the main procedure, the Ishihara test for color blindness (Ishihara, 1972) was administered, since the meaning space in the experiment consisted exclusively of colors. All participants showed typical color vision in the test.

### Materials

The meaning space of the experiment consisted of a continuous HSL color space (H, S, and L describing the colors in terms of Hue, Saturation, and Lightness, respectively), going full circle in  $360^\circ$  of hue. The saturation and lightness parameters were kept constant. For practical purposes, we constructed a total of 360 colors in this way (with a constant distance of  $1^\circ$  in hue, which makes neighboring colors indistinguishable). From this space, color arrays were constructed by randomly drawing a color and then choosing the three next colors with a fixed spread value of  $45^\circ$  in hue, respectively. Thus, the first color was at a distance of  $45^\circ$  from the second,  $90^\circ$  from the third, and  $135^\circ$  from the fourth color. As can be seen, the domain of color provides us with a flexible continuous meaning space that can be divided in certain ways, allowing some control over the relations between the discrete referents as well as creating similar portions of the total space participants experience (cf. experiment 2).

For the signal space, participants were presented with a selection of 39 pre-constructed black-and-white symbols (see Fig. 2.2) to choose from. The symbols had been selected on the level of ambiguity, such that they could become associated with several different colors. For instance, the “crystal” symbol in the second to last row of Fig. 2.2 could be treated as a gemstone of any color imaginable. Participants were not trained on any meanings that the symbols might have and saw them for the first time just before the experiment started. This and the arbitrariness of the symbols regarding their relation to specific colors ensured that participants had to form new conventions over the course of the experiment.



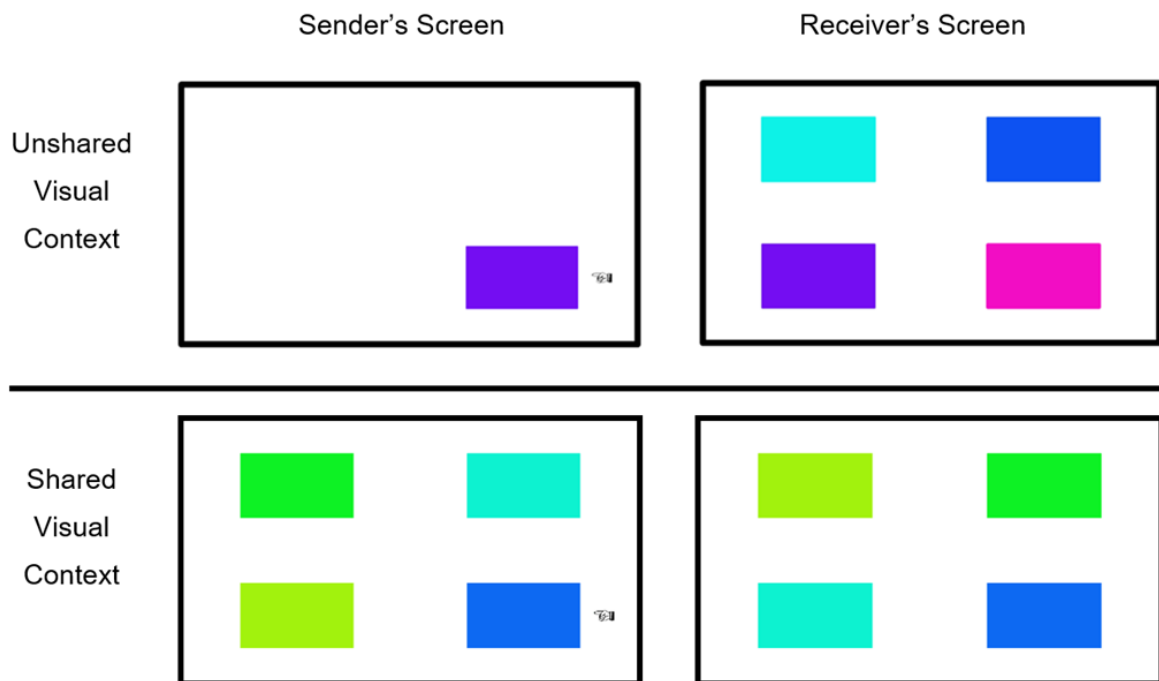
**Fig. 2.2. List of symbols available to senders in experiment 1.** The symbols can be roughly categorized as logical symbols, abstract shapes, and symbols depicting real-world objects (from top to bottom). Crucially, all symbols were chosen to be ambiguous with regard to their association with colors.

**Procedure**

Participants were randomly paired in dyads and randomly assigned the role of sender and receiver for the entire duration of the experiment. To minimize knowledge about each other, players were seated in separate sound-proof rooms during the experiment, and one participant was scheduled to arrive to the experiment 15 minutes earlier than the other participant to avoid contact between them. Upon arrival, participants read the general participant information, gave informed consent regarding the experiment, and completed a short demographic questionnaire. After that, the Ishihara test for color blindness (Ishihara, 1972) was administered. Just before the experimental task, participants read printed instructions that explained the rules of the game to them, and they were allowed to ask questions for clarification. Since this study is concerned with the emergence of communication, they then proceeded immediately with the first trial of the experiment, without a training phase.

## Experimental Task

On any given trial in the game, a random array of four colors was constructed in the way described above. Out of the four colors, a target color was chosen randomly and conveyed to the sender by a pointing finger next to it on the computer screen (for example screens, see Fig. 2.3). The sender's task always was to communicate this target color using only the symbols of the signal space; the goal for the receiver was to choose the correct color out of the array of four colors. Communication was only possible via a whiteboard application (Baiboard, created by Lightplaces Ltd.) running on two iPads that the participants were using. The iPads were connected by WLAN, allowing for live synchronization of changes made by the participants and for observation by the experimenter via a third connected iPad. Senders were free to arrange the symbols on the canvas in whatever position they wanted, and there was no limit on combinations or the number of symbols sent. The exception was that symbols were not allowed to overlap, as this would have prevented correct analysis of the messages.













**Fig. 2.3.** Example trials in the two conditions with the corresponding screens for sender and receiver. In the top example (*unshared visual context*), the sender has to communicate the purple color, and in the bottom example (*shared visual context*), the dark blue color.

Importantly, the receiver could not only passively watch the sender's message being created, but was also allowed to repair unclear messages using three simple responses that had been introduced to both players before the game started: Drawing a circle indicated that the message had been understood; drawing an arrow was used as a prompt for clarification; and drawing a schematized hourglass meant the receiver thought that the participants were running out of time in the current trial. Likewise, the sender had two possibilities to evoke feedback from the receiver: Sending a question mark was used as a prompt for the receiver to indicate whether the message was clear, and sending an exclamation mark was an indication that the message was complete for the sender. All of these possible responses can be seen in the example trials presented in Fig. 2.4. In this way, we wanted to allow for interaction in the task, using rules that were the same for every pair; apart from this, there was no direct feedback that indicated correct or wrong answers, nor were senders informed about the receiver's color choice at the end of the trial. This made sure that the possibility of receivers learning their sender's code by mere memorization of correct answers, on a trial-and-error basis, was minimized. Instead we wanted them to infer the sender's intended meaning and communicate about what they could understand and what they could not.

As displayed in Fig. 2.3, the experimental manipulation concerned the number of colors the sender knew about, which was either one (i.e. the sender sees only the target) or four (i.e. the sender has knowledge of the whole array). The receiver always saw all four colors. Participants were informed in the instructions about how many colors their partner in the experiment would see. The conditions varied between the 26 pairs of participants. Thus, 13 dyads experienced shared visual context (*shared* condition) and 13 dyads unshared visual context (*unshared* condition). The only technical change in the unshared condition was that the three distractor colors were removed from the sender's array, meaning that the target color still appeared in a random position in every trial and was marked by the finger, as in the shared condition. A consequence of this is that the sender saw a single color which was always present in the receiver's visual context.

In total, participant pairs were presented with 64 experimental trials, divided into eight blocks (of eight trials) that were separated by a short pause, respectively. After the main experimental task, participants completed a short questionnaire in which they listed all the symbols they remembered from the main experiment (i.e. in free recall) and their corresponding meanings (suspected meanings, in case of the receiver). Finally, participants were paid 10€ plus up to 6€,

depending on their success in the task, in compensation. Completion of an average experiment took between one and two hours in total.

|   | t   | t + 1   | t + 2  | t + 3   |
|---|---|---|--|---|
|  |  |  |  |  |
|  |  |  |  |  |

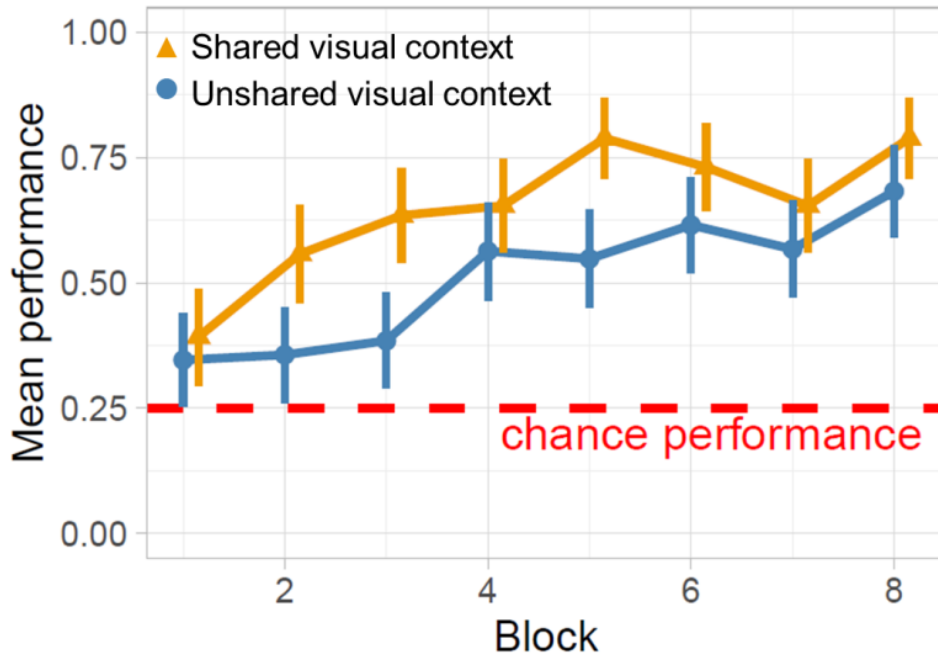
**Fig. 2.4. Two frame-by-frame examples of communication during a single trial in each condition, respectively.** The examples correspond to the screens presented in Figure 5.3. In the example for the unshared visual context condition (top), the receiver does not understand the “candy” symbol sent at time  $t$  and indicates so by drawing an arrow ( $t + 1$ ). The sender then tries to elaborate further by adding a “rainbow” symbol ( $t + 2$ ), but ultimately communication fails and the receiver indicates that the pair is running out of time (hourglass at  $t + 3$ ). In the example for the shared visual context condition (bottom), the sender specifies the precise meaning of “dark blue” using a combination of 3 different symbols ( $t$  until  $t + 2$ ). At  $t + 3$ , the receiver indicates they understand the intended message by drawing a circle.

## Results

### Does the shared visual context improve communicative success and do pairs improve over time?

Before the analysis, six trials (0.4% of the total sample) were excluded from the data for the following reasons: Three trials had reaction times below three seconds, probably due to accidental button presses; and three trials were lost due to a WLAN crash. All analyses were conducted using R version 3.4.0 (R Core Team, 2017).

The mean accuracy, which is the proportion of trials with correct answers by receivers, across both conditions was  $M = 0.58$  ( $SD = 0.49$ ). There were individual differences between dyads, with the lowest scoring pair only reaching  $M = 0.31$  and the highest scoring pair reaching  $M = 0.86$  in accuracy. In the subgroups of the shared and unshared conditions, the mean outcome was higher for the dyads with shared visual context ( $M = 0.65$  and  $M = 0.51$ , respectively). Fig. 2.5 illustrates the mean development over time in the two conditions.



**Fig. 2.5. Development of performance over time in experiment 1, in blocks of eight trials.** Error bars represent 95% confidence intervals. In both conditions, the mean trend is that pairs started out slightly above chance and generally improved in performance. However, the performance in the shared condition is elevated, compared to the unshared condition.

To test the predictions regarding shared visual context and communicative success (predictions 1 to 3), a logistic mixed effects model with the accuracy outcome was constructed. First, the two predictor variables were centered to remove collinearity between the main effects and the interaction, and thus to make parameters interpretable as the total main effects (Schielzeth, 2010). Then the model was estimated using the R package *lme4* (Bates et al., 2015). More precisely, accuracy was predicted by shared visual context (dummy-coded with 1 being *shared*), trial



number, and their multiplicative interaction, while the maximal random effects structure was included in the model (cf. Barr, Levy, Scheepers, & Tily, 2013). This maximal structure consisted of random intercepts for pairs and random slopes for trial number.

There were significantly positive estimates for the effects of shared visual context and trial number, but not for their interaction (see Table 2.1). This means that participants' performance was significantly better in the shared condition (prediction 1), and significantly better the more trials they had played in the game (prediction 2); however, participants in the shared condition did not progress faster in their overall performance (no evidence for prediction 3).

**Table 2.1. Estimates and p-values for the accuracy model in experiment 1.**

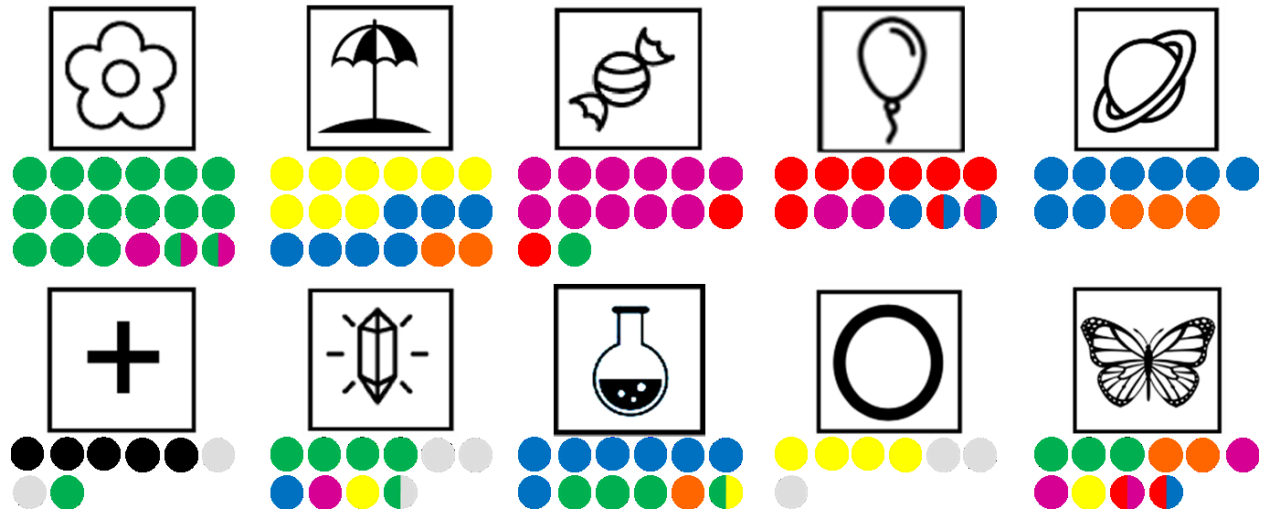
| Fixed effect                         | $\beta$ | SE    | $p^a$            |
|--------------------------------------|---------|-------|------------------|
| Intercept                            | 0.39    | 0.12  | <b>.002</b>      |
| Trial number                         | 0.03    | 0.003 | <b>&lt; .001</b> |
| Shared visual context                | 0.69    | 0.25  | <b>.005</b>      |
| Trial number * Shared visual context | 0.002   | 0.007 | .733             |

<sup>a</sup>p-values < .05 are marked in bold.

### Exploration of the Questionnaires

To get a better sense of how participants used the symbols in the task, we inspected the symbol lists they created after completing the main experiment. As can be seen exemplarily for the ten most frequently reported symbols in Fig. 2.6, symbols were used by most senders as substitutes for color terms. However, there were also other reported meanings, such as symbols indicating subjective brightness, mixing of colors, correct and wrong answers, or even the fact that the target

was a “basic” (i.e. primary) color. Different conventions arose in different pairs, and in fact no symbol was exclusively used for one color only.



**Fig. 2.6. Meanings of the ten most frequently reported symbols, as recalled by the senders.** Every dot below a symbol stands for one sender reporting that the symbol was used for the respective color. Double-colored dots indicate that the sender reported using it for two different colors, and black or light grey dots mean dark or bright colors. In addition to the meanings presented in the figure, senders reported using the “plus” symbol on the left for mixing colors in five cases, and the “circle” symbol on the right for “basic colors” in one case.

As a proxy for conventions, we counted the number of cases in which the sender and receiver of each pair reported the same symbol and agreed on the same meaning for it. This analysis should be seen as supplementary, as there was some vagueness involved in the free descriptions provided by the participants. Moreover, the free recall meant that the players did not necessarily remember the same symbols after the game. For these reasons, we refrained from computing more than descriptive values for the agreements and merely wanted to get a first indication about how many conventions arose in the two conditions. On average, pairs in the unshared condition agreed on  $M = 4.08$  ( $SD = 2.28$ ) symbol meanings and pairs in the shared condition agreed on  $M = 6.15$  ( $SD = 3.18$ ) symbol meanings.

## **Discussion**

The results of experiment 1 suggest that the shared visual context is helpful for the successful emergence of communication. Dyads in the shared condition outperformed dyads in the unshared condition. Participants managed to communicate above chance, and overall performance increased over time, indicating the formation of novel conventions. Participants mostly formed conventions for symbols by mapping them to color categories, although some were also used for the mixing of colors, lightness, or even more abstract meanings. Descriptively, these conventions also seemed to be more frequent for pairs in the shared condition.

However, this last result was merely explorative, lacking a rigorous test. We also did not investigate the generalizability of conventions in the two conditions. In addition, some methodological considerations can improve the experimental paradigm. For instance, the colors used as targets in the experimental trials (randomly chosen from a 360° space) might differ in their difficulty, limiting our control over this variable. The symbol space was also quite large, with some symbols clearly outperforming others and becoming very popular, while some were rarely used. We tried to address these issues in experiment 2.

### **2.4 Experiment 2**

Experiment 2 set out to replicate the main results of experiment 1 with a larger sample size, improving on the paradigm in several ways, especially with regard to the meaning and signal spaces. Additionally, we aimed to test for more frequent conventions in the shared condition more rigorously by including a systematic questionnaire at the end of the experiment (prediction 4). Lastly, we predicted that conventions in the shared condition should also become easier to generalize and use in a new referential context (prediction 5), because we expected communication to be more successful. We address this by switching to a different color space after the first half of the experimental task; assuming that it is functional for successful communication to re-use symbols, symbols from the first half of the experiment should be re-used more often in the shared condition.

## Method

Where not explicitly mentioned in the subsequent paragraphs, the experimental design was the same as in experiment 1.

## Participants

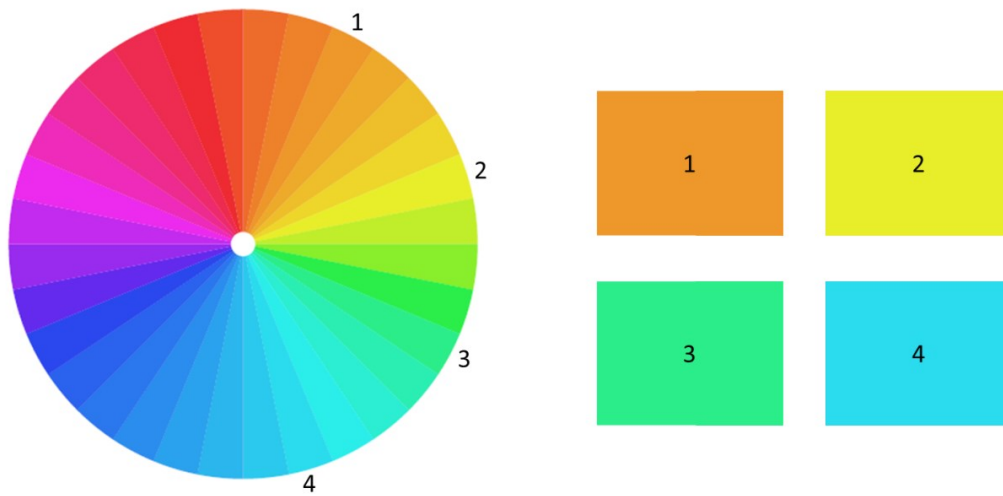
This time, 96 participants (89 students) were recruited. Their mean age was 24 ( $SD = 4.0$ ); 81 were female and 15 male. Since eleven participants did not report German as their native language, we made sure (before the experiment) that all participants were fluent speakers of German and had no problems understanding the printed instructions. All participants showed typical color vision in the test for color blindness, and none had taken part in experiment 1.

## Materials

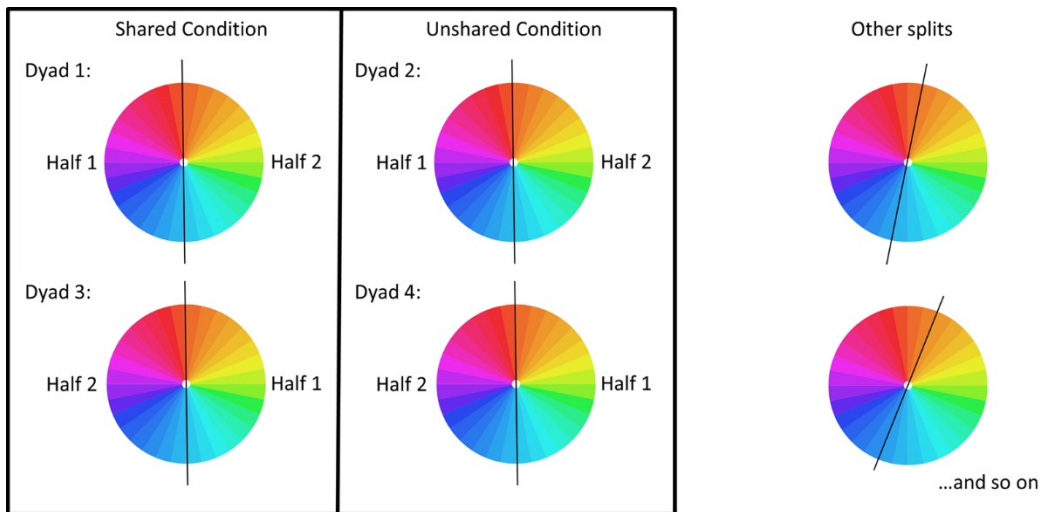
In this experiment, we improved our control over the difficulty of target referents (and their arrays) by using an artificially discretized set of colors. Similar to the first color space, physical saturation and lightness were kept constant and colors were only varied in hue. However, this time we chose a discrete space of 32 colors by applying the CIE2000 formula (created to reflect perceptual differences; cf. Luo et al., 2001) to create a circle of perceptually equidistant colors, each in a distance of  $\Delta E = 7.8$  ( $\Delta E$  representing the distance between colors in the CIE2000 space) from their two neighbors (see Fig. 2.7). Additionally, this color space was split in half for each dyad to allow testing for the generalization of conventions. This resulted in two half-circles each representing the color space for one half of the experiment, respectively. Since the location of this split in the color circle was arbitrary and might influence the results of the experiment, dyads started with different halves of the space, in total reflecting the full spectrum. This was counterbalanced between conditions (for a visualization, see Fig. 2.8).

The signal space in experiment 2 consisted of a subset of the symbols used in experiment 1; we removed those symbols that were used barely or almost constantly, leaving us with 23 symbols in total (see Fig. 2.9). The reasoning behind this was to remove the least ambiguous symbols

(enabling rather similar and easy communication) and the least useful symbols (with very low usage numbers, making them less comparable).



**Fig. 2.7. Left: All 32 colors comprising the color space in experiment 2.** The space can be split in half by drawing a straight line at any border between colors. **Right: A color array created from this space.**



**Fig. 2.8. Left box: Counterbalancing the color space between conditions.** Each split occurred four times: Once in each condition, and once in regular and reversed order. **Right: Visualization of how the space can be split in different ways.** Note that when the line has traversed half of the circular space ( $180^\circ$ ), we arrive at the same splitting pattern as at the example on the left ( $0^\circ$ ).

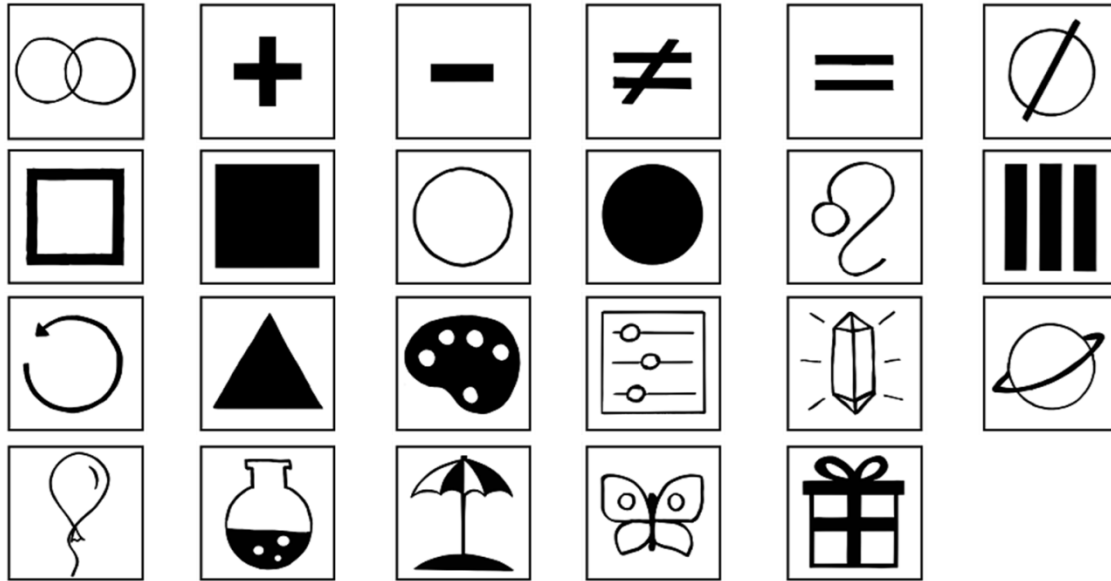


Fig. 2.9. List of symbols available to senders in experiment 2.

### Procedure

The procedure closely followed the design outlined for experiment 1.<sup>1</sup> 24 pairs each played in one of the shared visual context conditions. Without notice to the participants, the color arrays presented in the second half of the experiment (i.e. the second set of 32 trials, or the last 4 blocks out of 8) were switched and only drawn from the half of the color space the dyad had not encountered previously. Because the meaning space was discrete, we counterbalanced the color arrays presented and the targets chosen from them, such that each color appeared as the target twice. After completion of the experiment, an “alignment questionnaire” was handed to both participants, in which they had to tick a description for each of the 32 colors presented in the experiment.

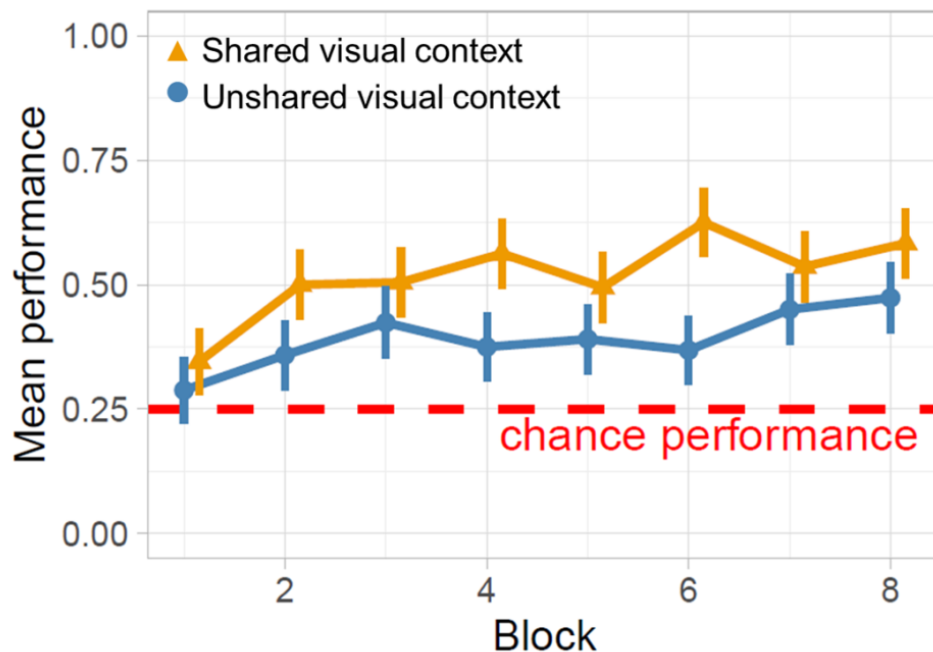
---

<sup>1</sup> There were two different experimenters this time, who followed the same parallel procedure when conducting the study.

## Results

### Can the results of experiment 1 be replicated?

Before the analyses, 15 trials (0.5% of the total sample) were excluded from the data for the following reasons: 13 trials were lost due to crashes, and two trials had reaction times below three seconds (same guideline as in experiment 1). Compared to experiment 1, the task was slightly harder: The mean accuracy across both conditions was  $M = 0.46$  ( $SD = 0.50$ ). This time, the lowest scoring pair only reached  $M = 0.17$  in accuracy, while the highest scoring pair reached  $M = 0.80$ . In the subgroups of the different conditions, shared pairs were again better on average than unshared pairs ( $M = 0.52$  and  $M = 0.39$ ). Fig. 2.10 illustrates the development of performance over time.



**Fig. 2.10. Development of performance over time in experiment 2, in blocks of eight trials.** Error bars represent 95% confidence intervals. Again, pairs in both conditions generally improved in performance, but performance in the shared condition is elevated. Pairs in the shared condition showed decreased performance in block 5 (right after the change in color arrays), but recovered in the remaining blocks.

To replicate the results of experiment 1, a logistic mixed effects model with trial accuracy as the outcome variable was constructed, following the analytic strategy of the previous model. This

time, we were able to also control for random effects of the color arrays in addition to random participant effects because of the discrete color space. There were significantly positive estimates for the effects of shared visual context and trial number, but not for their interaction, replicating the results of experiment 1 (see Table 2.2).<sup>2</sup> This means that participants' performance on the accuracy outcome was significantly better in the shared condition (prediction 1), and significantly better the more trials they had played in the game (prediction 2). Again, participants in the shared condition did not progress faster in their overall performance (no evidence for prediction 3).

**Table 2.2. Estimates and p-values for the accuracy model in experiment 2.**

| Fixed effect                         | $\beta$ | SE    | $p^a$            |
|--------------------------------------|---------|-------|------------------|
| Intercept                            | - 0.18  | 0.09  | .059             |
| Shared visual context                | 0.56    | 0.17  | <b>&lt; .001</b> |
| Trial number                         | 0.01    | 0.003 | <b>&lt; .001</b> |
| Shared visual context * Trial number | 0.003   | 0.007 | .597             |

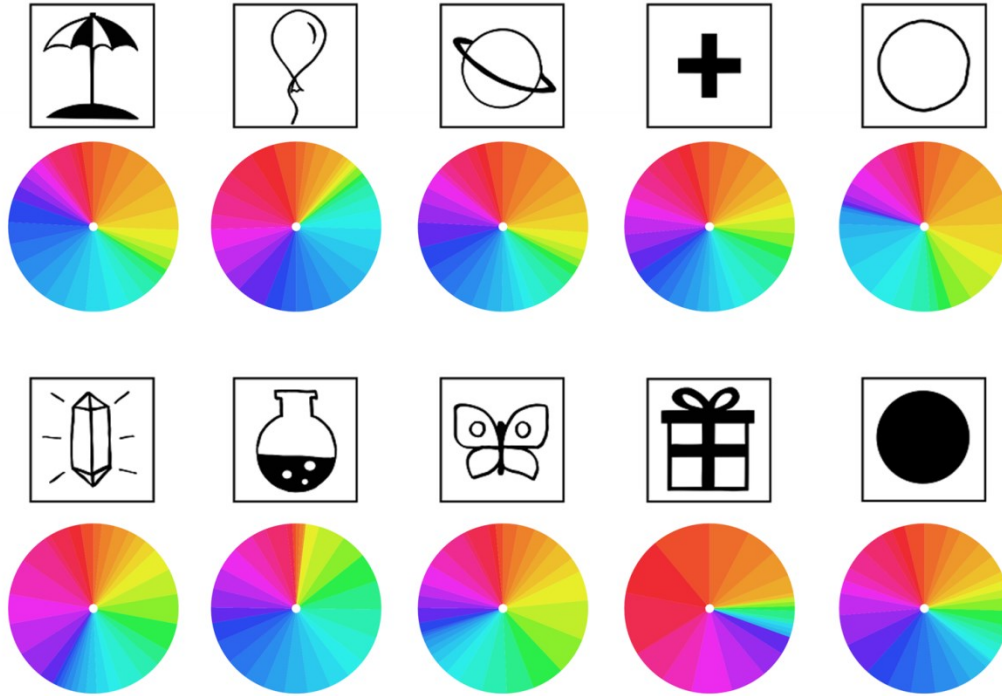
<sup>a</sup>p-values < .05 are marked in bold.

### **Does the shared visual context increase the number of conventions?**

This time, the more systematic questionnaires allow us to separate the reported colors one by one instead of relying on categories chosen by the senders. This highlights the diversity of conventions in different pairs even more (cf. Fig. 2.11). Again, symbols were mapped to specific color hues, but also to other features such as perceived brightness levels.

<sup>2</sup> Adding an effect for the two experimenters that had conducted the study did not reveal any differences between them.






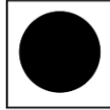

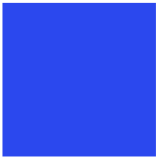


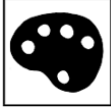


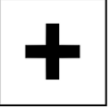
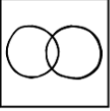
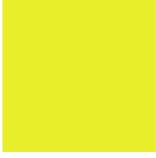

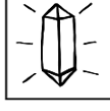
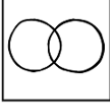





**Fig. 2.11. Sender reports about the usage of symbols in experiment 2 with regard to all 32 colors, for the symbols presented in Figure 5.6 (+ two new ones).** The size of the area occupied by every color corresponds to the number of reported uses in the questionnaires. Distributions similar to experiment 1 can be observed, like the “planet” symbol (top middle) being used mainly for blue and orange colors, or the “flask” symbol (second from the left in the bottom) being used mainly for blue and green. The two circles (at the right in both rows) show a clear distinction for brightness: The top one is mainly used for subjectively brighter colors, and the bottom one is mainly used for subjectively darker colors.

We computed the normalized Levenshtein distance within dyads to measure their alignment after the experiment. This was done using the R package *stringdist* (Van der Loo, 2014) for each of the 32 colors assessed in the post-experiment questionnaire, with missing values for each color in a given pair if either of the participants had not chosen any symbol for the color. This conservative approach produced a large amount of missing values (274 cases or 17.8% of the sample). The strings compared were composed of a single (unique) letter for each symbol used for the respective color. For example distances on highly aligned and lower aligned strings, see Fig. 2.12. This string distance did not take the order of symbols into account, as that information was not available from our study design: During the task, symbols could be arranged freely on the whiteboard space, and

thus order information was not obtained in the questionnaires. We constructed a linear mixed effects model in which the distance was predicted by the shared visual context, with random intercepts for pairs and colors. There was a positive estimate for the effect of shared visual context ( $\beta = 0.08$ ,  $SE = 0.05$ ), but it was nonsignificant ( $p = .118$ ). This means we could find no support for our prediction (4).

| Target color  | Sender description  | Receiver description  | Normalized Levenshtein distance |
|---|---|---|---------------------------------|
|    |    |     | 0                               |
|    |    |     | 0.5                             |
|  |     |    | 1                               |

**Fig. 2.12. Example Levenshtein distances capturing sender-receiver alignment from the self-reported symbols describing specific colors in the post-experiment questionnaire.** Example strings of senders and their respective receivers are in the middle columns. If a pair is perfectly aligned (top example), the normalized Levenshtein distance is 0, if strings differ completely (bottom example), it is 1.

### **Are conventions developed by shared visual context pairs more generalizable?**

To measure the generalization of conventions to new contexts, we look at functional symbol re-use in the second half of the experiment. Since our hypothesis was based upon the assumption that re-use was functional, we had to test this first. We computed the number of symbol types re-used (in the same pair) for every single trial of the experiment in half 2, i.e. whether a symbol type that

had been used at any time in half 1 appeared in the relevant trial. If symbol re-use were functional, this variable should be a significant predictor of accuracy in those trials. Before implementing these variables in a model, symbol re-use values were normalized by the total amount of types (re-use and novel use) appearing in the trial to account for differences in length of messages. The average proportion of symbols in the messages in half 2 that had been used in half 1 already was  $M = 0.91$  ( $SD = 0.18$ ).

**Table 2.3. Estimates and p-values for the fixed effects in the model for functional re-use.**

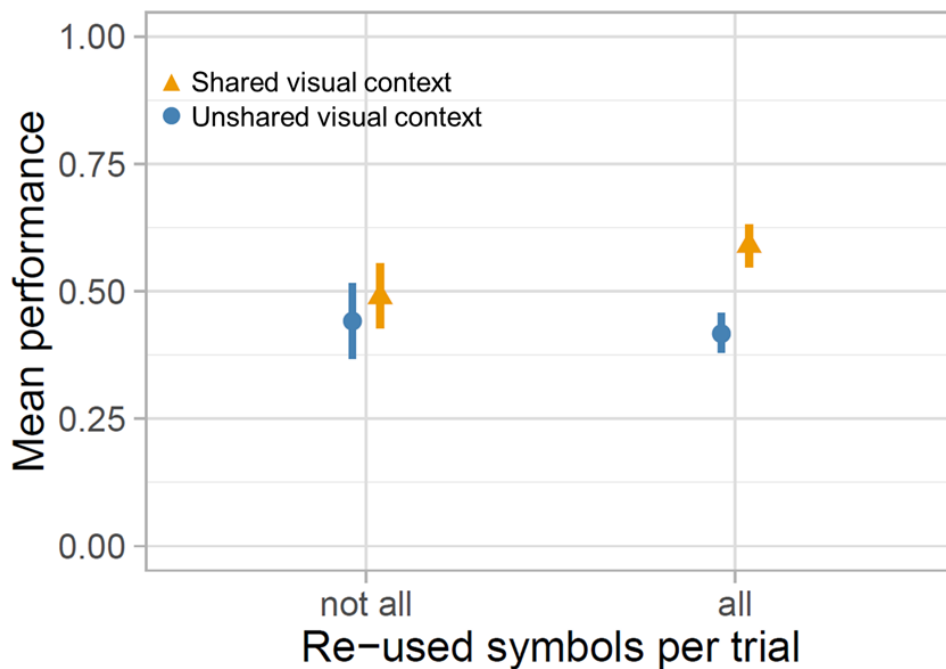
| Fixed effect                   | $\beta$ | $SE$ | $p^a$       |
|--------------------------------|---------|------|-------------|
| Intercept                      | -0.02   | 0.13 | .905        |
| Shared visual context          | 0.69    | 0.24 | <b>.005</b> |
| Re-use                         | 0.07    | 0.58 | .905        |
| Shared visual context * Re-use | 2.31    | 1.15 | <b>.045</b> |

<sup>a</sup>p-values < .05 are marked in bold.

We used a logistic mixed effects model in which accuracy was predicted by re-use and shared visual context, with random intercepts for pairs and target colors and a random slope for re-use (only on the intercept for pairs; for colors, a random slope was not as feasible design-wise, since color spaces varied between pairs), to test for functional re-use. There was a significantly positive estimate for shared visual context (see Table 2.3), replicating the accuracy result from above, and a positive but nonsignificant estimate for re-use. Additionally, there was a significant positive interaction between shared visual context and the amount of re-use, indicating that pairs in the shared condition were performing even better when they were re-using more symbols, whereas pairs in the unshared condition were performing worse when they were re-using more symbols

(for a visualization, see Fig. 2.13). Thus, re-use was functional for shared condition pairs, but not so for unshared condition pairs.

We then computed a new variable indicating whether any symbol used in half 1 was re-used by the same pair in the second half, coded in a binary fashion ( $M = 0.64$ ,  $SD = 0.48$  for shared condition;  $M = 0.73$ ,  $SD = 0.44$  for unshared condition). This variable was predicted, in a logistic mixed effects model, by shared visual context, with random intercepts for pairs and symbols. There was a significant negative effect for shared visual context ( $\beta = -0.47$ ,  $SE = 0.20$ ,  $p = .017$ ). This means that unshared condition pairs re-used more symbols in half 2 than shared condition pairs (although it was not functional for them; this is contrary to prediction 5).



**Fig. 2.13. Relationship between symbol re-use and accuracy in the two conditions.** Because of the generally high amounts of re-use in the data, it was dichotomized into trials that consisted entirely of re-used symbols and trials that saw some or no re-use of symbols, for purposes of visualization. Error bars represent 95% confidence intervals. It can be seen that shared context pairs outperform unshared context pairs whether they re-use symbols or not. Furthermore, an interaction effect is visible: Shared context pairs performed even better when they were re-using symbols, whereas unshared context pairs performed even worse when they were re-using symbols.

## Discussion

Experiment 2 replicated the main results of experiment 1, demonstrating for the second time the importance of the shared visual context for the successful emergence of novel communication. This could still be shown whilst controlling for color difficulty and the perceptual distance between colors, and with a reduced set of symbols. The task was harder overall, but dyads managed to improve in performance over time, even when generalizing to a novel reference space. We found no evidence that conventions were more frequent for pairs in the shared condition. Contrary to our expectations, pairs in the unshared condition were re-using more of their symbols in the second half; however, at the same time their re-use did not appear to be as functional as it was for pairs in the shared condition.

### 2.5 General Discussion

In two experiments, this study demonstrates that the shared visual context between two interlocutors is useful for the emergence of communication, enabling more success when developing a novel communication system. This is shown by our main result: the higher accuracy for dyads in the shared visual context condition, compared to the unshared condition. We demonstrated the importance of the shared context for the successful emergence of conventions even in the absence of training on symbol meanings and external feedback. Instead, participants had to rely on inference and interaction, which potentially amplified the contextual effects we wanted to investigate. Our results represent direct empirical evidence for theories emphasizing the importance of context for successful communication (e.g. Clark, 1996; Sperber & Wilson, 1996), and extend these considerations to the study of language emergence.

The *emergence* of conventions implies we saw novel conventions arising within dyads during the experiments. Participants could not achieve a high level of success without endowing vague symbols with novel meanings. Whilst we acknowledge that providing participants with pre-established symbols means they already bring formed associations into the experiment, we argue that this cannot be the main factor behind the conventionalization. At most, it biases participants to prefer certain colors while the desired ambiguity in the selection of our symbol space remains. In other words, symbols may carry a little information from the start, but over the course of the

task, they acquire much more. It is this increase in informational value that we were interested in, not possible prior associations for symbols.<sup>3</sup> The evidence we provide for this in the two experiments is twofold: First, conventions between pairs differed drastically, as seen in the results from the questionnaires in both experiments. Second, performance continually increased over time in both conditions and in both experiments, implying that participants built up a shared conversational history – and thus, conventions.

However, our alignment questionnaire in experiment 2 failed to show a difference between the two conditions for the number of conventions arising, even though their success differed. It could be the case that the shared context does not facilitate the creation of conventions but merely boosts the successful emergence of communication. We would argue that the lack of evidence for the predicted effect is likely due to a methodological problem, however, especially since we could observe the desired pattern in the explorative results of experiment 1. The rigorous questionnaire we employed in experiment 2, prompting participants to tick a description for every single one of the 32 colors in the meaning space, unfortunately produced a lot of missing values. The participants felt overwhelmed by the precision required for this task, and in many cases reported being unable to remember or make an educated guess about how the symbols were used for a particular color. This was particularly true of receivers, who had to infer which meanings their sender had associated with which symbol. Because of our conservative approach in the analysis, even one missing message from either of the participants led to an exclusion of the alignment for the whole color for the pair (i.e. a given row in the data). Future studies would be well placed to investigate the importance of contextual knowledge for the formation of conventions much in the same way as we did in experiment 1, but employ more controlled (i.e. fixed) categories instead of free recall, while not presenting participants with the entire meaning space.

In neither of the experiments did we find evidence that dyads in the shared condition progressed faster in performance than unshared visual context pairs. Following this result, we have to assume that in our specific task, shared visual context pairs were not able to capitalize more successfully on their conventions and their performance, which was already higher from the start. This would suggest that the effect of the shared visual context is fixed, elevating pairs' performance rather

---

<sup>3</sup> For a supplementary analysis regarding the effect that biased associations might have on successful or unsuccessful communication, see the appendix.

than multiplying it with more experience. On the flip side, it means that pairs in the unshared condition were able to improve equally well in the task, just overall below the performance of the shared condition. It would be interesting to change our experimental design to include a within-pair manipulation of the context to investigate whether dyads would immediately profit or suffer from switching to shared or unshared contexts.

Experiment 2 tested the prediction that senders in the shared condition would be more likely to re-use their symbols, with the assumption that such re-use should be functional for all dyads. We made this prediction because we believed shared information would foster the evolution of more generalizable conventions. This prediction could not be tested, because re-use turned out to be functional only for dyads in the shared condition; in the unshared condition, re-use did not help performance. Though less functional, re-use was more frequent in the unshared condition — surprisingly, in light of our initial prediction.

We do not know what made senders in the unshared condition re-use symbols more than pairs in the shared condition. Potentially, the access to the shared visual context might have tempted pairs to create conventions that rely on it to carry a good part of their intended meaning; in other words, shared visual context pairs might have made use of their opportunity for contextual enrichment (cf. Winters et al., 2018), leading to a greater need for novel symbols once the contexts changed. Alternatively, we suspect that the shared visual context could have made the transition between half 1 and half 2 more salient, encouraging senders to change their repertoire of symbols.<sup>4</sup> In one respect, however, experiment 2 verified our expectation that the conventions evolved in the shared condition would be more generalizable: Symbol re-use resulted in better performance in the shared condition, and in that condition only. This is consistent with the general view that linguistic conventions emerge by being used in ostensive-inferential communication (Höfler, 2009), and with the specific claim that the shared visual context makes for more efficient communication, yielding more generalizable conventions.

At the center of our study was the manipulation of the shared visual context. As described in the introduction, this is merely one aspect of the general notion of context, and ignores other types

---

<sup>4</sup> An additional suggestion brought forward during review is that re-use in shared context pairs might have focused on symbol *combinations* rather than single symbols. We address this idea with a supplementary analysis in the appendix.

such as the historical context (e.g. Yoon, Benjamin, & Brown-Schmidt, 2016) or the basic community membership (e.g. Clark, Schreuder, & Buttrick, 1983) of participants, interesting objects of study in themselves. Interestingly, there is a case to be made for our manipulation also concerning the historical context: Dyads in the unshared condition were limited to a history of unshared contexts in addition to their immediate situation, and also switched to a new set of unshared contexts in experiment 2. As such, we cannot separate the effects of the immediate shared context and the shared context accumulating over time. However, this is less problematic since our main interest lay in the evaluation of the shared effect, which entails both of these confounded aspects.

Our manipulation to the shared context was achieved by removing all distractor colors from the arrays of senders in the task, so that they only knew what the target in the current trial was. It is important to note that different operationalizations of the shared context would have been possible: For instance, another option would have been to present entirely different contexts to sender and receiver (with the same target color), but keep the amount of colors the same. We decided not to do this because i) it is difficult to keep the differences between and within conditions constant with this design and ii) participants would probably have to be deceived about them not seeing the same colors in this case. In contrast, we settled for an open and informed quantitative manipulation of the shared context, such that only the amount of colors varied. As such, we expected and found quantitative differences in the performance of dyads as well; nevertheless, the question whether this result generalizes to other operationalizations of the shared context would need to be addressed empirically by future studies.

Another open question concerns the cognitive representations underlying the more successful communication in the shared condition. Do interlocutors take their partner's knowledge into account to communicate accurately? Some theories suggest that they should (Clark, 1996; Lewis, 1969). However, as outlined in the introduction, this point has been challenged by a line of research studying reference resolution with the eye-tracking method (e.g. Horton & Keysar, 1996; Keysar et al., 2000). As such, the results of our experiments are also in line with a more parsimonious explanation (Keysar, 1997): Senders could simply be better at the task because there is more knowledge available to them. It is important to note that this is still in agreement with Sperber and Wilson's relevance theory (Sperber & Wilson, 2002). Here, shared representations are not always



necessary for communication, but the individual representation of contextual information for both interlocutors is often sufficient. In our case, the senders' messages could simply be built on their contextual information, and likewise, the receivers' inferences could be drawn from the message combined with what they see on their screen. Minimally, then, we have shown the benefits of the shared context for the successful emergence of communication, but are not making any claims as to how this context is used by the interlocutors exactly.

We grounded our experimental design and the general research question about contextual influences on the emergence of language in an ostensive-inferential model of communication. This led to a number of design choices, most notably the absence of training for any symbol meanings and the reliance on repair mechanisms combined with a lack of external feedback. By doing this, we aimed to come closer towards the emergence problem of communication. Participants were encouraged to create novel conventions through interaction. Although we acknowledge that there might still be biases from interference with their natural language (a general problem for artificial language experiments), we think it is necessary to eliminate as many alternative mechanisms for the formation of novel conventions as possible in an experimental setting. All in all, we think the current study provides a firm basis for how future studies can utilize the ostensive-inferential framework to investigate the emergence of language.

## **2.6 Conclusion**

In this paper, we set out to investigate the influence of the shared visual context on the successful emergence of communication. To this end, we combined pragmatic concepts with the methods of experimental semiotics. We constructed two artificial language experiments and found that participants performed better in a referential task when they had access to the visual context. This has implications for the emergence of language, and is in accordance with an ostensive-inferential model of communication: To successfully create and interpret a novel convention, interlocutors build on the contextual information. In the second experiment, we also found that participants sharing the visual context adapted their conventions more successfully to new contexts than those lacking the context. At the same time, unshared visual context pairs re-used more of their conventions, the reasons for which remain unclear. On the methodological side, our experiments

demonstrate how an ostensive-inferential framework can be used to inform choices in the designs used by artificial language experiments, emphasizing inferential processes and interaction.

### **Data and Code Availability**

All data and R code is available on the Open Science Framework: <https://osf.io/ts4ka/files/>.

### **Acknowledgments**

We would like to thank Helene Kreysa and Dana Schneider for their role in supervising Thomas Müller's master thesis, partly reflected in experiment 1. A special thanks goes to Lisa Jeschke for her help with organizing and conducting experiment 2.

### **2.7 Appendix**

Here, we report two supplementary and exploratory analyses suggested during review. First, we tried to address the question of whether a reason for shared condition pairs re-using fewer symbols in the second half of experiment 2 might be that they develop conventions for *combinations* of symbols rather than the single symbols themselves. We investigated this by treating symbol use in the second experiment on the level of entire messages, i.e. unique symbol combinations instead of single symbols, ignoring reduplications of the same symbol. Descriptively, the diversity of unique combinations used in the second half of the experiment does not differ between the two conditions, as suggested by both the relative proportion of combinations that are duplicates of previous messages (45% vs. 47%) and the conditional entropy of combinations given participant pairs (3.68 bits vs. 3.64 bits). We also repeated our analysis on symbol re-use on the level of unique combinations (as opposed to individual symbols), predicting the re-use of every combination used in half 1 of the experiment (as a binary variable) by condition while adding a random intercept for participant pairs. This model revealed no effect of condition ( $\beta = -0.39$ ,  $SE = 0.34$ ,  $p = .25$ ), leading us to tentatively conclude that pairs in both conditions re-used unique combinations of symbols at a similar rate, based on this post-hoc analysis.

Second, we ran an exploratory analysis based on experiment 2 to find out whether players are more successful when using symbols that usually show strong associations with a given colour range (cf. Fig. 2.11). Biased associations were assessed by computing the conditional entropy on the frequencies of colors given symbols: Here, for each symbol, higher values mean more diversity in symbol associations (i.e. a less biased distribution of colors). Interestingly, most symbols end up with very high values of entropy, calculated this way ( $>.9$  on the normalized variable, which takes values between 0 and 1). This can be seen as further evidence that symbol associations were not straightforward for participants and not limited in reference to a selective part of the color space. What we find in the model is that we can replicate the known effects for condition and trial number, but do not see a significant effect for the mean entropy per trial, even though the parameter points into the expected direction (i.e. higher accuracy for trials that inhibit symbols with more biased associations).

### **3. Compression in Cultural Evolution: Homogeneity and Structure in the Emergence and Evolution of a Large-Scale Online Collaborative Art Project**

This chapter represents the following study published in *PLoS one*:

Müller, T. F., & Winters, J. (2018). Compression in cultural evolution: Homogeneity and structure in the emergence and evolution of a large-scale online collaborative art project. *PloS one*, *13*(9), e0202019. <https://doi.org/10.1371/journal.pone.0202019>

Author contributions: Both authors contributed to all parts of the study. The conceptualization of the study and writing as well as editing of the manuscript were equally shared between the authors. The data curation, analyses and figures were mostly conducted by me, with support from James Winters.

#### **3.1 Abstract**

Cultural evolutionary theory provides a framework for explaining change in population-level distributions. A consistent finding in the literature is that multiple transmission episodes shape a distribution of cultural traits to become more compressible, i.e., a set of derived traits are more compressed than their ancestral forms. Importantly, this amplification of compressible patterns can become manifest in two ways, either via the homogenization of variation or through the organization of variation into structured and specialized patterns. Using a novel, large-scale dataset from Reddit Place, an online collaborative art project, we investigate the emergence and evolution of compressible patterns on a 1000x1000 pixel canvas. Here, all Reddit users could select a colored pixel, place it on the canvas, and then wait for a fixed period before placing another pixel. By analyzing all 16.5 million pixel placements by over 1 million individuals, we found that compression follows a quadratic trajectory through time. From a non-structured state, where individual artworks exist relatively independently from one another, Place gradually transitions to a structured state where pixel placements form specialized, interdependent patterns.

## 3.2 Introduction

Explaining population-level distributions of cultural traits across both spatial and temporal dimensions is a central goal of cultural evolutionary theory (Boyd & Richerson, 1988; El Moudden et al., 2014; Henrich, 2004; Henrich et al., 2008; Kandler et al., 2017; Kirby, 1999; Laland & Brown, 2011; Mesoudi, 2017; Powell et al., 2009; Smith & Kirby, 2008). Gaining traction within this framework is the idea that cultural traditions organize themselves into compressible patterns in response to simplifying pressures (Bartlett, 1932; Brighton, Kirby, et al., 2005; Chater & Vitányi, 2003; Culbertson & Kirby, 2016; Henrich, 2004; Kemp & Regier, 2012; Kirby, 2017; Kirby et al., 2008, Kirby et al., 2015; Tamariz & Kirby, 2015; Zipf, 1949). Compressible patterns are found in many cultural domains, from language and music to technology and art (Kirby et al., 2008; Ravignani et al., 2017; Tamariz et al., 2016). What makes a set of cultural traits compressible is if the length of the set is describable by a rule shorter than simply listing each individual trait. Islamic geometric art, for instance, is highly compressible in that many artworks within this tradition are built on simple, generalizable rules, involving the repeated use of basic shapes to generate novel, open-ended patterns (Dabbour, 2012). Increasing compression is therefore characterized by the organization of cultural information into more predictable patterns; the less irreducible unpredictability remaining in the data, the greater the amount of compression (Culbertson & Kirby, 2016).

Compressible patterns can become manifest in a distribution of cultural traits via two processes. The first is through spreading homogeneity in a population of cultural traits (Reali & Griffiths, 2009). Consensus formation is an example where, through the differential amplification of behaviors, variation is regularized via the spreading of homogeneity (Baronchelli, 2018). The second is to organize variation into structured, interdependent patterns (Claidiere et al., 2014; Kirby, 2017). Heraldry is structured because it can generate novel designs by using underlying rules to recombine a finite set of components (e.g., motifs combine with tinctures of metals, colours, and furs to create a design; Morin & Miton, 2018). The main difference between the two is that the first type of compression removes variation and encourages homogeneity, whereas the second type of compression maintains variation and imposes structure.

Using a novel, large-scale dataset from *Reddit Place*, a collaborative pixel art project involving over one million participants, we investigate the emergence and evolution of compressible patterns

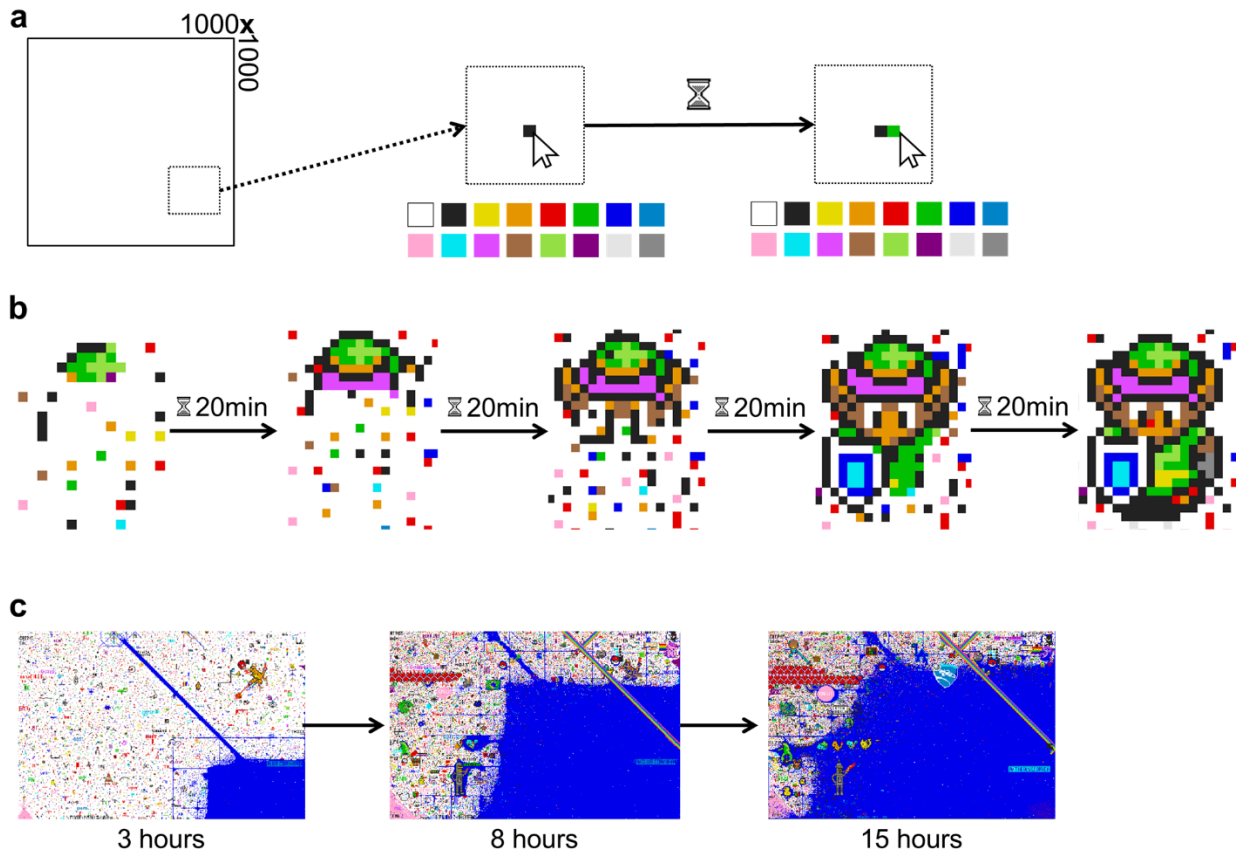
on an unprecedented scale. Our general hypothesis is that compressible patterns are present in Place and driven by artworks being organized into structured, interdependent arrangements. By measuring compression, and controlling for changes in the frequency distribution of variation, we discriminate between homogenizing and structuring processes. If changes in variation are driving compression, then decreases in variation should result in increases in compression (i.e., homogeneity spreads). Conversely, if variation is being organized into structured and specialized patterns, then increases in compression are decoupled to some extent from changes in the frequency distribution of variation (i.e., structure spreads).

### **From Pixels to Artworks: Reddit Place**

For 3 consecutive days (72 hours), starting from March 31st 2017, Reddit opened a white 1000x1000 pixel canvas to its users. When visiting the page, all users that had been registered prior to the start of this event were allowed to place pixels on this canvas (see Fig. 3.1a for details). This was accompanied only by the following developer post:

*There is an empty canvas. You may place a tile upon it, but you must wait to place another. Individually you can create something. Together you can create something more.*

Crucially, after Reddit users had placed one pixel anywhere on the canvas, a fixed period (initially 5 minutes) was imposed before a user could place another pixel. Thus the *collaborative* aspect of the project: On their own, users were severely limited with regard to their possible creations. Because of this, it is of no surprise that cooperation was present from early on, even though many initial pixel placements were disorganized, chaotic and lacking in overall direction. Through these collaborations, taking place via posts on Reddit, simple artworks began to form (Fig. 3.1b), corresponding to one way of organizing variation into structured patterns. Creations covered a wide range of themes, from national flags and sports teams to video games and general *geek culture*, with some instances being geared toward goals internal to Place. For example, members of *Blue Corner* tried to dominate as much space as possible by repeatedly placing blue pixels in non-blue regions (Fig. 3.1c).



**Fig. 3.1. The basic mechanisms of Place.** **a)** Reddit users could select a single pixel from a set of 16 colors, place it anywhere on the 1000x1000 canvas, and then wait for a fixed period until they could place another pixel. This represents the introduction of variation to a homogeneous space on an individual level. **b)** Through collaboration, simple artworks can form, with variation being organized into structured, compressible patterns. **c)** Some groups pursued goals internal to Place, like dominating as much of the canvas as possible, which highlights how compressible patterns can arise via the removal of variation.

### Cultural Evolution in Place

Culture exists as information in the minds of individuals (e.g., an idea for an artwork or artistic techniques) and is expressed in a population as cultural traits via observable behaviors (e.g., the act of placing paint on a canvas) as well as tangible artefacts (e.g., artworks; Ferdinand, 2015; Hurford, 2002). Place meets all the necessary prerequisites of a cultural evolutionary process: there is reproduction, variation, and change in a population of cultural traits (Richerson & Boyd, 2008).

Artworks form lineages of descent, colored pixels constitute cultural traits, and both deterministic and stochastic factors can govern the rise and fall in the frequency of these traits.

The core mechanism of Place is relatively simple, as it involves choosing and placing a colored pixel on a canvas. This process fulfils the minimal requirements of cultural transmission through the repeated observation and production of behaviors (Kirby et al., 2014): Individual users are able to observe the behavior of others via previous pixel placements, and to then use these observations to update their beliefs and make a decision about which colored pixel to choose and where to place it on the canvas. What makes Place unique in this respect is that the transmission dynamics and population structure are far more complicated than standard cultural evolution experiments (Kirby et al., 2008) and simulations (Griffiths & Kalish, 2007): Individuals are free to organize themselves into groups, and the sheer number of individuals maps more closely to the scale of transmission events and group structures we observe in modern, large-scale societies.

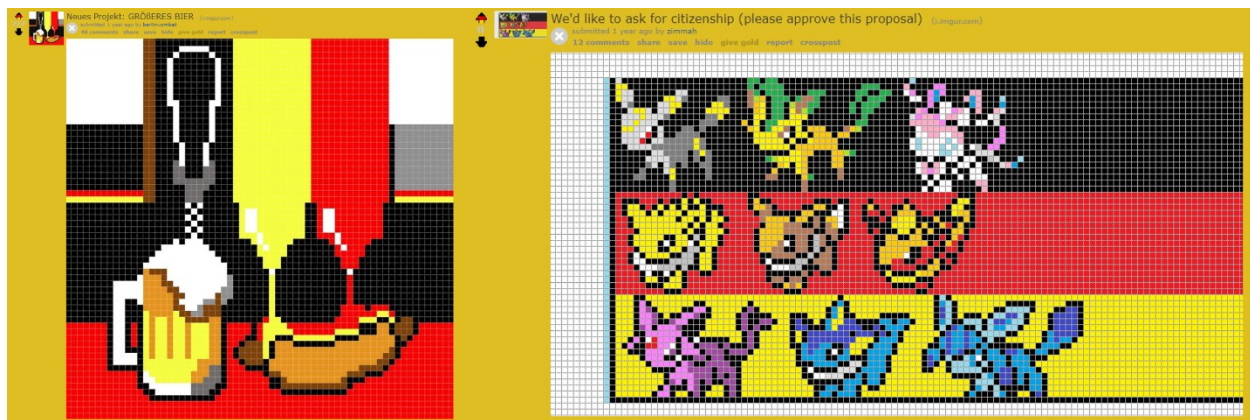
The nature and scale of the task highlights several additional differences between Place and standard cultural transmission experiments (Mesoudi & Whiten, 2008). First, there is perfect retention of pixel placements between time-steps, removing the need to remember or reproduce pixel placements at a previous time-step. Second, due to the complexity of the transmission dynamics, discerning generations is not straightforward at the level of users. A single individual can appear at many different time steps, meaning there are several possible transmission patterns (e.g., one-to-many, many-to-many, oblique and horizontal transmissions; see Mesoudi & Whiten, 2008 for overview). Lastly, users were not provided with an overt goal in choosing and placing colored pixels. Instead, the motivation for creating and participating in artworks was determined by the preferences of individuals. Compare this to most experimental tasks where participants are generally provided with a training regime and given explicit instructions about the goal (e.g., to successfully communicate: Winters et al., 2018; or asked to memorise a set of sequences: Kirby et al., 2008).

Overlaid on top of the task itself are the subreddit communities. Users could come together to strategically plan pixel placements and organize the resources of their respective subreddits in maintaining, expanding, and sabotaging artworks. As such, Place captures individual variability of users as well as social factors emerging from within-group goals (e.g., to mark group identity). An example of this is found in the *PlaceDE* subreddit (see Fig. 3.2): users often proposed new



pixel art projects to the community, which were then subject to discussion and voting. This process played a within-group role for determining which artworks eventually made it onto the canvas. So, even before users placed a pixel, there were cultural evolutionary dynamics governing the selection of ideas for artworks.

In many respects, Place forms a microcosm of cultural evolution in a similar manner to Petri dish bacteria in biological evolution (Blount et al., 2008). It is also worth noting that the dissimilarities between Place and other cultural phenomena are not as great as their similarities in terms of evolutionary dynamics. Just as *e.coli* differ from humans in that they can reproduce asexually and transmit genetic information horizontally as well as vertically (Barton et al., 2007), so too does Place differ from other, well-studied examples of cultural evolution (e.g., language, technology etc.).



**Fig. 3.2.** Two proposals for artworks in PlaceDE subreddit. **Left:** Proposal for artwork which successfully made it onto the canvas with 225 votes. **Right:** An unsuccessful proposal with 0 votes. Subreddits allows for selection of artworks based on up-voting or down-voting as well as via comments below the proposals.

### Research Questions, Hypotheses and Predictions

Any evolutionary system where resources are finite requires solving the problem of competition and coexistence (Barton et al., 2007; Darlington, 1972). Like in the biological domain, cultural traits (i.e., colored pixels) that enter into competition with one another cannot coexist indefinitely (see *competitive exclusion principle*; Hardin, 1960; Roberts & Fedzechkina, 2018): Either one trait

drives the other to extinction or competing traits come to occupy distinct niches (see *niche differentiation*; Altmann et al., 2011; Scheffer & van Nes, 2006). Competition resulting in extinction is essentially a homogenizing process (as highlighted by cases like the Blue Corner; Fig. 3.1c) and niche differentiation maintains and organizes variation by structuring it into specialized patterns (e.g., Fig. 3.1b). What remains unknown is the extent to which these processes shape the overall distributional patterns of the canvas (i.e., the complete set of pixel placements). Specifically, we ask: (a) Does compression follow a predictable time course? (b) Is this overall compression driven by variation being reduced or structured? (c) Does the canvas become increasingly stable?

Place provides a particularly apt and novel dataset for answering questions pertaining to cultural evolution. A central feature of Place is its fixed, bounded space of 1000x1000 pixels, placing a hard constraint on the maximum amount of variation in a population (i.e., exactly a million pixels). Having a finite space allows us to investigate whether the density of the canvas interacts with evolutionary dynamics in shaping the distribution of pixel placements. This leads us to two predictions about the time course progression of Place:

- When Place is sparsely populated with artworks, the canvas will decrease in compression, have low stability, and increase in variation.
- When Place is densely populated with artworks, the canvas will increase in compression, have higher stability, and plateau in variation.

Initially, users solve the competition-coexistence problem by creating different artworks in unused regions of the canvas, offsetting competition between artworks of different groups. Such dynamics are even formally codified in some cases, exemplified by rule 1 of the *Green Lattice* subreddit:

*Protect art, do not destroy it. Work around existing pixel art. Any new art must be OUTSIDE our borders unless we approve it first.*

The overall effect of this strategy results in the canvas becoming less compressed over time as distinct groups create artworks in unused regions. Growth in diversity should also be reflected in a more uniform distribution of colored pixels and lower levels of stability.

The viability of such a solution depends on there being a large number of unused regions to mitigate between artwork competition and for innovations to promulgate. This is not the case at

the latter stages of Place. Now, the canvas is densely saturated with artworks, making the number of unused regions a rarity and increasing the probability of competition. Compressible patterns are advantageous because these are easier to produce, maintain, and generalize. If two artworks compete for pixel placements, then the more compressible of the two is better adapted to this niche as it can expand quicker and is more readily reconstructed than its less compressible counterpart.

Group structure acts as a countervailing force to an increased pressure for compressible artworks (perhaps in an analogous fashion to the role of communication in language evolution; Kemp & Regier, 2012; Kirby, 2017). First, each group has its own internal goals for creating artworks, and it becomes increasingly costly for one group to dominate the canvas. Even with a maximally homogeneous artwork, maintaining and expanding an artwork at the expense of other artworks requires a significant number of users (relative to the total population). Second, by modifying pre-existing artworks to create derivative innovations, groups can mitigate competition by preserving elements of old artworks whilst facilitating the creation of new ones. This second point explains why we predict structured, stable, and diverse patterns to emerge: the specialization and integration of artworks injects structure into the canvas via the creation of shared features.

### **3.3 Materials and Methods**

#### **Procedure**

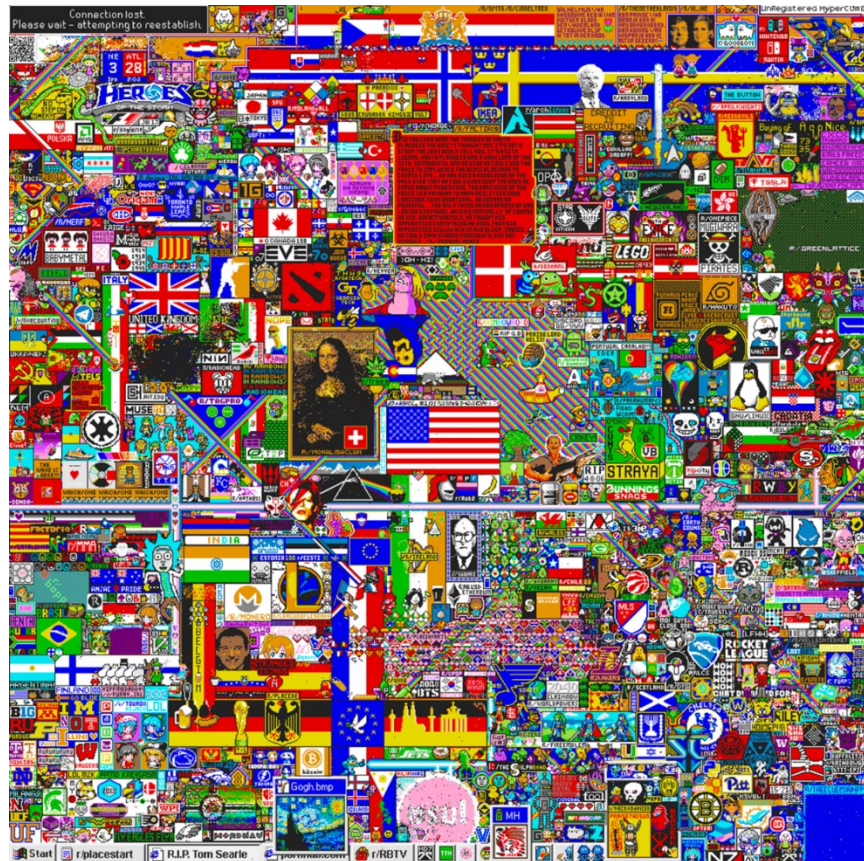
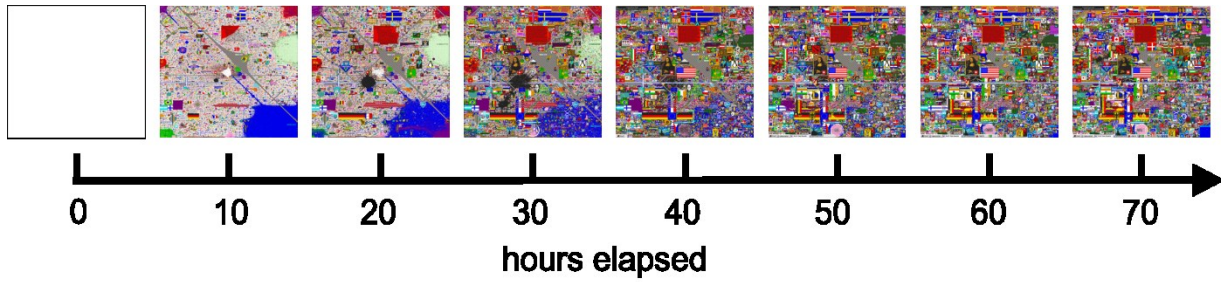
When placing a pixel on the canvas, Reddit users could navigate and zoom into areas of interest to allow for greater precision (especially important for mobile devices) and overview (for the technical details on how Place was programmed, see <https://redditblog.com/2017/04/13/how-we-built-rplace/>). As mentioned in the introduction, one crucial feature of the project was that a waiting period was enforced after a pixel had been placed. This period had initially been set to 5 minutes, but was changed by the administrators several times while the canvas was live to accommodate the incoming traffic on Reddit's web page. The creators of Place estimate that approximately 80,000 users were connected to the canvas simultaneously at peak time. In the total runtime of 72 hours, about 16.5 million pixels were placed by the Reddit community, corresponding to approximately 1.2 million active unique users. Some caution is warranted, however, with this number of users, since Place was deliberately designed to also allow the

programming of bots. Still, users had to adhere to the waiting period imposed by the rules of the project and stick to their account registered prior to the start of Place, irrespective of whether they placed the pixels manually or automatically. The only exception to this was a minority of users who discovered a loophole that allowed for multiple pixel placements at once if they were sent to the client at the exact same time; however, this exploit was only used to change around 15,000 pixels, i.e. roughly 0.09% of the total placements (<https://redditblog.com/2017/04/13/how-we-built-rplace/>). All in all, we do not see any of the issues described as problematic for our data, since they mean that some individuals were able to have more impact than others (through technological means), while the general notion of Place as a model for cultural evolution still holds.

## **Data**

The raw data of Place has been made freely available online by the Reddit administration: [https://www.reddit.com/r/redditdata/comments/6640ru/place\\_datasets\\_april\\_fools\\_2017/](https://www.reddit.com/r/redditdata/comments/6640ru/place_datasets_april_fools_2017/).

However, as of 2nd August 2018, the link to this file is not working properly (see Data Availability below for how it can be accessed on the OSF). In it, every pixel placement is provided with the corresponding color chosen, user id, and a time stamp. The time stamps exhibit a resolution of one second. Since the full data also includes a period before the project went live, during which the creators of Place conducted some final tests, we reduced this raw file to a runtime of exactly 72 hours by cutting off this testing period. From the resulting dataset, we prepared our analysis by constructing a separate bitmap image (.bmp) for each unique time step showing the current state of the canvas. This resulted in 259,194 images to work from. Fig. 3.3 gives examples of the canvas for every 10 hours in the time course of the project, as well as showcasing the final outcome of Place. When the 72 hours had elapsed, the canvas was frozen in this state, and Reddit users were aware of this fact for the duration of the project.



**Fig. 3.3. The development of the Place canvas. Top:** State of the canvas in 10-hour-intervals. **Bottom:** The final image after 72 hours, frozen in its state.

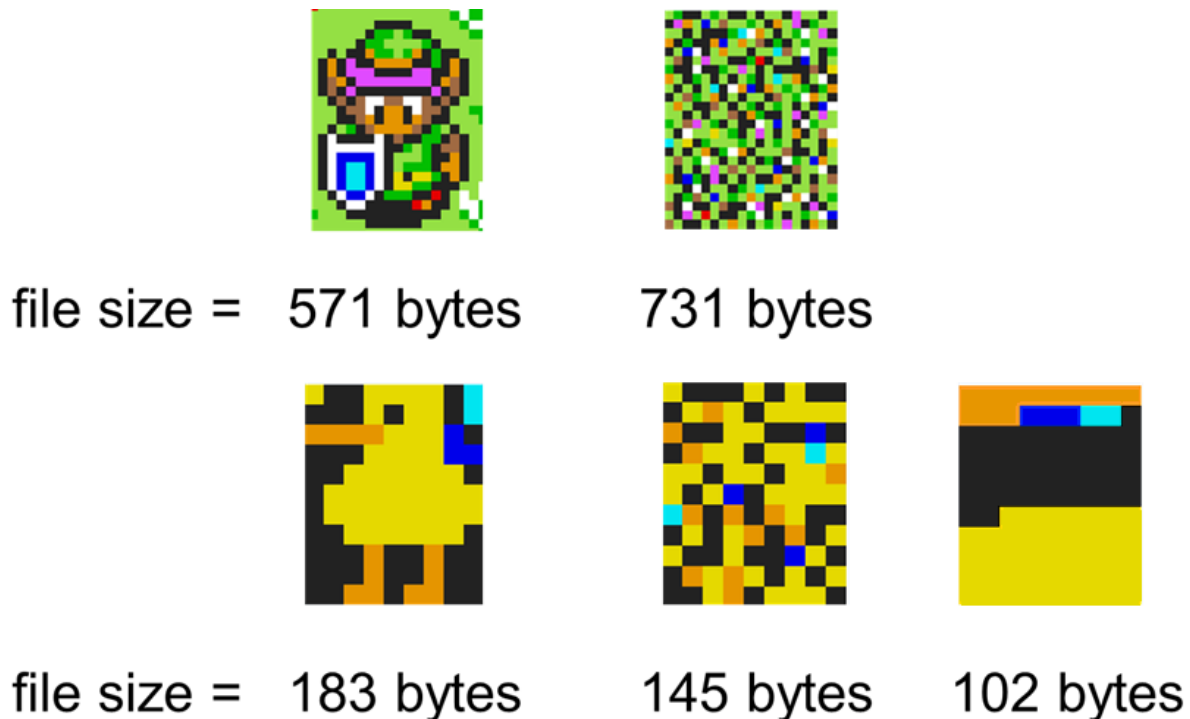
### Code and Data Availability

All code and data, including the raw data, is available at the following repository:  
[https://osf.io/qjkc/?view\\_only=ccc612e3b65d455fb5f1e6d843fb19df](https://osf.io/qjkc/?view_only=ccc612e3b65d455fb5f1e6d843fb19df).

## Measures

### Compression

Compression was measured by converting all of our .bmp images to png images (.png). We used the DEFLATE compression algorithm found in the PIL package in Python to generate a set of maximally compressed .png images (for two simple examples, see Fig. 3.4). As the images generated are always 1000x1000 pixels, we can compare the compression sizes of each image with one another based on their time stamp. In line with the original .bmp images, we obtained 259,194 images this way. The measure of compression equates to the file size of the .png images, read out automatically in Python.



**Fig. 3.4. Compression measure applied to an image.** **Top:** How a .png format can be used to assess the compressibility of an image: The pixels on the image to the right are identical to the left image in color value, but shuffled randomly in their position. Still, the file sizes indicate that the left image is more compressed. **Bottom:** Randomization can still create compressible patterns, as demonstrated by the image in the middle, which is more compressible than the original. Note, however, that a strongly structured image like the one on the right will always be more compressible than the average randomized image.

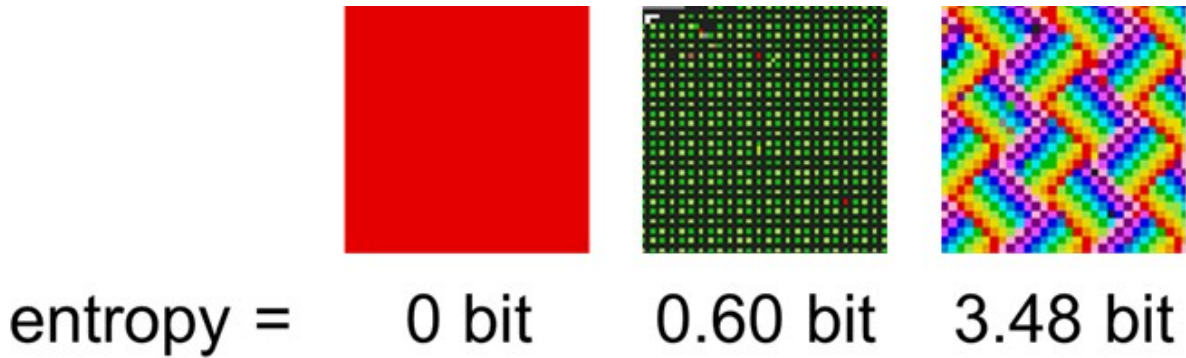
There are two ways compressed patterns can emerge in Place. The first is that changes in compression correspond to changes in the frequency distribution of color tiles (i.e., our measure of global entropy). If this is the case, compression should closely track the time course trajectory of global entropy. The second possibility is that variation is being maintained and structured into regular and predictable patterns. In this scenario, the trajectory of compression should be decoupled to some extent from changes to the frequency distribution of colors, with compression increasing relative to the amount of global entropy.

### Global Entropy

The emergence of compressible patterns is often coupled to changes in the frequency distribution of pixel placements. To measure this, we first calculated a frequency distribution for each of the 16 colors available to users on each of the image files (e.g., if Blue has a frequency of 100, then this corresponds to 100 unique pixels of Blue in Place). Next, we computed the conditional entropy (Cover & Thomas, 1991) of colors given time-steps,  $H(C|T)$ :

$$H(C|T) = - \sum_{c \in C} P(c) \sum_{t \in T} P(t|c) \log_2 P(t|c)$$

where  $C$  refers to the set of 16 colors  $\{c_1, \dots, c_{16}\}$  and  $T$  is the set of time-steps  $\{t_1, \dots, t_{259,194}\}$ .  $H(C|T)$  therefore measures the predictability of a color at a specific time-step (i.e., the non-uniformity of the frequency distribution; see Fig. 3.5). As there is a fixed space of 1000x1000 pixels, a maximally unpredictable state corresponds to a uniform distribution: each color is represented at an equal frequency (62,500 pixels) and  $H(C|T) = 4$  bit. A maximally predictable state is a non-uniform distribution, where  $H(C|T) = 0$  bit and consists of a single color at a frequency of 1,000,000 pixels (with the other 15 colors at a frequency of 0).



**Fig. 3.5. Some real examples from the canvas and their corresponding entropy values.** As can be seen, a completely non-uniform distribution of colors (i.e., a single color) results in an entropy of 0 bit, whereas a strongly uniform distribution of colors (such as a rainbow pattern) results in entropy values close to the maximum of 4 bit.

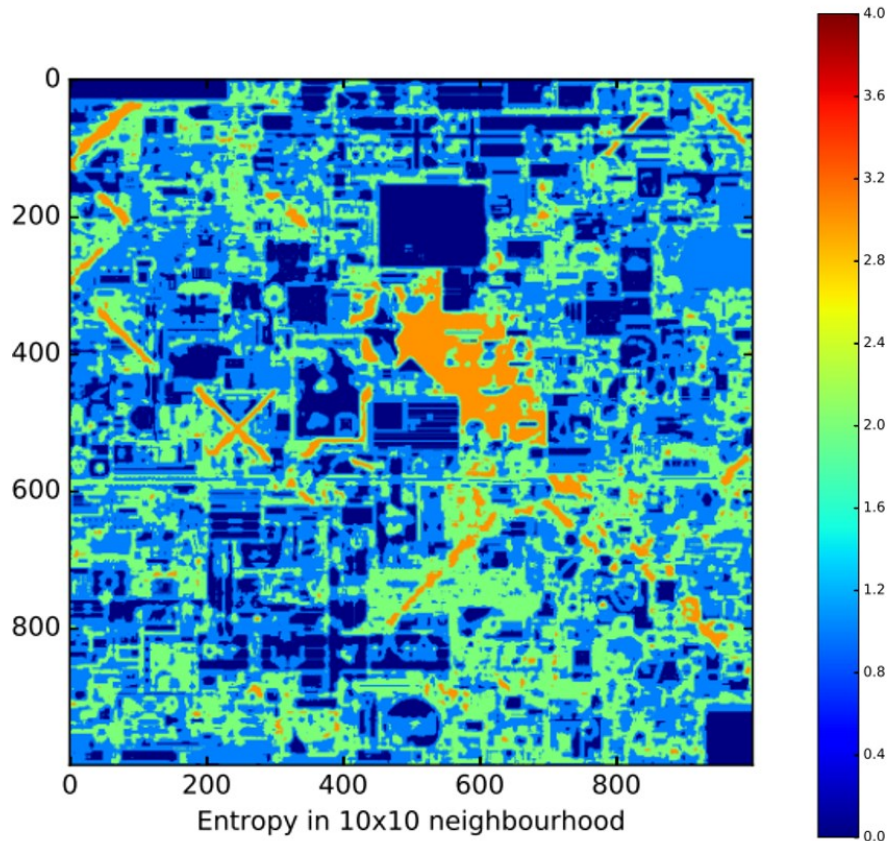
### Local Entropy

The local entropy between pixel placements can be measured by calculating at each pixel position  $(x, y)$  the entropy of pixel placements within a 2-dimensional region (Kadir & Brady, 2001). Using 10x10 regions, we flatten the 2-dimensional region into a 1-dimension array, which is then passed to the conditional entropy function:

$$H(C_R|T) = - \sum_{c_r \in C_R} P(c_r) \sum_{t \in T} P(t|c_r) \log_2 P(t|c_r)$$

where  $C_R$  is the set of 16 colors within a specific 10x10 region. Due to the large amount of computation power required for 1 million pixels, we restricted our use of this measure to 1-hour-steps (72 slices). As with *global entropy*, a maximally predictable pattern has an entropy of 0 and a maximally unpredictable pattern has an entropy of 4 (see Fig. 3.6). Importantly, homogeneous patterns have a low global entropy and a low local entropy, whereas structured patterns should have a high global entropy and a low local entropy. This is because structured patterns maintain variation (hence high global entropy), with regularities being restricted to local level patterns (hence low local entropy).

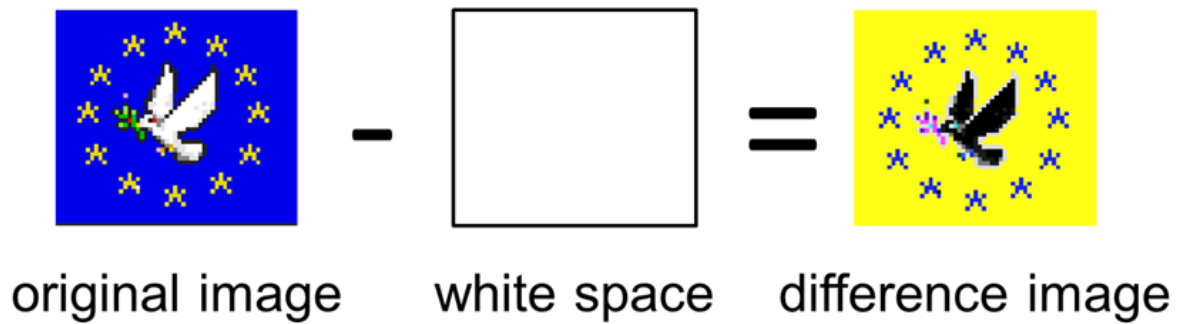




**Fig. 3.6. Local entropy of the final canvas.** Blue regions correspond to 10x10 neighborhoods with low entropy and orange regions correspond to 10x10 neighborhoods with high entropy.

### Stability

Stability was assessed by computing the pixel-by-pixel difference between different slices of the data (see Fig. 3.7). A pixel value is seen as stable if this difference amounts to zero. The total number of stable – i.e., unchanged – pixels indicates the stability of an image. Since the resolution of 1-second-slices is too fine-grained to find meaningful development within a single time step, we ran our analyses on three different resolutions (pre-registered before analysis): a resolution of 1-hour-steps (72 slices), 2-hour-steps (36 slices), and 4-hour-steps (18 slices). Because of the way this measure is computed, coarse resolutions allow us to observe more general trends (since the amount of change between steps is larger), but these resolutions also run the risk of losing power (due to the reduction in the number of data points).



**Fig. 3.7. How the stability measure is applied.** To compare two states of the canvas, the difference between the values of the pixels is computed. On the resulting difference image, stable pixels appear in black, since all their values are zero.

### Pre-Registration

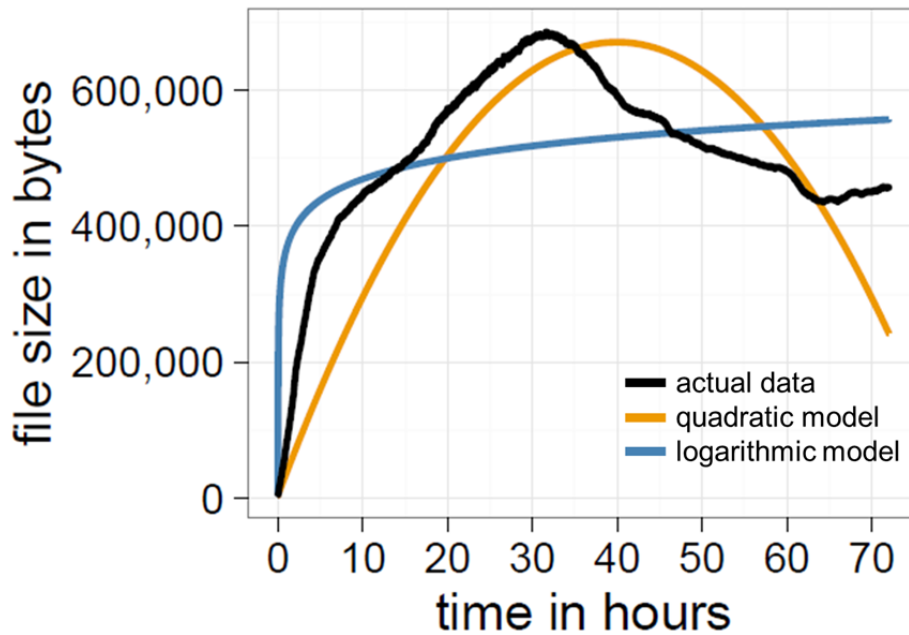
All of our predictions, with the specific time courses for the measures, have been registered on the Open Science Framework before preparing the raw data (<https://osf.io/cx67g/>).

## 3.4 Results

### Does compression follow a predictable time course?

To test for whether compression follows a predictable time course in Place we specified three competing models: a linear model, a logarithmic model, and a quadratic model. Compressed file sizes were predicted by time only, but the models differed in the way this relation was formulated. The intercept in all models was set to 4,372 bytes, as that was the (*a priori* known) compressed file size of the initial empty canvas. After the models had been fit, they were compared using the Akaike information criterion (AIC); an estimator of the model fit, taking into account the number of parameters for a given set of data. For the linear and logarithmic models, the canvas is not predicted to become more compressed as file sizes increase monotonically. Only the quadratic model predicts higher levels of compression: file sizes initially increase, reach a peak, and then decrease at the latter stages. As predicted, we find that a quadratic model provides the best fit of the data (linear AIC: 7,199,987; logarithmic AIC: 6,701,407; quadratic AIC: 6,687,914;  $\Delta$ AIC two

best models: 13,493; Fig. 3.8), supporting our contention that Place does become more compressed and follows a predictable trajectory.

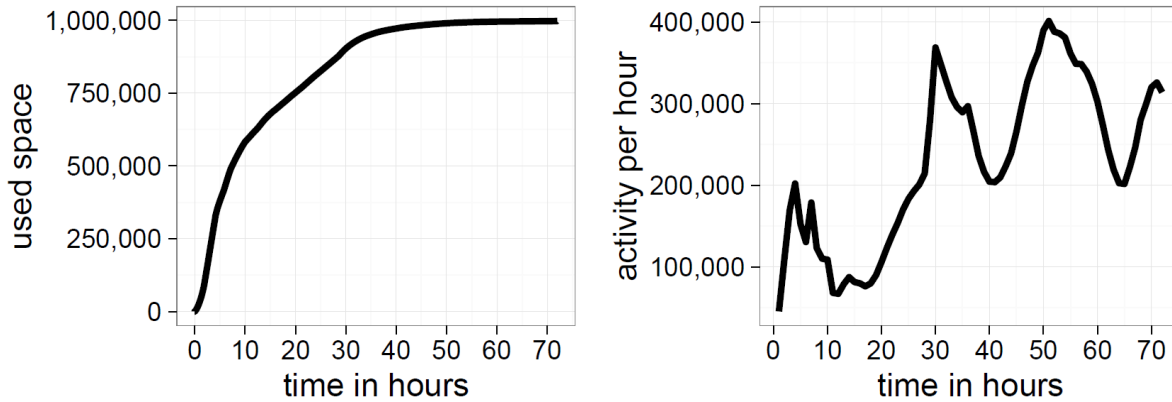


**Fig. 3.8. The time course of compression.** File sizes in bytes are plotted in black. The quadratic and logarithmic model are plotted in orange and blue, respectively. The quadratic model showed the best fit to the data.

One explanation for this trajectory is that the density of the canvas modulates the pressure for compressible patterns. If the spread of compressible patterns is contingent on the density of the canvas, then we predict that compression only spreads when the canvas is densely saturated with artworks. Our reasoning is that a higher density increases the probability of competition, with more compressible artworks having an advantage as these are easier to produce, maintain, and generalize than less compressible ones.

To explore this hypothesis further, we constructed a more complex regression model, using the availability of space on the canvas (limited by the boundedness of Place), the activity levels of individuals, and time as a linear predictor of file size (plus associated interactions). We measured the amount of space by calculating the frequency of used pixel placements at a given time-step. A pixel qualifies as used if any user had previously placed a pixel at that particular coordinate on the

canvas. As such, the number of used pixels should increase as a function of time (see Fig. 3.9). Activity refers to the number of pixels placed on the canvas at a specific time-step (see Fig. 3.9). Time was included to separate the variables of interest from its trajectory of a constant increase.



**Fig. 3.9. User activity and number of used pixels.** **Left:** the amount of used space (i.e., pixels) for the duration of Place. **Right:** the amount of user activity for the duration of Place (total number of changes per hour).

In isolation, more used space should increase variation on the canvas, resulting in less compressible images. However, the interaction between used space and activity captures the increased pressure for compression, resulting in a reversal of the effect (i.e., the direction of the effect changes in the interaction when compared to used pixels as a predictor). Our rationale for this is that higher rates of activity and more used space act as a proxy for increased levels of competition. The model confirms our expectation that levels of competition predict compression: the interaction between higher rates of activity and more used space is a significant predictor of decreases in file size (cf. Table 3.1).

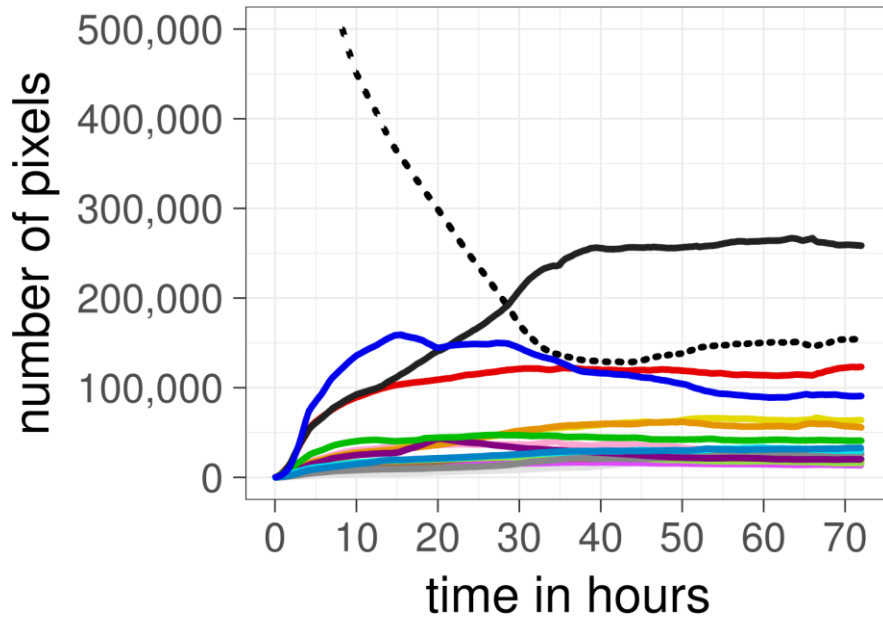
**Table 3.1. Results of regression model with activity, used pixels, and time (plus interactions) as predictors of compression (file sizes).**

|                           | Estimate $\beta$ | Std. Error     | t value | p value |
|---------------------------|------------------|----------------|---------|---------|
| (Intercept)               | 50,940           | 379            | 134.41  | <0.001  |
| Used pixels               | 0.9032           | 0.001345       | 671.51  | <0.001  |
| Activity                  | -123             | 10.07          | -12.21  | <0.001  |
| Time                      | -7.696           | 0.04325        | -177.92 | <0.001  |
| Used pixels:Activity      | -0.002131        | 0.00001910     | -111.57 | <0.001  |
| Used pixels:Time          | 0.000005548      | 0.00000003718  | 149.22  | <0.001  |
| Activity:Time             | 0.2103           | 0.0006544      | 321.33  | <0.001  |
| Used pixels:Activity:Time | -0.0000002009    | 0.000000005795 | -346.61 | <0.001  |

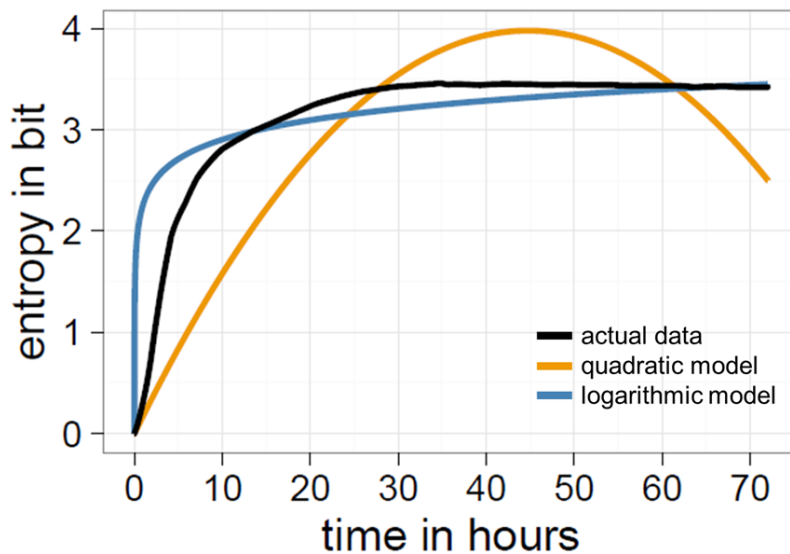
### **Is the trajectory of compression explained by homogenizing or structuring processes?**

One possibility is that the observed decrease in compression is being driven by changes to the frequency distribution of colors. A canvas with a uniform distribution (i.e., all 16 colors are represented at an equal frequency) is less compressible than a non-uniform distribution (e.g., the canvas is homogeneous and consists of a single color). To capture changes in the frequency distribution of colors we measured the amount of global entropy. In parallel with the analysis for compression, we specified three models: a linear model, a logarithmic model, and a quadratic model. Again, all models had in common that global entropy values were predicted by time only, but differed in the way this relation was formulated. Additionally, the intercept in all models was set to zero, as it was known that the white space at the beginning would show an entropy of 0 bit. Models were then compared using the AIC.

If changes in global entropy account for changes in compression, the trajectory of this distribution should approximate a quadratic function. Instead, the distribution of colors moves away from the non-uniform distribution at the start (see Fig. 3.10) and approaches an asymptote of approximately 3.4 bit, with a logarithmic model providing the best fit of the data (linear AIC: 874,314; logarithmic AIC: 264,064; quadratic AIC: 485,931;  $\Delta$ AIC two best models: 221,867; Fig. 3.11). This shows the opposite to what we would expect if the color distribution alone was the main source of compression: at the start, the distribution of colors in Place is highly non-uniform, as white tiles dominate the canvas, but as more colors are used the frequency distribution moves towards a uniform configuration (although it never becomes entirely uniform; global entropy for final time-step: 3.42 bit).



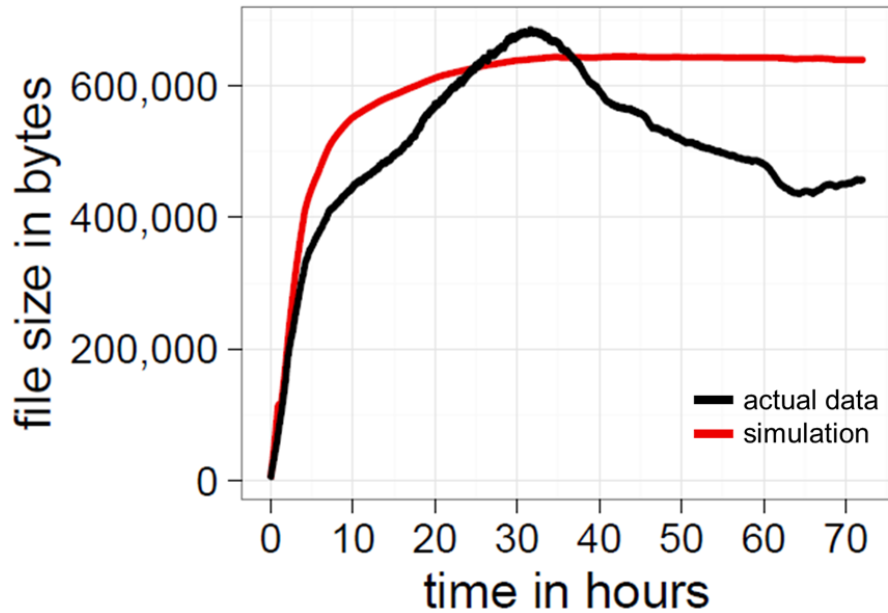
**Fig. 3.10. Frequency distribution of the 16 colors in Place over time.** White pixels are represented by the dotted line and the remaining 15 colors are depicted in their original hue. Note that the canvas started with all of the one million pixels in white (a completely non-uniform distribution). Over time, each color approaches its own asymptotic level and stabilizes at approximately the 50-hour mark.



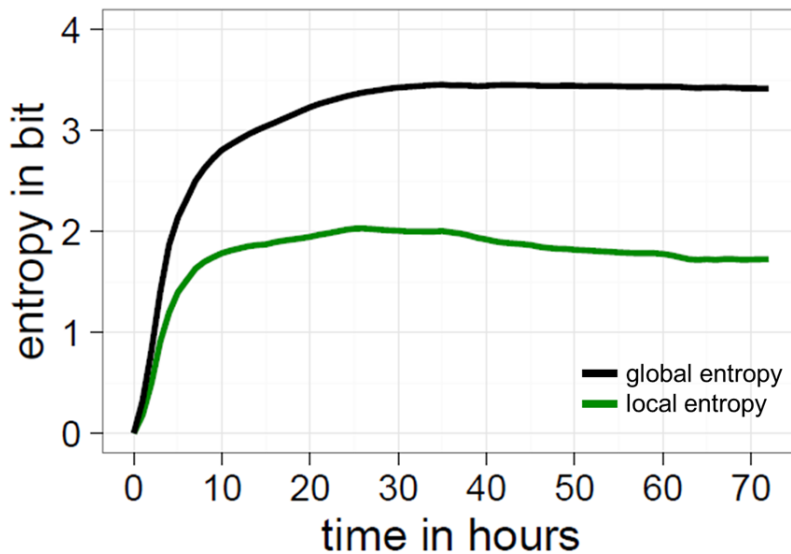
**Fig. 3.11. The time course of global entropy.** Entropy values in bit are plotted in black. The quadratic and logarithmic model are plotted in orange and blue, respectively. The logarithmic model showed the best fit to the data.

Another possibility is that compression is being driven by the organization of variation into structured, predictable patterns. We devised two tests for this. The first is to compare the actual compression data to simulations of randomly shuffled versions of the entire canvas, where the frequency distribution of colored pixels is controlled for at each time-step (compare with Fig. 3.4). Specifically, 100 shuffled images were created for each of the 259,194 time-steps, using the original color frequencies of a specific time-step and assigning them to a randomly chosen position on a new 1000x1000 canvas. These randomly shuffled canvases were then compressed to .png files as described in the compression section. Randomly shuffling the canvas in this way allows us to preserve the frequency distribution of colors and remove the contribution of structure in making the canvas more compressible. This comparison reveals that early on, but, more importantly, also during the entire second half of the run-time of Place, the canvas was more compressible than random placement of the same colors (Fig. 3.12). In fact, the compression outcomes in the simulations mirror the logarithmic trajectory of global entropy, confirming that the DEFLATE algorithm is also extracting structural regularities in pixel placements to increase compression (and not just relying on the frequency distribution of colors).

For our second test, we divided up the space into 10x10 pixel regions and measured the local entropy per pixel. Local entropy tells us whether the pixel placements within a given region are organized into more or less predictable frequency distributions. Contrasting our local and global measures of entropy therefore provides a proxy for discriminating between homogeneity and structuring processes. Homogeneity should show low global and low local entropy, whereas structure should show high global and low local entropy. At the early stages of Place local and global entropy are low, confirming our original expectation that compression is mainly being driven by the homogeneity of the space (i.e., white, unused pixels dominate). Yet, at the latter stages of Place, we find that global entropy is far higher than the local entropy (see Fig. 3.13 and Fig. 3.14). Taken together, both the random simulation and local entropy show that the evolution of compressible patterns in the latter half of Place is mainly due to the underlying structure of artworks, and not simply the spreading of homogeneity.

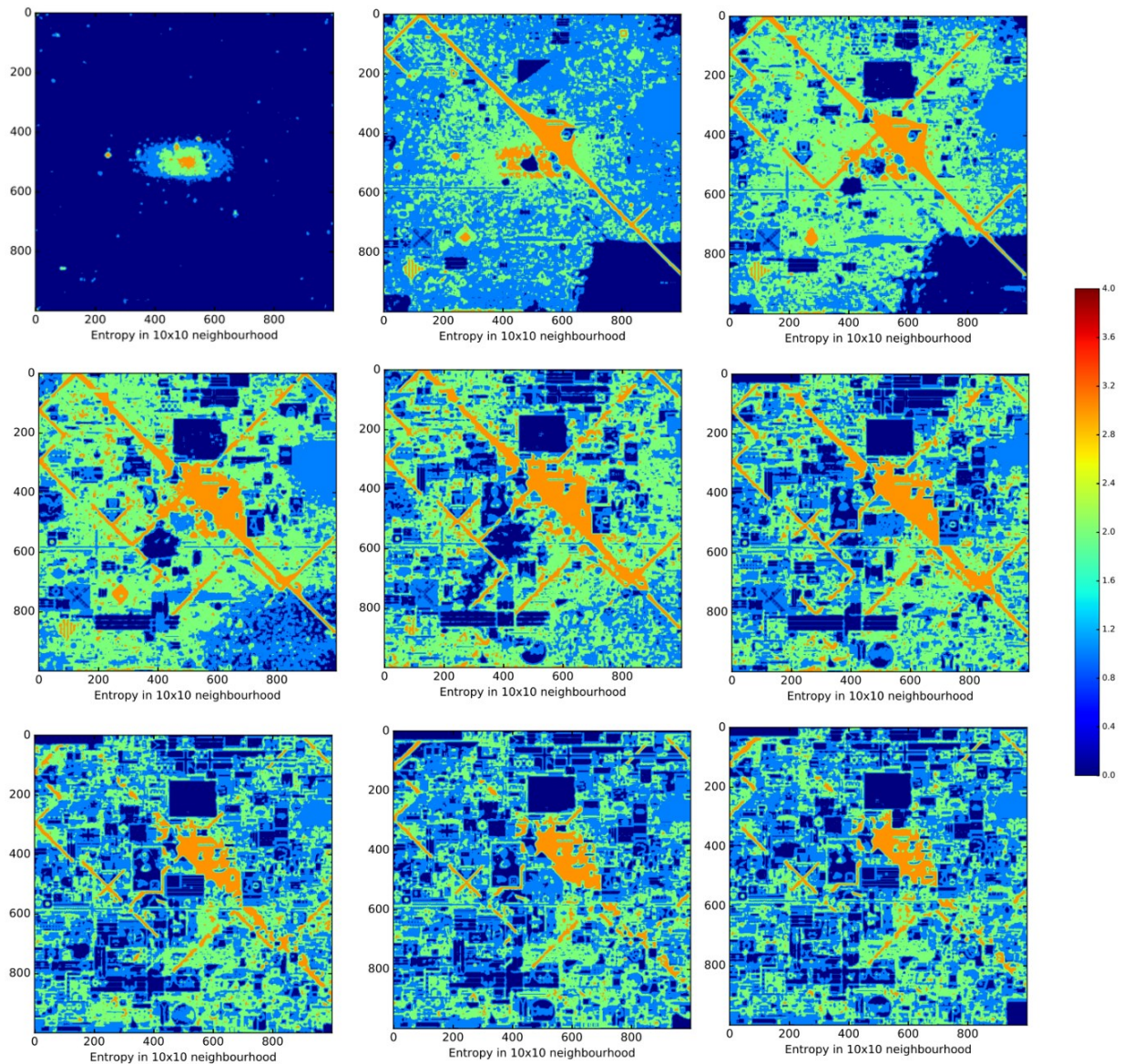


**Fig. 3.12. Comparison of compression between the actual data and the simulation.** The black line represents the original file sizes and the red line the mean values for simulated images at each time step. For Place, the actual images are more compressible than randomly shuffled versions for a period during the first half and during the entire second half of the run-time.



**Fig. 3.13. Comparison of summary entropy values both globally and locally.** Whilst both types of entropy follow similar trajectories through time, global entropy (black line) remains much higher than local entropy (green line).

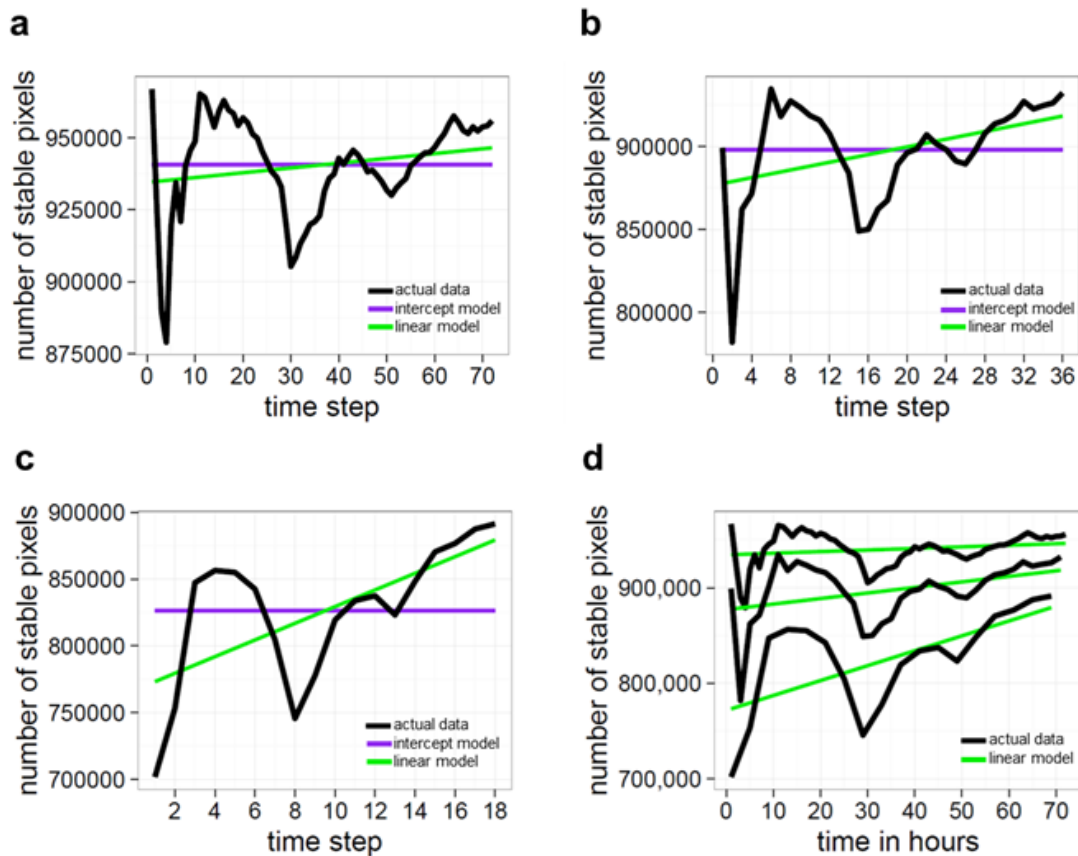




**Fig. 3.14. Heatmap of local entropy for nine time-stamps.** Several heatmaps showing the local entropy for the following time-stamps (from top-left to bottom-right): 1-hour, 9-hours, 17-hours, 25-hours, 33-hours, 41-hours, 49-hours, 57-hours, 72-hours. The axis on the right shows the color-coded entropy values (0 bit: dark blue and 4 bit: dark red). At 1-hour, the canvas is highly predictable at a local level, as the dark blue values indicate, and gradually becomes less predictable as time progresses. Between 25- and 41-hours Place reaches its highest levels of entropy locally (as highlighted by the large regions of orange and green), before gradually decreasing the entropy again (as the growing number of blue regions shows).

### Does the canvas become more stable over time?

If stability is increasing, we should see this across three different time resolutions, with the signal becoming weaker as the resolution narrows (the ability to detect differences in stability diminishes due to the perfect retention of pixel placements between adjacent time-steps; see Methods for more details). Measuring stability in this way provides a robust measure for detecting a trend in the data. To test this hypothesis, we compared intercept-only models to linear models, and predicted that increasing slopes in the linear models outperform a simple intercept-only.



**Fig. 3.15. The development of stability over time.** The black lines represent the number of stable pixels between two time slices, plotted for three different resolutions. Purple and green lines show intercept-only and linear models fitted to the data, respectively. **a)** Number of stable pixels over time for 1-hour-intervals. **b)** Number of stable pixels over time for 2-hour-intervals. **c)** Number of stable pixels over time for 4-hour-intervals. **d)** All three levels of granularity in one figure. There is a positive slope in the linear models for each, but the slope is steeper at coarser time intervals (as indicated by the lower numbers of stable pixels, more change can occur in these within a single time step).

As predicted, all model comparisons show a positive increase in the trajectory of stability (1-hour-slices: intercept only AIC: 1,609; linear AIC: 1,608;  $\Delta$ AIC: 1; direct comparison:  $F = 3.11$ ,  $p = .082$ ; 2-hour-slices: intercept only AIC: 848; linear AIC: 844;  $\Delta$ AIC: 4; direct comparison:  $F = 6.58$ ,  $p = .015$ ; 4-hour-slices: intercept only AIC: 445; linear AIC: 437;  $\Delta$ AIC: 8; direct comparison:  $F = 11.20$ ,  $p = .004$ ; Fig. 3.15a-d). Although the effect in 1-hour-slices is marginal, this is consistent with our expectation that the signal becomes weaker as the time between slices decreases. Overall, these results show the latter stages of Place were less chaotic, suggesting participants were more likely to retain (and build upon) pre-existing artworks.

### 3.5 Discussion

Using a novel, large-scale dataset, which links the interactions of more than one million individuals to the emergence of population-level pixel artworks, our results demonstrate the following: (i) That compression follows a predictable trajectory through time; (ii) Compression at the latter stages of Place is mainly driven by the canvas becoming more structured; (iii) The canvas becomes increasingly stable. From a non-structured state, where individual artworks exist relatively independently from one another, Place gradually transitions to a structured state where pixel placements form specialized, interdependent patterns. We provide three lines of evidence for this. First, the trajectories for variation and compression become decoupled at the latter stages of Place (as indexed by our random simulation results). Second, the canvas becomes increasingly stable, lending weight to the idea that pre-existing artworks are being maintained and extended. Lastly, the local regions of pixel placements remain low in entropy, whereas globally the entropy remains high, suggesting that variation is being maintained at a global level and organized into predictable distributions at a local level.

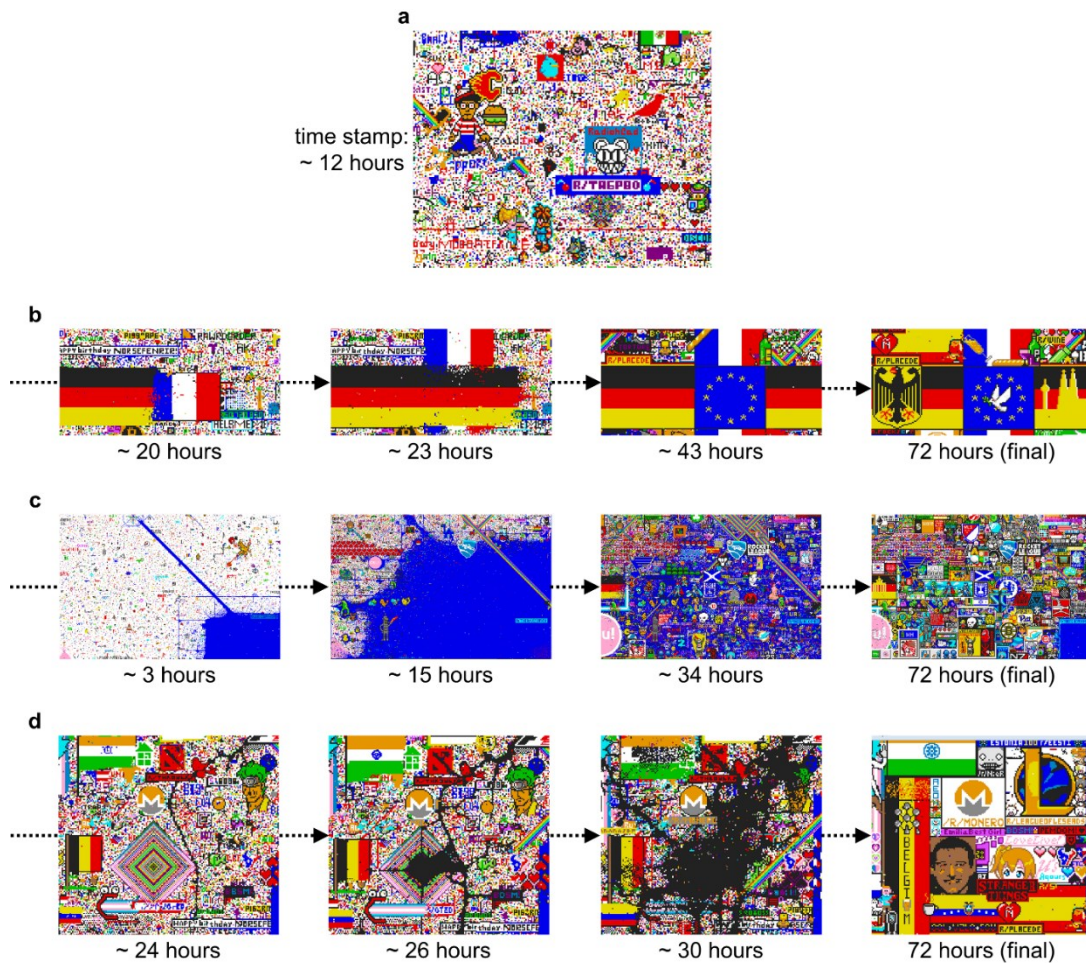
A major implication of our results is that the overall trajectory of compression is dependent on the density of the canvas. This is reflected in the non-linear time-course of compression: during the early phases, when the canvas is sparsely populated with artworks, Place becomes less compressed as more pixels are placed, yet it becomes more compressed at the latter stages when the canvas is densely saturated with artworks. Additional evidence is found in the direction of the interaction between activity and the number of used pixels: compression increases with a greater

number of used pixels (i.e., a more densely populated canvas) *and* higher rates of activity (i.e., the number of pixels placed on the canvas at a specific point in time).

Cultural evolutionary dynamics can explain these population-level patterns in terms of competition and coexistence. Borrowing a simple explanatory concept from ecology, known as the *competitive exclusion principle* (i.e., two species competing for the same resource will result in the extinction of one competitor; Darlington, 1972), we argue that artworks must use different resources in order to coexist; otherwise competition takes place, and inevitably leads to the extinction of the losing artwork. In the case of Place, resources correspond to pixel positions, and competition takes place both within and between artworks via the placement of colored pixels by users.

During the early phases, users have the capability to offset competition by creating artworks in unused regions, resulting in a canvas comprised of relatively idiosyncratic and isolated artworks (see Fig. 3.16a). However, it is worth noting that compressible patterns are present at these early phases, as highlighted by the large regions occupied by Blue Corner and Green Lattice (especially around the 10-20 hour mark; Fig. 3.16c). Such instances demonstrate that compressible patterns can emerge and spread, but importantly the overall canvas is still decreasing in compression. This is in contrast to the latter phases where the canvas follows a trajectory of increasing compression, even though it is densely saturated with a complex, interdependent ecosystem of artworks. We hypothesize that one consequence of fewer unused regions is that solutions to the competition-coexistence problem now amplify a preference for compressible patterns: artworks which are easier to produce, maintain, and generalize are favored over other artworks. Compressible artworks are more amenable to preservation, as aspects of the artwork can be repaired and reproduced using a simple, generalizable rule. Having a generalizable rule also facilitates innovation: new artworks can be generated by incorporating regularities and patterns of one artwork into another.

Simplicity is not the only factor underpinning the survival of artworks, otherwise the canvas would be populated with homogeneous solutions such as Blue Corner (Fig. 3.16c). Linking the low levels of homogeneity to the competition-coexistence problem is our second major finding: that increased compression is mainly driven by the canvas becoming more structured, integrated and specialized. By modifying pre-existing artworks to create new innovations, users are able to preserve elements of old artworks whilst facilitating the creation of new ones. We observe this in



**Fig. 3.16. Selected examples in Place and their development over time.** **a)** Some isolated, simple and rather independent artworks in an early phase of the event. Note the chaotic single pixels and white space still dominating most of the space. **b)** The development of the German-French (and later European) region. Germany naturally expanded its (horizontal) flag to the right, taking over French territory. The French (vertical) flag expanded to the top and bottom, and after diplomatic talks, a European alliance was formed; the two artworks became interdependent. Following the formation of this agreement, there was still constant modification within the two flags, adding national icons and sights. **c)** The rise and fall of the Blue Corner. As one of the earliest groups, a subreddit formed that was determined to cover as much ground as possible in blue pixels, starting from the bottom right corner. This expansionary strategy was very successful early on, but later the group had to incorporate small artworks into their territory to be able to maintain it. Only a small portion of the Blue Corner remained upon completion of Place. **d)** Destruction and innovation caused by the actions of the Void. Determined to destroy other creations, members of the Void group strategically attacked pieces of the canvas with black pixels. Ultimately, however, most of their efforts were overwritten over the time course of the event. Note that some, but not all of the original art destroyed by the Void resurfaces at the end; some creations are entirely novel.

Fig. 3.16b: here, the extension of old artworks injects structure into new ones, mitigating competition between artworks in a manner analogous to *niche differentiation* in ecology (i.e., artworks are able to coexist through integration and specialization). Importantly, it is this spreading of structure which accounts for the increased compression in our results.

What remains unanswered is *why* Place becomes more structured as opposed to homogeneous. Arguably, structure seems to provide a more stable and less costly solution to the competition-coexistence problem than homogeneity. One reason for this is due to the diversity of the subreddit groups and their endogenous goals for creating artworks. We hypothesize that it is this diversity which results in diminishing returns for homogeneous artworks. To dominate the canvas, a homogeneous artwork would need to expend resources on growing as well as repairing damage to the artwork, with both factors minimizing the competitive advantage of simplicity. In this sense, spreading homogeneity becomes increasingly costly because it requires large and well-organized groups to mitigate the countervailing effects of other groups (see Fig. 3.16c-d). Had there been an overarching task for placing pixels, such as dominating as much space as possible, then it might have made homogeneous solutions more viable as groups could work together towards a common goal.

It is important to address the role played by white pixels in Place. One potential confound is that the initial state was a million white pixels and users could select white as a color to place on the canvas. This raises the question as to whether or not the white space should be considered part of the culturally evolving artefact. We decided to fully include white as a color in our analyses for the following reasons: First, it is unclear which cases of unused space were actually incorporated into artworks by users and which were simply not in use. Second, the potential confound is only really problematic during the first 30 hours of the runtime of Place, before the space is saturated with artworks. In contrast, our results mainly focus on the trajectories during latter phases of Place. Lastly, our predictions already factored in the effects of white pixels to some extent, which is why we formulated specific time-course trajectories for compression and variation in our models.

Another confound is in the use of the DEFLATE compression algorithm. In principle, randomness should act as an upper bound on the compression of image, yet we clearly see that our simulations produce randomly shuffled images which are more compressible than the original image (at approximately the 30-hour mark; see Fig. 3.12). Such disparities suggests there is a bias

in the way lossless compression algorithms exploit statistical regularities. It is already known that there are sequences with low algorithmic complexity which are not compressible by statistical estimators (Zenil et al., 2018). An example of this is the Thue-Morse sequence where a string is repeatedly appended by taking the Boolean complement of the current sequence, i.e., 0 becomes 01, 01 becomes 0110, and 0110 yields 01101001 etc. This provides a possible explanation as to why the compression algorithm might make the actual image less compressed than it appears. But it is also the case that randomly shuffled pixels will produce statistical regularities by chance. Randomly shuffling a non-uniform distribution of colors can create statistical regularities where there were previously none, with the detection and extraction of these regularities determined by the size of the sliding window used by the DEFLATE algorithm. It is possible that in our case the algorithm is both failing to capture some patterns in the actual image and compressing randomly generated regularities in the shuffled images.

Place provides a powerful model for investigating how cultural evolutionary processes can link spatial and temporal dynamics in producing predictable patterns. Future work is now well-placed to establish whether these findings can be reproduced and generalized to other cultural phenomena. Computational models (Epstein, 1999) and mobile apps (Griffiths, 2015) provide two complementary routes for reproducing the scale of Place. In terms of generalizability, perhaps the most relevant comparison is to artistic traditions. Just as artworks are used to mark the identity of subreddits in Place, so too is art often employed to mark familial lineage via heraldic emblems (Morin & Miton, 2018) and group membership via graffiti tags (Adams & Winter, 1997). Another point of comparison with graffiti is the collaborative nature of Place; a single artwork comprises the contributions of numerous individuals. Lastly, artistic traditions can also borrow from one another, as seen in the influence of Japanese woodblock prints on impressionist and post impressionist styles (Sullivan, 1989), which parallels the direct and indirect cross-fertilization of ideas and styles in Place.

### **3.6 Conclusion**

Despite the cross-cultural heterogeneity of human societies, cultural evolutionary processes often deliver surprisingly convergent outcomes. We set out to investigate one of these proposed outcomes: the spread of compressible patterns. In particular, we used a novel, large-scale dataset

to witness the *de novo* emergence of a culture in real-time: from an initially white canvas, where patterns of activity were idiosyncratic and independent, over a million individuals came together to produce a complex, interdependent ecosystem of artworks. Place not only demonstrates that the spread of compressible patterns is a signature of cultural evolution, it also highlights the interaction of temporal and spatial dynamics in determining the trajectories of change.

### **Acknowledgments**

We would like to thank Olivier Morin for reading and commenting on the pre-registration draft. Additional thanks to the Max Planck IT staff (Marcel Sommer and Jürgen Rosenstengel) for their assistance in setting up a server for the simulation part of the study.



## **4. Color Terms: Natural Language Semantic Structure and Artificial Language Structure Formation in a Large-Scale Online Smartphone Application**

This chapter represents a study currently under revision at the *Journal of Cognitive Psychology*, authored by T. F. Müller, J. Winters, T. Morisseau, I. Noveck, and O. Morin (in this order).

Date of submission: July 18, 2020

Author contributions: The Color Game application was conceptualized by Olivier Morin, with the help of James Winters, Tiffany Morisseau, and myself. The concept of this study was developed by me, and revised with the help of all authors. The additional online survey was conceptualized by James Winters, Olivier Morin, Tiffany Morisseau, and me, and programmed and conducted by James Winters. Data curation for the general Color Game data was performed by Olivier Morin and Tiffany Morisseau, and specific curation for this study was performed by me. The analyses were entirely performed by me, and figures were created by me and Olivier Morin. The manuscript was written by me, reviewed by all authors and then revised by me.

### **4.1 Abstract**

Human language exhibits structure on many different levels. We investigate semantic structure, i.e. the organization of meaning spaces into discrete categories, in an artificial language game distributed through a large-scale smartphone application. Artificial language studies in the past have addressed the evolution of structure in many ways, but mostly ignored a potential natural language bias. We compare color terms from natural to artificial language to assess the similarity of their semantic structures, and investigate the influence of the semantic structure on artificial language communication. The goal for our participant pairs was to communicate the correct color from an array of four colors, using only black-and-white symbols. We compare the in-game communication to a separate online naming task providing us with the natural language structure. Our results show that natural and artificial language structure overlap at least moderately. Furthermore, communicative performance and the number of sent symbols were influenced by the shared semantic structure, but only for English pairs. These results imply a cognitive link between participants' semantic structures and artificial language structure formation.

## 4.2 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 that are 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 summarized with 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.

However, the exact nature of and the processes involved in the evolution of these different structural features of language are still not fully explained and merit further investigation. 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 the natural languages. This is important because almost none of the past studies that evolved structure in an artificial language game have been concerned with a possible *bias for natural 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 natural 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 highly 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). Typically, the experiments involve either i) repeated interaction between the same individuals, to observe the convention formation in this closed communication; or ii) repeated transmission of the artificial language from one individual to another, set up in a chain; or iii) both interaction and transmission.

To put one example into more concrete terms, the study by Nölle et al. (2018) investigated how structure in an artificial language arises under different environmental circumstances. To this purpose, the authors recorded the silent, spontaneous gestures participants used to describe cartoon characters to one another, chosen from a selection that changed in every trial. Thus, this task involved repeated interaction between two partners, but no transmission of the signals to new pairs. By coding whether any two gestures used by a pair were the same, and manipulating (among other things) the structure of the selections that participants were provided with, the authors could show that communication systems adapt to different environments by systematically coding meaningful features of the stimuli (dependent on the selections) with the same gesture.

Because language exhibits structure on so many different levels, it is important to distinguish the linguistic structure that is evolved and measured in any specific experiment from other kinds of structure. One 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, acquires systematic and conventional rules about their

combination and mapping to the dimensions of the meaning space (e.g. Christensen et al., 2016; Kirby et al., 2008; Nölle et al., 2018; Selten & Warglien, 2007; Winters et al., 2015, 2018). This refers 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 structure can evolve.

Less attention has been devoted to the latter, the 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 (but see Perfors & Navarro, 2014; Silvey et al., 2015; Xu et al., 2013). This refers to the semantic structure, this time not of a natural but an artificial language. Experiments conducted in this fashion circumvent one possible criticism against studies concerned with structuring the signals: 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 (Nölle et al., 2018; Perfors & Navarro, 2014; Silvey et al., 2015; Winters et al., 2015, 2018), but becomes a hindrance whenever signal 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 compositional structure (one morpheme for black/white and one for cat/dog). This introduces a confound into the first line of experiments (on structuring the signals), making the origin and shape of an evolved signal structure less clear. By employing continuous meaning spaces, the few studies focusing on the semantic structure (Carr et al., 2017; Perfors & Navarro, 2014; Silvey et al., 2015; 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 that of natural language interference with the evolution of semantic structure in these experiments. Participants are already fluent in 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), to our knowledge no study has systematically set out to address this question. The study by Xu et al. (2013) is particularly relevant here: In their artificial language task, they demonstrate that repeated

learning and transmission of initially random partitions of color spaces (i.e. subdivisions of a color space which are named with a single color term) will result in partitions close to color term systems found in the World Color Survey (representing data of over 100 unwritten languages; Kay et al., 2009). Since all the participants in this experiment were native speakers of English and Xu et al. (2013) want to rule out this potential native language bias, they compare 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 find 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 conclude "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 natural language of the studies' participants?

Another relevant topic here is that of crosslinguistic influence of a native language on second language learning (see Kellerman, 1995, for a summary). Second language learners can both suffer and benefit from interference with their native language (Gass, 1987). In particular, the grammatical structures of the native language can be transferred to the second language performance (Zobl, 1980). Crucially, however, participants in artificial language games on the semantic structure (Carr et al., 2017; Perfors & Navarro, 2014; Silvey et al., 2015; Xu et al., 2013) are not tasked with learning the existing rules of a linguistic system, but create these rules themselves to form a set of shared conventions. This lets us abstract away from the special case of a specific native language coming together with a specific second language (e.g. "Do learners of English profit from knowing German?") and ask more generally whether semantic structures evolving in these experiments reflect the semantic structure of participants' native languages. This is our first research question: (1) How similar are the semantic structures in natural language 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 point 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; 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 (see Method).

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 for this top-down application of the broader term of categorical perception.

One special task that might engage this high-level processing of the color terms that has seen not much attention is referential communication. This task is characterized not by mere naming of colors by sole individuals, but by communicating a target color successfully to an interlocutor. 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. A systematic test in a communication paradigm is difficult within the boundaries of a natural language, since communicating colors accurately is then almost trivial; instead, an artificial language is required. Communication in such a task is hard since there is high uncertainty about the messages, because the conventions are not yet established. 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 natural language influence communication with an artificial language?

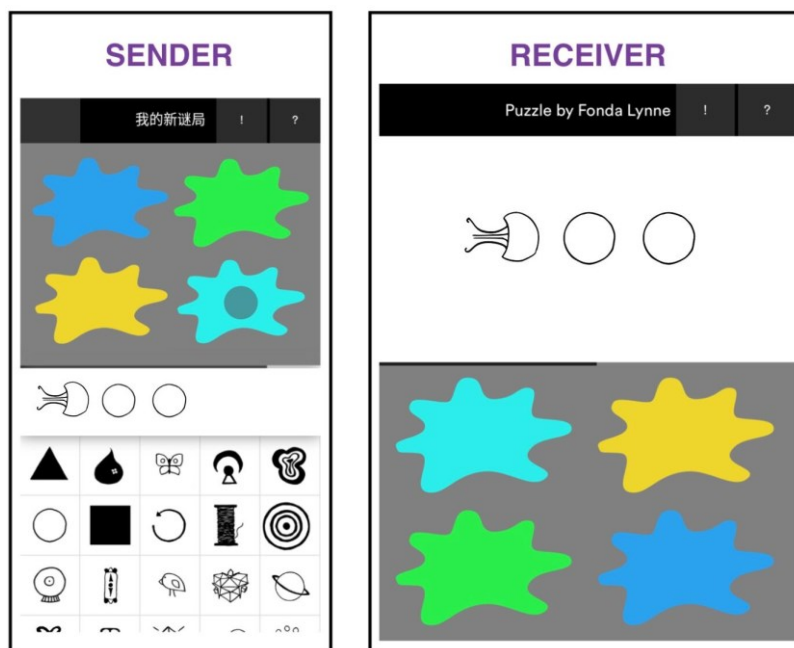
Lastly, sharing a common language obviously makes communication easier. Relying on shared conventions, interlocutors profit in their interaction, since they are closer to mutual understanding already (Lewis, 1969). Transferring this to signaling systems in artificial language games, it applies to the 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). What follows is 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 natural language structure could mean that speakers of different native 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 natural 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 natural languages experience more problems in communication?

### 4.3 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. During this time, anyone could download the game and, after a short tutorial, play the game with another player in one of several game modes. These game modes all shared the basic structure of a referential communication task: One of the players (the sender) was tasked to communicate a color, using an artificial language comprised of black-and-white symbols, to the other player (the receiver), who then had to pick this color out of an array of four colors in total. Fig. 4.1 shows an example trial and the view from both sides. For the entire color space, see Fig. 4.2, and for the set of symbols, Fig. 4.3.

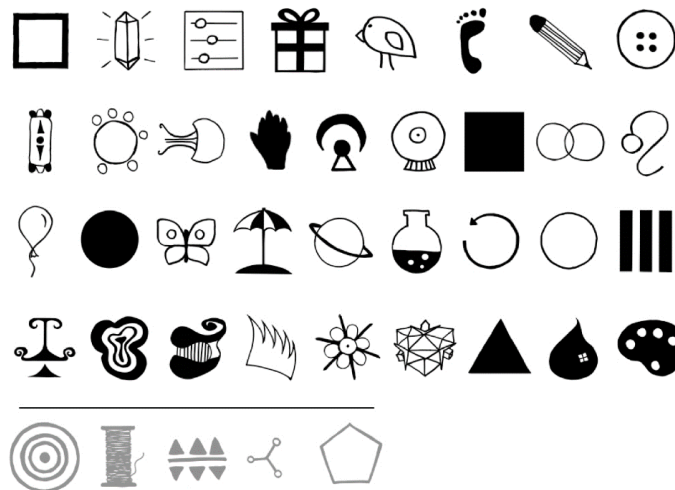


**Fig. 4.1. An example trial in synchronous mode.** The sender in the current trial communicates with a receiver to help find the target color (here, the brighter shade of blue), marked for the sender by a dot.





**Fig. 4.2. 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 (cf. Luo et al., 2001), chosen 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; this includes the first and last color, meaning that the space is circular and can be represented as on the right).



**Fig. 4.3. 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. The symbols were chosen so that they had ambiguous associations with regard to the colors they could be used for to encode, but at the same time allowed the players to solve the communication task above chance level (see Müller et al., 2019, for a discussion of this approach).

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. The app's source code is open and the full raw data will be made available after a period of embargo. The processed data files and code of the current study can be accessed on the Open Science Framework (<https://osf.io/a8bge/>). 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/>. 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 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: <https://osf.io/9pdzk/files/>.

Participants 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 of four colors. 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). Upon entering the game, players were presented with a lobby showing the pseudonyms of all available other players, how many puzzles these players currently offered for others to solve as receiver, and whether they were currently online. They could then decide to enter a game in one of the different game modes. Given that players were free to choose with whom to interact in the game, but also played with one partner for at least a single set of 10 trials, we simulate repeated interaction between partners as well as transmission of the artificial languages that get formed to new players within the game.

Regardless of the mode, the game did not provide feedback to the players, 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 series 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 trial, both players were awarded points, depending on their performance. With increasing points, players also unlocked more symbols that could be used as a sender (in random order), and after all symbols had been unlocked, an additional speed game mode. The different

game modes were included to give more variability to players other than simply the basic task, and in particular to assess the importance of live interaction, and also to allow for content that players can 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.

Crucial for our purposes are the colors that receivers had to choose from. Using the CIE2000 color space, we constructed a circular selection of 32 colors that are perceptually equidistant (Fig. 4.2). 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. On every trial, an array was chosen randomly, as was the target color. This left 3 remaining colors of the array in the role of *distractors* for the current trial, since they were incorrect responses for the receiver. Here, an additional manipulation was implemented, which was relevant for a different project (<https://osf.io/qz597/>): Senders did not always have access to all 4 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.

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).

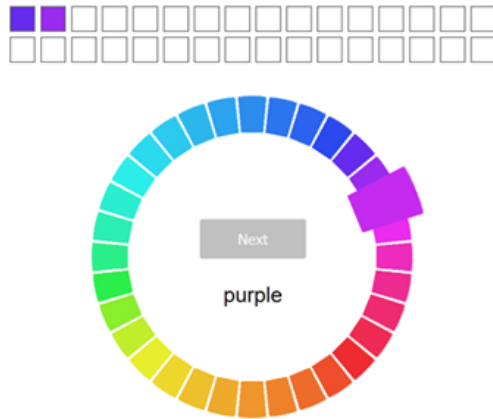
## **Online Survey**

We need baseline data to find the semantic structure that speakers of the three languages use for our (unique) set of colors. For this purpose, we set up an additional and separate online study on Prolific (an online portal to conduct scientific surveys), somewhat similar to the method of the

World Color Survey (Kay et al., 2009). We operationalize the natural language semantic structure in the form of basic color terms much like Berlin and Kay (1969), an approach that has been 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 of the world 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).

This study and the Color Game study used a different sample of subjects. Only participants that were a native speaker of one of the specific languages were able to access the respective online survey. The 32 colors of the application’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 (with a randomized starting point; see Fig. 4.4). The participants must then provide a label for every color by associating it with at least one basic color term (presented one at a time). The terms were presented one by one and in random order.

Basic color terms 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 Fig. 4.5. Given the close relatedness between the three languages, this seemed appropriate.



**Fig. 4.4. 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. After submitting choices for a term with the “next” button in the middle, participants were presented with the next term in their language, until all colors had been named.

|         | 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 4.5. Color terms used for the online survey in the three languages.** 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 of the way 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.

Participants in the online survey 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. For each of the three languages, we got survey data for 50 individuals. Bringing together the data on the natural 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. Based on the research questions, we made the following predictions:

**Research Question 1: How similar are the semantic structures in natural language and an artificial language?**

**Prediction 1.** We predicted that the factorial structure (assessed by exploratory factor analysis) of the natural language should not invalidate a confirmatory model on the structure of the artificial language. This would indicate that the artificial language structure reflects the natural language structure, to some degree.

**Research Question 2: Does the semantic structure inferred from the natural 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 factors to be easier to solve for participant pairs (of the same native language) as compared to color arrays that cross fewer of these boundaries.

**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 factor in the natural 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-factor 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-factor 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-factor 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 factor 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-factor 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-factor 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 natural languages experience more problems in communication?**

**Prediction 3.** Here, the prediction is that the coordination problems arising from different natural 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.

#### **4.4 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 material), 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 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 trails played together for individual pairs (regardless of the distribution of role in the pair as sender or receiver). The number of users and trials in the specific subsets of relevant languages (from the preregistration) can be seen in Table 4.1.

**Table 4.1. Number of senders and same-language pairs that reached the preregistered thresholds of trials for each of the 5 languages we intended to use.**

| 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           |

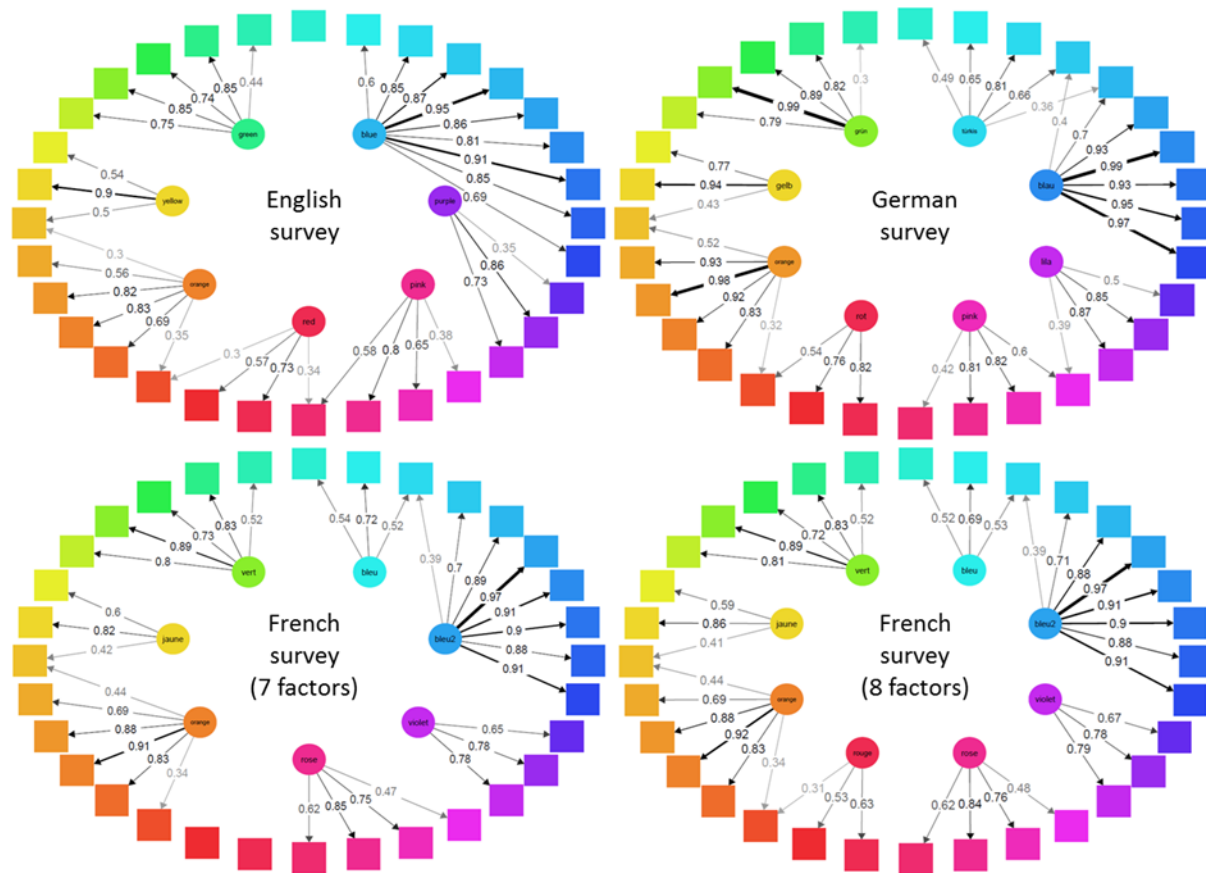
Although we did not reach the cutoffs preregistered for an early analysis in most cases, there is still enough data to perform the planned analyses fully for English, German, and French. Spanish and Chinese had to be dropped due to the lack of data of players with a high number of trials; since it was 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. As per the preregistration, the resulting data



was restructured to represent one row for each participant-term pair (compare also Jäger, 2012). Then the exploratory factor analysis (*EFA* from here on) was applied using R version 3.5.1 (R Core Team, 2017). More specifically, the number of factors to be extracted was set to the number of basic color terms determined in the piloting study (see Method section) and used throughout the survey, and promax rotation was applied to facilitate interpretation of the results. The goal of these analyses was to see whether the data would reflect our assumptions concerning the structure, resulting in clean factors that represent the natural language terms and show low cross-loadings, with clear boundaries which then could be used for the further analyses.

The factorial structure of the data from the naming task for the three languages is visualized in Fig. 4.6. In English and German, it reflects the natural 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 factors, which are easily identifiable as the respective color terms. Likewise, we observe the highest loadings on colors that are in the center of factors 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. However, 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” factor in favor of a “light blue” factor (such as in German). Still, cross-loadings are also low and boundaries clear-cut. The “red” factor 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 4.6.** Visualizations of the factorial structure resulting from the EFA on the data of the online survey. Boxes represent the 32 colors used in our study. Circles represent factors, named with the term that was most frequently associated with the colors loading on them; these factors 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. 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 as per the preregistration. As can be seen in Table 4.2, 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 factor), and boundaries that showed in the EFA are not contradicted by the loadings of the CFA.

**Table 4.2. Goodness of fit measures (robust estimates) for the CFA on the three languages.**

| 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         |

### **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 factor 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 factors 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 factors, 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.

**Table 4.3. 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  $\Delta AIC$  column implies the simpler model minimized the AIC and was favored over the model including the number of boundaries.

| Language | n     | AIC simple model | AIC model including #boundaries | $\Delta AIC$ | 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                            |

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  $\Delta AIC$  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 4.3. 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 factors (according to English color terms). For German and French, no such effect could be detected.

## **Prediction 2.2**

We then turned to the analyses investigating whether same-factor distractors would impact performance, given the natural 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-factor colors. We then coded each trial for whether the color array presented the sender with a distractor that was part of the same factor as the target, given their natural 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 4.4. 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-factor 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-factor distractor was present in a given trial.

**Table 4.4. AIC values and p-values of the likelihood-ratio tests computed from the models testing the effect of the presence of a same-factor distractor on the number of symbols sent in the given trial.**

| Language          | n    | AIC simple model | AIC model including “distractor” variable | $\Delta$ AIC | 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            |

### Prediction 2.3

After that, we also predicted performance by the presence of same-factor 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-factor 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 ( $\Delta$ AIC of 1.9 and p-value of .049; see Table 4.5). This means that English pairs also tended to perform worse when a same-factor distractor was present in a trial, but this result is less robust. No such effect could be found for German or French, nor for the interaction effect in any of the three languages.

**Table 4.5. AIC values and p-values of the likelihood-ratio tests computed from the models testing the effect of the presence of a same-factor 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  $\Delta$ AIC 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.

| Language | n    | AIC simple model | AIC model including “distractor” variable | AIC model including interaction | $\Delta$ AIC          | Likelihood-ratio tests: p-value |
|----------|------|------------------|---|---------------------------------|-----------------------|---------------------------------|
| English  | 3323 | 3881.8           | 3879.9                                    | 3881.4                          | 1.9;<br>negative      | .049*; .458                     |
| German   | 9356 | 10905.9          | 10907.6                                   | 10909.6                         | negative;<br>negative | .602; .985                      |
| French   | 995  | 1307.5           | 1309.4                                    | 1310.7                          | negative;<br>negative | .725; .408                      |

### 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 4.6). 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.

**Table 4.6. 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.**

| Data subset                              | n     | AIC simple model | AIC model including effect | $\Delta$ AIC | 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                   |

#### 4.5 Discussion

By combining the results of the online survey with our smartphone application, this study was able to compare the Color Game’s artificial 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 natural language of human 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 natural 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, factor 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 factors 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 factor showed very high loadings overall, whereas peripheral colors showed the lowest loadings. This is in agreement with central colors being prototypes within their factor (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 factor for the blue-green colors instead of a “rouge” factor, 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).

An important point is that while our results on the CFA speak for good similarities between natural and artificial language semantic structure, they do not necessarily imply a direct causal link. It might be tempting to argue that natural 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 natural 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 natural 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 natural 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 natural 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 natural language, but not on the contrasting language, this would imply that potential cognitive 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 natural language semantic structure. They also communicated less successfully and sent more symbols when an array contained a distractor that belonged to the same factor 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-factor distractors are harder for participants to delineate clearly in communication than distractors from a different factor. 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 natural 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-factor 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 natural 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-factor 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 factor 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 factor over the expected red factor. 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). This was, however, 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.

#### **4.6 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 natural language structure of native speakers of English, German, and French. To do so, we combined the results of a large-scale online smartphone application specifically programmed for this purpose 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 natural language semantic structures to a moderate to good degree, confirming our expectations. This does not necessarily imply a causal effect of natural 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 natural and artificial language influenced the performance and pragmatics of the artificial language, however

only for native 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 natural languages. Methodologically, we argue 1) that potential biases towards natural 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>.

### **Acknowledgments**

We would like to thank all people that contributed to the Color Game, either in its creation or as a participant. A summary of acknowledgments regarding these contributions can be found here: <https://osf.io/nsxu4/>. We would also like to thank Yoolim Kim for her comments on the L2 interference literature.

## 5. Summary and Conclusion

How do communicative signals emerge and become organized? In this thesis, I set out to answer this central question, with three empirical studies related to smaller questions. As a starting point, I outlined the special features underlying human interaction, namely intention-reading, a special structure focusing on turn-taking and repair, and a particularly cooperative nature, and linked this to how communication represents an extraordinary example for all of these. I continued by introducing the ostensive-inferential approach to communication, an attempt at explaining the mechanisms of communication at the interactional level, adopting it as a model of communication for the thesis. Likewise, I presented an account of cultural evolution as an attempt to explain how cultural traits (such as communicative signals) can be transmitted over multiple generations of learners, potentially shaping the overall population-level patterns. I also outlined how a central goal here is to link processes of micro- and macro-evolution, i.e. to link individual-level interactions to population-level patterns.

The first empirical study of the thesis asked how the shared context affects the successful emergence of communicative conventions at the micro-evolutionary scale (Chapter 2). The study answers this question by showing that in both experiments, access to the shared visual context for the sender enabled higher rates of success when dyads of participants were developing novel conventions through communicative interaction. The second experiment was also able to show that dyads that shared the visual context were more successful in re-using their shared conventions in a new communicative environment, thus generalizing them more efficiently. This is experimental evidence for theories emphasizing the role of context for successful communication, and especially for its emergence (Clark, 1996; Sperber & Wilson, 1996).

In the context of the current thesis, one point that should be emphasized is how relying on an ostensive-inferential framework of communication led to a number of design choices for the experiments, most notably the absence of external feedback and training on specific symbol meanings for the participants. Instead, participants were encouraged to rely on inference and interaction to successfully solve the task. This increases the confidence in arguing that proper novel conventions were emerging between participants in the task, rather than them being limited to using pre-established associations, a question that all language evolution experiments must be concerned about. Further confidence is added in light of the results that participant pairs differed

drastically in their performance and conventions, and that their performance significantly increased over time. The possible limitation and skepticism regarding the novelty of conventions developing in these tasks is also one of the central questions behind the research presented in Chapter 4; and importantly, the question of whether any conventional associations of this sort are necessary or not to achieve communication is one of the key differences between the ostensive-inferential model and code models of communication, as outlined in the introduction.

Chapter 2 showcases an example of how theoretical assumptions from pragmatic frameworks can fruitfully be translated and tested within an artificial language paradigm. Future work would be well placed to focus, for instance, on more detailed aspects of the shared context such as the impact of the contextual information being available to either senders or receivers; or in fact on other special features of human communication. One particularly interesting feature here that has seen some interest in the context of artificial language games (Garrod et al., 2007; Healey et al., 2007) but deserves more attention with respect to its potential importance for the emergence of communicative signals is the phenomenon of conversational repair, i.e. the correction of misunderstandings. It could very well be that while being culturally universal (Dingemanse et al., 2013, 2015) and a key part of the human interaction engine (Levinson, 2006), repair mechanisms might boost the success and efficiency of emerging communicative signals.

The second study of the thesis tried to answer the question of how cultural traits can get organized through interaction, at a macro-evolutionary level (Chapter 3). A major finding was that in the case of the large-scale online art collaboration, compressible patterns developed and their evolution through time could be predicted, building on the idea of a rising pressure for simplification due to increasing competition. The second major finding was that this compression trajectory could be shown to occur due to the evolution of systematic structure, and not mere increasing homogeneity. These results are relevant in light of the known compression effects that occur in communication systems as a response to other simplifying pressures such as the need to memorize the cultural traits (Kirby et al., 2015; Tamariz & Kirby, 2015), expanding the picture by presenting data from an art collaboration in which pressure increased due to the limited amount of space for the expression of cultural traits.

The organization of visual information into predictable patterns is only a single example for how cultural systems can exhibit structure, however. As such, the study naturally is limited in the

breadth of its application as a single case study. Especially in language, structure is ubiquitous (Everaert et al., 2015), present in grammar, semantics, and phonology, to name some examples (it is the semantic structure of natural and artificial languages that was also at the center of Chapter 4). Another limitation that stems from the study being an observational case study is that it remains unclear why structure rather than homogeneity prevails as an outcome in Chapter 3. A plausible explanation is that the inherent group structure of the art collaboration might have contributed greatly to this; rather than working towards a single common goal, the large diversity of subgroups with different goals might have prompted participants to split the canvas into structured subregions. However, without a control condition due to the collaboration being a one-time past event that cannot be repeated faithfully, this cannot be tested thoroughly. What would be needed for future studies to address this question is a clean manipulation by seeding a population of participants into different group structures, e.g. one big group versus several smaller ones, or keeping the participants in a similar structure but letting them work towards a common goal in one condition. This would disentangle some of the possible explanations that could lead to either structure or homogeneity. However, it is improbable if not impossible that any controlled scientific project could be created that would reach similar numbers in sheer scale, with over a million participants.

In this way, the study is a unique showcase for the phenomenon of compression as a repeated outcome of cultural evolutionary processes. This is its main implication in the context of the present thesis; it also suggests similar potential patterns in the organization of communicative signals, especially in light of the existing literature on this topic (e.g. Kirby et al., 2015). Future studies, including those that are concerned with communication systems more directly, could especially make use of the distinction between structure and homogeneity, both theoretically defined and methodologically applied within Chapter 3. By simulating random distributions of signals that are equal in variation and by considering also local (in addition to global) patterns, future work could get a more nuanced picture than simply distinguishing between more and less compressible patterns; for example by showing that the emerging structure is organized in a non-random way, given its variation. Another strength of the chapter lies in the way it considers both temporal and spatial dynamics and their effects on the overall outcomes on the collaborative space. Although purely focused on macro-evolutionary patterns, the data allows for high-resolution analyses of their trajectories over time.



The third and last study presented in the thesis tried to answer the question of how existing communicative conventions compare to novel jointly created conventions and their usage in interaction, from a macro- to a micro-evolutionary scale (Chapter 4). Comparing the semantic structures developing within an artificial language game to the respective natural language ones of speakers of English, German, and French, the chapter found a good to moderate correspondence. As a second main result, it was also found that the performance and the number of sent symbols were predicted by this shared semantic structure, but only for English participant pairs.

These results imply that there is a cognitive link between natural language semantic structure and artificial language semantic structure, however the precise nature of this link is still unclear. Due to the large overlap between the semantic structures of the three closely connected Indo-European languages in the sample, a direct comparison from one structure to another was impossible. Future work could directly compare two languages that differ drastically in their semantic structures. Here, a possible insightful result could be that native speakers of one language make use of an artificial language structure similar to the natural language that is a bad fit on the second language's semantic structure. This would suggest that artificial language structures are directly affected by the respective natural language semantic structure. Whatever the exact link, one major implication for artificial language games from the study presented in Chapter 4 is that a potential bias towards natural language structure should not be neglected. In particular when the main theoretical focus of a study relies on the structure that develops during the experiment, it is crucial to show that it did not develop due to biases other than the supposed explanatory variable.

Chapter 4 is of central importance for the thesis overall as it links together both previous studies, theoretically as well as methodologically. On the theoretical side, the study is concerned with the emergence of novel signals in a communication task, just as is the case for Chapter 2; but the focus is on their relation to natural language rather than the contextual dependencies of their success. Additionally, it is concerned with the organization of these cultural traits into structured patterns, similar to Chapter 3; but the focus is on the alignment of semantic structures rather than the compressibility of the visual structure in that chapter. Ultimately, Chapter 4 comes closest to bridging the gap between the micro-evolutionary interactions of participant pairs and the macro-evolutionary patterns of cultural evolution, a small contribution towards addressing the problem of linkage outlined in the introduction (section 1.2). On the methodological side, Chapter 4

compromises between the approaches of the laboratory experiment (Chapter 2) and the analysis of massive data on online behavior (Chapter 3) by employing an online smartphone application. This helps lift several of the limitations of the restricted nature of experiments (such as limited sample size, strongly constrained environments, closed transmission chains, etc.) and the uncontrolled nature of observable online behaviors (limiting the causal explanatory power without any clear manipulations).

Overall, the thesis takes a highly interdisciplinary approach, integrating concepts from psychology, linguistics, and computer science into a cognitive science framework. This is also mirrored in the diversity of the methods that are applied, chosen with regard to their fit to the research question of the respective study. Although the diversity helps with adjusting a specific study to, for example, the micro- or macro-level of human interaction, and with gaining or relaxing experimental control when necessary, there is one common limitation that the different methods used in the thesis share: None of the studies featured face-to-face interaction between participants, instead opting to let them interact only virtually. Human interactions are adapted to this face-to-face situation (Clark, 1996; Schegloff, 2006), and thus a case could be made that to recreate true interactions as faithfully as possible, participants should also be allowed to interact face-to-face in experimental studies. Although examples of communication outside of this setting are abundant in human cultures (for instance, writing or traffic signs), they are arguably low in productivity or rely mostly on linguistic encoding (Morin et al., 2020). It is important to note that the three studies strived to preserve as many features of face-to-face interaction as possible. In particular, live interaction between the participants, referring to activity in the same time frame, was enabled in the three studies (although not that common in Chapter 4). Furthermore, repair mechanisms were explicitly introduced into the paradigms presented in Chapters 2 and 4 (although rarely used in the latter), and the framework in Chapter 3 was unrestricted with regard to how participants interacted outside of the canvas. Yet this leaves participants still short of several options available to them in “true” face-to-face interaction, in particular the ability to communicate multi-modally and to build on an immediate shared environment.

One reason to resort to virtual tasks without face-to-face interaction is simply practical: Studies in the scale of the one presented in Chapter 3 would be impossible to conduct face-to-face, and an enormous effort in the case of Chapter 4. Yet at least for artificial language games, face-to-face

interactions can be allowed for, and examples of such approaches already exist (e.g. Garrod et al., 2007; Nölle et al., 2018). Here, another crucial factor to resort to virtual tasks nevertheless is that they are easier to control. Allowing for the full gamut of face-to-face situations to be made use of means that information will be carried by other means than the intended communication device (in Chapters 2 and 4, the black-and-white symbols). This information is then hard to quantify; in fact, the main point of Chapter 2 was to keep shared information other than in the manipulation limited to precisely estimate the effect of this provided information and as little else as possible. Ultimately, the more important ecologically valid part of the studies presented in the thesis is that they do involve real interactions between human participants, whether these happen virtually or face-to-face. It is this basic interaction between humans that has been the cornerstone underlying the concepts discussed and researched throughout the thesis, theoretically and methodologically.

Taken together, the three studies paint a picture of the emergence and organization of communicative signals being driven by human interaction. This picture is consistent with and logically follows from some of the main concepts outlined in the introduction. Building their ostensive communicative signals on the shared context, interlocutors are able to establish novel conventions without prior conventional codes (Chapter 2). These conventions emerge from individual interactions, and their semantic structure is reflected to some degree in both natural and artificial communication (Chapter 4). By repeated production and transmission, cultural traits get organized at the macro-evolutionary level due to different constraints, exemplified by the outcome of compression (Chapter 3). Overall then, this thesis presented an account of how the individual interactions are the baseline behaviors driving both the emergence and organization of communicative signals. As a result, an entire communication system can emerge and evolve.

## 6. References

- Adams, K. L., & Winter, A. (1997). Gang graffiti as a discourse genre. *Journal of Sociolinguistics*, 1(3), 337–360. <https://doi.org/10.1111/1467-9481.00020>
- Altmann, E. G., Pierrehumbert, J. B., & Motter, A. E. (2011). Niche as a determinant of word fate in online groups. *PLoS ONE*, 6(5), e19009. <https://doi.org/10.1371/journal.pone.0019009>
- Archer, D. (1997). Unspoken diversity: Cultural differences in gestures. *Qualitative Sociology*, 20(1), 79–105. <https://doi.org/10.1023/A:1024716331692>
- Atlas, J. D. (2005). *Logic, meaning, and conversation: Semantical underdeterminacy, implicature, and their interface*. Oxford University Press.
- Barnett, H. G. (1953). *Innovation: The basis of cultural change*. McGraw-Hill.
- Baronchelli, A. (2018). The emergence of consensus: A primer. *Royal Society Open Science*, 5(2), 172189. <https://doi.org/10.1098/rsos.172189>
- 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>
- Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a “theory of mind.” *Cognition*, 21(1), 37–46. [https://doi.org/10.1016/0010-0277\(85\)90022-8](https://doi.org/10.1016/0010-0277(85)90022-8)
- Barr, D. J., & Keysar, B. (2002). Anchoring comprehension in linguistic precedents. *Journal of Memory and Language*, 46(2), 391–418. <https://doi.org/10.1006/jmla.2001.2815>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>

- Bartlett, F. C. (1932). *Remembering: An experimental and social study*. Cambridge University Press.
- Barton, N. H., Briggs, D. E. G., Eisen, J. A., Goldstein, D. B., & Patel, N. H. (2007). *Evolution*. Cold Spring Harbor Laboratory Press.
- 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 University Press.
- Berwick, R. C., Pietroski, P., Yankama, B., & Chomsky, N. (2011). Poverty of the stimulus revisited. *Cognitive Science*, 35(7), 1207–1242.  
<https://doi.org/10.1111/j.1551-6709.2011.01189.x>
- Blount, Z. D., Borland, C. Z., & Lenski, R. E. (2008). Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 105(23), 7899–7906.  
<https://doi.org/10.1073/pnas.0803151105>
- Bornstein, M. H. (1987). Perceptual categories in vision and audition. In S. A. Harnad (Ed.), *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), 61-74. <https://doi.org/10.2307/3773722>
- Boyd, R., & Richerson, P. J. (1988). *Culture and the evolutionary process*. University of Chicago Press.

- Boyd, R., & Richerson, P. J. (1996). Why culture is common, but cultural evolution is rare. In W. G. Runciman, J. M. Smith, & R. I. M. Dunbar (Eds.), *Evolution of social behaviour patterns in primates and man*. (pp. 77–93). Oxford University Press.
- Boyd, R., & Richerson, P. J. (2006). Culture and the evolution of the human social instincts. In N. J. Enfield & S. C. Levinson (Eds.), *Roots of Human Sociality: Culture, Cognition and Interaction* (pp. 453–477). Berg.
- Breheny, R. (2006). Communication and folk psychology. *Mind and Language*, 21(1), 74–107. <https://doi.org/10.1111/j.1468-0017.2006.00307.x>
- Brennan, S. E. (2005). How conversation is shaped by visual and spoken evidence. In J. C. Trueswell & M. K. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions* (pp. 95–129). MIT Press.
- Brennan, S. E., & Hanna, J. E. (2009). Partner-specific adaptation in dialog. *Topics in Cognitive Science*, 1(2), 274–291. <https://doi.org/10.1111/j.1756-8765.2009.01019.x>
- Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial Life*, 8(1), 25–54. <https://doi.org/10.1162/106454602753694756>
- Brighton, H., Kirby, S., & Smith, K. (2005). Cultural selection for learnability: Three principles underlying the view that language adapts to be learnable. In M. Tallerman (Ed.), *Language origins: Perspectives on evolution* (pp. 291–309). Oxford University Press.
- Brighton, H., Smith, K., & Kirby, S. (2005). Language as an evolutionary system. *Physics of Life Reviews*, 2(3), 177–226. <https://doi.org/10.1016/j.plrev.2005.06.001>

- 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), 1–21. <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>
- Brown-Schmidt, S. (2009). Partner-specific interpretation of maintained referential precedents during interactive dialog. *Journal of Memory and Language, 61*(2), 171–190. <https://doi.org/10.1016/j.jml.2009.04.003>
- Brown-Schmidt, S. (2012). Beyond common and privileged: Gradient representations of common ground in real-time language use. *Language and Cognitive Processes, 27*(1), 62–89. <https://doi.org/10.1080/01690965.2010.543363>
- Brown-Schmidt, S., Gunlogson, C., & Tanenhaus, M. K. (2008). Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition, 107*(3), 1122–1134. <https://doi.org/10.1016/j.cognition.2007.11.005>
- Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*(2), 261–304. <https://doi.org/10.1177/0049124104268644>
- Caldwell, C. A., & Smith, K. (2012). Cultural evolution and perpetuation of arbitrary communicative conventions in experimental microsocieties. *PLoS ONE, 7*(8), e43807. <https://doi.org/10.1371/journal.pone.0043807>
- 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>

- Cavalli-Sforza, L. L., & Feldman, M. W. (1981). *Cultural transmission and evolution: A quantitative approach*. Princeton University Press.
- Charbonneau, M. (2015). All innovations are equal, but some more than others: (Re)integrating modification processes to the origins of cumulative culture. *Biological Theory*, *10*(4), 322–335. <https://doi.org/10.1007/s13752-015-0227-x>
- Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(1), 19–22. [https://doi.org/10.1016/S1364-6613\(02\)00005-0](https://doi.org/10.1016/S1364-6613(02)00005-0)
- 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>
- Christiansen, M. H., & Kirby, S. (2003). *Language evolution*. Oxford University Press.
- Claidiere, N., Smith, K., Kirby, S., & Fagot, J. (2014). Cultural evolution of systematically structured behaviour in a non-human primate. *Proceedings of the Royal Society of London B: Biological Sciences*, *281*(1797), 20141541. <https://doi.org/10.1098/rspb.2014.1541>
- Clark, A. (2015). Radical predictive processing. *The Southern Journal of Philosophy*, *53*, 3–27. <https://doi.org/10.1111/sjp.12120>
- Clark, H. H. (1996). *Using language*. Cambridge University Press.
- Clark, H. H., & Carlson, T. B. (1981). Context for comprehension. In J. Long & A. Baddeley (Eds.), *Attention and performance IX* (pp. 313–330). Lawrence Erlbaum.



- Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, *50*(1), 62–81.  
<https://doi.org/10.1016/j.jml.2003.08.004>
- Clark, H. H., & Schaefer, E. F. (1987). Collaborating on contributions to conversations. *Language and Cognitive Processes*, *2*(1), 19–41.  
<https://doi.org/10.1080/01690968708406350>
- Clark, H. H., Schreuder, R., & Buttrick, S. (1983). Common ground at the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior*, *22*(2), 245–258. [https://doi.org/10.1016/S0022-5371\(83\)90189-5](https://doi.org/10.1016/S0022-5371(83)90189-5)
- Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39. [https://doi.org/10.1016/0010-0277\(86\)90010-7](https://doi.org/10.1016/0010-0277(86)90010-7)
- Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. Wiley-Interscience.
- Craycraft, N. N., & Brown-Schmidt, S. (2018). Compensating for an inattentive audience. *Cognitive Science*, *42*(5), 1504–1528. <https://doi.org/10.1111/cogs.12614>
- Culbertson, J., & Kirby, S. (2016). Simplicity and specificity in language: Domain-general biases have domain-specific effects. *Frontiers in Psychology*, *6*, 1964.  
<https://doi.org/10.3389/fpsyg.2015.01964>
- Dabbour, L. M. (2012). Geometric proportions: The underlying structure of design process for Islamic geometric patterns. *Frontiers of Architectural Research*, *1*(4), 380–391.  
<https://doi.org/10.1016/j.foar.2012.08.005>
- Dale, R., Fusaroli, R., Duran, N. D., & Richardson, D. C. (2013). The self-organization of human interaction. *Psychology of Learning and Motivation*, *59*, 43–95.  
<https://doi.org/10.1016/B978-0-12-407187-2.00002-2>

- Darlington, P. J. (1972). Competition, competitive repulsion, and coexistence. *Proceedings of the National Academy of Sciences*, 69(11), 3151–3155.  
<https://doi.org/10.1073/pnas.69.11.3151>
- 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>
- Dawkins, R. (1979). *The selfish gene*. Oxford University Press.
- Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., Gisladdottir, R. S., Kendrick, K. H., Levinson, S. C., Manrique, E., Rossi, G., & Enfield, N. J. (2015). Universal principles in the repair of communication problems. *PLOS ONE*, 10(9), e0136100. <https://doi.org/10.1371/journal.pone.0136100>
- Dingemanse, M., Torreira, F., & Enfield, N. J. (2013). Is “Huh?” a universal word? Conversational infrastructure and the convergent evolution of linguistic items. *PLOS ONE*, 8(11), e78273. <https://doi.org/10.1371/journal.pone.0078273>
- Dufau, S., Duñabeitia, J. A., Moret-Tatay, C., McGonigal, A., Peeters, D., Alario, F.-X., Balota, D. A., Brysbaert, M., Carreiras, M., Ferrand, L., Ktori, M., Perea, M., Rastle, K., Sasburg, O., Yap, M. J., Ziegler, J. C., & Grainger, J. (2011). Smart phone, smart science: How the use of smartphones can revolutionize research in cognitive science. *PLoS ONE*, 6(9), e24974. <https://doi.org/10.1371/journal.pone.0024974>
- Dunbar, R. I. M. (1998). *Grooming, gossip, and the evolution of language*. Harvard University Press.

- El Mouden, C., André, J.-B., Morin, O., & Nettle, D. (2014). Cultural transmission and the evolution of human behaviour: A general approach based on the Price equation. *Journal of Evolutionary Biology*, 27(2), 231–241. <https://doi.org/10.1111/jeb.12296>
- Enfield, N. J. (2014). *Natural causes of language: Frames, biases, and cultural transmission*. Language Science Press. [https://doi.org/10.26530/OAPEN\\_533873](https://doi.org/10.26530/OAPEN_533873)
- Enfield, N. J., & Levinson, S. C. (2006). Human sociality as a new interdisciplinary field. In N. J. Enfield & S. C. Levinson (Eds.), *Roots of Human Sociality: Culture, Cognition and Interaction* (pp. 1–35). Berg.
- Epstein, J. M. (1999). Agent-based computational models and generative social science. *Complexity*, 4(5), 41–60. [https://doi.org/10.1002/\(SICI\)1099-0526\(199905/06\)4:5<41::AID-CPLX9>3.0.CO;2-F](https://doi.org/10.1002/(SICI)1099-0526(199905/06)4:5<41::AID-CPLX9>3.0.CO;2-F)
- 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>
- Fay, N., Garrod, S., & Roberts, L. (2008). The fitness and functionality of culturally evolved communication systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1509), 3553–3561. <https://doi.org/10.1098/rstb.2008.0130>
- Fay, N., Garrod, S., Roberts, L., & Swoboda, N. (2010). The interactive evolution of human communication systems. *Cognitive Science*, 34(3), 351–386. <https://doi.org/10.1111/j.1551-6709.2009.01090.x>
- Ferdinand, V. A. (2015). *Inductive evolution: Cognition, culture, and regularity in language*. [Doctoral dissertation]. University of Edinburgh.

- Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084), 998. <https://doi.org/10.1126/science.1218633>
- Fusaroli, R., & Tylén. (2012). Carving language for social coordination: A dynamical approach. *Interaction Studies*, 13(1), 103–124. <https://doi.org/10.1075/is.13.1.07fus>
- Fussell, S. R., & Krauss, R. M. (1989). The effects of intended audience on message production and comprehension: Reference in a common ground framework. *Journal of Experimental Social Psychology*, 25(3), 203–219. [https://doi.org/10.1016/0022-1031\(89\)90019-X](https://doi.org/10.1016/0022-1031(89)90019-X)
- 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](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, 11. <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>
- Galati, A., & Brennan, S. E. (2010). Attenuating information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language*, 62(1), 35–51. <https://doi.org/10.1016/j.jml.2009.09.002>
- Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of representation: Where might graphical symbol systems come from? *Cognitive Science*, 31(6), 961–987. <https://doi.org/10.1080/03640210701703659>

- Garrod, S., Fay, N., Rogers, S., Walker, B., & Swoboda, N. (2010). Can iterated learning explain the emergence of graphical symbols? *Interaction Studies*, *11*(1), 33–50.  
<https://doi.org/10.1075/is.11.1.04gar>
- 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>
- Gass, S. M. (1987). The resolution of conflicts among competing systems: A bidirectional perspective. *Applied Psycholinguistics*, *8*(4), 329–350.  
<https://doi.org/10.1017/S0142716400000369>
- 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>
- Gorman, K. S., Gegg-Harrison, W., Marsh, C. R., & Tanenhaus, M. K. (2012). What's learned together stays together: Speakers' choice of referring expression reflects shared experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 843–853. <https://doi.org/10.1037/a0029467>
- Grice, P. (1989). *Studies in the way of words*. Harvard University Press.
- Griffiths, T. L., Christian, B., & Kalish, M. (2008). Using category structures to test iterated learning as a method for identifying inductive biases. *Cognitive Science*, *32*(1), 68–107.  
<https://doi.org/10.1080/03640210701801974>
- Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition*, *135*, 21–23. <https://doi.org/10.1016/j.cognition.2014.11.026>
- Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, *31*(3), 441–480. <https://doi.org/10.1080/15326900701326576>

- Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, 49(1), 43–61. [https://doi.org/10.1016/S0749-596X\(03\)00022-6](https://doi.org/10.1016/S0749-596X(03)00022-6)
- Hardin, G. (1960). The competitive exclusion principle. *Science*, 131(3409), 1292–1297. <https://doi.org/10.1126/science.131.3409.1292>
- Harnad, S. A. (1987). Psychophysical and cognitive aspects of categorical perception: A critical overview. In S. A. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 1–52). Cambridge University Press.
- Healey, P. G., Swoboda, N., Umata, I., & King, J. (2007). Graphical language games: Interactional constraints on representational form. *Cognitive Science*, 31(2), 285–309. <https://doi.org/10.1080/15326900701221363>
- Heller, D., Gorman, K. S., & Tanenhaus, M. K. (2012). To name or to describe: Shared knowledge affects referential form. *Topics in Cognitive Science*, 4(2), 290–305. <https://doi.org/10.1111/j.1756-8765.2012.01182.x>
- Heller, D., Grodner, D., & Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition*, 108(3), 831–836. <https://doi.org/10.1016/j.cognition.2008.04.008>
- Henrich, J. (2004). Demography and cultural evolution: How adaptive cultural processes can produce maladaptive losses—The Tasmanian case. *American Antiquity*, 69(02), 197–214. <https://doi.org/10.2307/4128416>
- Henrich, J., Boyd, R., & Richerson, P. J. (2008). Five misunderstandings about cultural evolution. *Human Nature*, 19(2), 119–137. <https://doi.org/10.1007/s12110-008-9037-1>
- Hockett, C. F. (1960). The origin of speech. *Scientific American*, 203(3), 88–97.

- Höfler, S. H. (2009). *Modelling the role of pragmatic plasticity in the evolution of linguistic communication* [Doctoral dissertation]. University of Edinburgh.
- Höfler, S. H., & Smith, A. D. M. (2009). The pre-linguistic basis of grammaticalisation: A unified approach to metaphor and reanalysis. *Studies in Language*, 33(4), 886–909. <https://doi.org/10.1075/sl.33.4.03hoe>
- Holler, J., & Wilkin, K. (2009). Communicating common ground: How mutually shared knowledge influences speech and gesture in a narrative task. *Language and Cognitive Processes*, 24(2), 267–289. <https://doi.org/10.1080/01690960802095545>
- 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.
- Hopper, P. J., & Traugott, E. C. (2003). *Grammaticalization*. Cambridge University Press.
- Horn, L. (2004). Implicature. In L. Horn & G. Ward (Eds.), *The Handbook of Pragmatics* (pp. 3–28). Blackwell.
- Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, 59(1), 91–117. [https://doi.org/10.1016/0010-0277\(96\)81418-1](https://doi.org/10.1016/0010-0277(96)81418-1)
- Hurford, J. R. (2002). Expression/induction models of language evolution: Dimensions and issues. In T. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models*. (pp. 301–344). Cambridge University Press.
- Hurford, J. R. (2007). *Origins of meaning*. Oxford University Press.
- Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology: General*, 116(1), 26–37. <https://doi.org/10.1037/0096-3445.116.1.26>

- Ishihara, S. (1972). *The series of plates designed as a test for colour-blindness*. Kanehara & Co., Ltd.
- 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>
- Kadir, T., & Brady, M. (2001). Saliency, scale and image description. *International Journal of Computer Vision*, 45(2), 83–105. <https://doi.org/10.1023/A:1012460413855>
- Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, 14(2), 288–294. <https://doi.org/10.3758/BF03194066>
- Kandler, A., Wilder, B., & Fortunato, L. (2017). Inferring individual-level processes from population-level patterns in cultural evolution. *Royal Society Open Science*, 4(9), 170949.  
<https://doi.org/10.1098/rsos.170949>
- 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, P., Berlin, B., Maffi, L., Merrifield, W. R., & Cook, R. (2009). *The world color survey*. CSLI Publications.
- Kay, P., & 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>



- Kellerman, E. (1995). Crosslinguistic influence: Transfer to nowhere? *Annual Review of Applied Linguistics*, 15, 125–150. <https://doi.org/10.1017/S0267190500002658>
- Kelly, P., Winters, J., Miton, H., & Morin, O. (in press). The predictable evolution of letter shapes: An emergent script of West Africa recapitulates historical change in writing systems. *Current Anthropology*, 1–38. <https://doi.org/10.31235/osf.io/eg489>
- Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, 336(6084), 1049–1054. <https://doi.org/10.1126/science.1218811>
- Kendrick, K. H., Brown, P., Dingemanse, M., Floyd, S., Gipper, S., Hayano, K., Hoey, E., Hoymann, G., Manrique, E., Rossi, G., & Levinson, S. C. (2020). Sequence organization: A universal infrastructure for social action. *Journal of Pragmatics*, 168, 119–138. <https://doi.org/10.1016/j.pragma.2020.06.009>
- Keysar, B. (1997). Unconfounding common ground. *Discourse Processes*, 24(2–3), 253–270. <https://doi.org/10.1080/01638539709545015>
- Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, 11(1), 32–38. <https://doi.org/10.1111/1467-9280.00211>
- Keysar, B., Barr, D. J., Balin, J. A., & Paek, T. S. (1998). Definite reference and mutual knowledge: Process models of common ground in comprehension. *Journal of Memory and Language*, 39(1), 1–20. <https://doi.org/10.1006/jmla.1998.2563>
- Kirby, S. (1999). *Function, selection, and innateness: The emergence of language universals*. Oxford University Press.

- Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. In T. Briscoe (Ed.), *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge University Press.
- Kirby, S. (2017). Culture and biology in the origins of linguistic structure. *Psychonomic Bulletin & Review*, 24(1), 118–137. <https://doi.org/10.3758/s13423-016-1166-7>
- 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., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, 104(12), 5241–5245. <https://doi.org/10.1073/pnas.0608222104>
- Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, 28, 108–114. <https://doi.org/10.1016/j.conb.2014.07.014>
- 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>
- Krauss, R. M., & Fussell, S. R. (1991). Perspective-taking in communication: Representations of others' knowledge in reference. *Social Cognition*, 9(1), 2–24. <https://doi.org/10.1521/soco.1991.9.1.2>

- Krauss, R. M., & Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, *1*, 113–114. <https://doi.org/10.3758/BF03342817>
- Krauss, R. M., & Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, *4*(3), 343–346. <https://doi.org/10.1037/h0023705>
- Krauss, R. M., & Weinheimer, S. (1967). Effect of referent similarity and communication mode on verbal encoding. *Journal of Verbal Learning and Verbal Behavior*, *6*(3), 359–363. [https://doi.org/10.1016/S0022-5371\(67\)80125-7](https://doi.org/10.1016/S0022-5371(67)80125-7)
- Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago press.
- Laland, K. N., & Brown, G. R. (2011). *Sense and nonsense: Evolutionary perspectives on human behaviour* (2nd ed). Oxford University Press.
- Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites* (Vol. 1). Stanford University Press.
- Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press.
- Levinson, S. C. (2006). On the human “interaction engine”. In N. J. Enfield & S. C. Levinson (Eds.), *Roots of Human Sociality: Culture, Cognition and Interaction* (pp. 39–69). Berg.
- Levinson, S. C. (2019). Interactional foundations of language: The interaction engine hypothesis. In P. Hagoort (Ed.), *Human language: From genes and brains to behavior* (pp. 189–200). MIT Press.
- Lewis, D. (1969). *Convention*. Harvard University Press.

- 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](https://doi.org/10.1016/S0749-596X(03)00021-4)
- Maynard-Smith, J., & Harper, D. (2003). *Animal signals*. Oxford University Press.
- Mesoudi, A. (2017). Pursuing Darwin’s curious parallel: Prospects for a science of cultural evolution. *Proceedings of the National Academy of Sciences*, 114(30), 7853–7860.  
<https://doi.org/10.1073/pnas.1620741114>
- Mesoudi, A., & Whiten, A. (2008). The multiple roles of cultural transmission experiments in understanding human cultural evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1509), 3489–3501. <https://doi.org/10.1098/rstb.2008.0129>
- Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, 7(3), 221–237. <https://doi.org/10.1177/1745691612441215>
- Millikan, R. G. (1984). *Language, thought, and other biological categories: New foundations for realism*. MIT Press.
- Millikan, R. G. (2005). *Language: A biological model*. Oxford University Press.
- Miton, H., & Morin, O. (2019). When iconicity stands in the way of abbreviation: No Zipfian effect for figurative signals. *PLOS ONE*, 14(8), e0220793.  
<https://doi.org/10.1371/journal.pone.0220793>

- Moreno, M., & Baggio, G. (2015). Role asymmetry and code transmission in signaling games: An experimental and computational investigation. *Cognitive Science*, 39(5), 918–943. <https://doi.org/10.1111/cogs.12191>
- Morin, O. (2015). *How traditions live and die*. Oxford University Press.
- Morin, O., Kelly, P., & Winters, J. (2020). Writing, graphic codes, and asynchronous communication. *Topics in Cognitive Science*, 12(2), 727–743. <https://doi.org/10.1111/tops.12386>
- Morin, O., & Miton, H. (2018). Detecting wholesale copying in cultural evolution. *Evolution and Human Behavior*, 39(4), 392–401. <https://doi.org/10.1016/j.evolhumbehav.2018.03.004>
- 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*, 3(2), 91–93. <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>
- Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science*, 13(4), 329–336. <https://doi.org/10.1111/j.0956-7976.2002.00460.x>
- 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>

- Origgi, G., & Sperber, D. (2000). Evolution, communication, and the proper function of language. In A. C. Peter Carruthers (Ed.), *Evolution and the Human Mind: Language, Modularity and Social Cognition* (pp. 140–169). Cambridge University Press.
- 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>
- Powell, A., Shennan, S., & Thomas, M. G. (2009). Late Pleistocene demography and the appearance of modern human behavior. *Science*, *324*(5932), 1298–1301. <https://doi.org/10.1126/science.1170165>
- Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526. <https://doi.org/10.1017/S0140525X00076512>
- R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Ravignani, A., Delgado, T., & Kirby, S. (2017). Musical evolution in the lab exhibits rhythmic universals. *Nature Human Behaviour*, *1*(1), 0007. <https://doi.org/10.1038/s41562-016-0007>
- Real, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*(3), 317–328. <https://doi.org/10.1016/j.cognition.2009.02.012>
- Regier, T., & 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>
- Rendall, D., Owren, M. J., & Ryan, M. J. (2009). What do animal signals mean? *Animal Behaviour*, *78*(2), 233–240. <https://doi.org/10.1016/j.anbehav.2009.06.007>

- Richerson, P. J., & Boyd, R. (2008). *Not by genes alone: How culture transformed human evolution*. University of Chicago Press.
- 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>
- Roberts, G., & Fedzechkina, M. (2018). Social biases modulate the loss of redundant forms in the cultural evolution of language. *Cognition*, *171*, 194–201.  
<https://doi.org/10.1016/j.cognition.2017.11.005>
- Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, *50*, 696–735.

- Scheffer, M., & van Nes, E. H. (2006). Self-organized similarity, the evolutionary emergence of groups of similar species. *Proceedings of the National Academy of Sciences*, *103*(16), 6230–6235. <https://doi.org/10.1073/pnas.0508024103>
- Schegloff, E. A. (1996). Turn organization: One intersection of grammar and interaction. In E. Ochs, E. A. Schegloff, & S. A. Thompson (Eds.), *Interaction and Grammar* (pp. 52–133). Cambridge University Press.
- Schegloff, E. A. (2006). Interaction: The infrastructure for social institutions, the natural ecological niche for language, and the arena in which culture is enacted. In N. J. Enfield & S. C. Levinson (Eds.), *Roots of Human Sociality: Culture, Cognition, and Interaction* (pp. 70–96). Berg.
- Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, *53*(2), 361–382.
- Schelling, T. C. (1960). *The strategy of conflict*. MIT Press.
- Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, *1*(2), 103–113. <https://doi.org/10.1111/j.2041-210X.2010.00012.x>
- Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, *21*(2), 211–232. [https://doi.org/10.1016/0010-0285\(89\)90008-X](https://doi.org/10.1016/0010-0285(89)90008-X)
- 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. (2017). Pragmatics and the aims of language evolution. *Psychonomic Bulletin & Review*, *24*(1), 186–189. <https://doi.org/10.3758/s13423-016-1061-2>



- 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>
- Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. University of Illinois Press.
- Shennan, S. (2002). *Genes, memes, and human history: Darwinian archaeology and cultural evolution*. Thames & Hudson.
- 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>
- Skyrms, B. (2010). *Signals: Evolution, learning, and information*. Oxford University Press.
- Smith, A. D. M. (2008). Protolanguage reconstructed. *Interaction Studies*, 9(1), 100–116. <https://doi.org/10.1075/is.9.1.08smi>
- Smith, A. D. M., & Höfler, S. H. (2015). The pivotal role of metaphor in the evolution of human language. In J. E. Díaz-Vera (Ed.), *Metaphor and Metonymy across Time and Cultures*. De Gruyter. <https://doi.org/10.1515/9783110335453.123>

- Smith, K., & Kirby, S. (2008). Cultural evolution: Implications for understanding the human language faculty and its evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1509), 3591–3603. <https://doi.org/10.1098/rstb.2008.0145>
- Sperber, D. (1996). *Explaining culture: A naturalistic approach*. Blackwell Publishers.
- Sperber, D. (2000). Metarepresentations in an evolutionary perspective. In D. Sperber (Ed.), *Metarepresentations: A multidisciplinary perspective* (pp. 117–137). Oxford University Press.
- Sperber, D. (2006). Why a deep understanding of cultural evolution is incompatible with shallow psychology. In N. J. Enfield & S. C. Levinson (Eds.), *Roots of Human Sociality: Culture, Cognition and Interaction* (pp. 431–449). Berg.
- Sperber, D., & Wilson, D. (1996). *Relevance: Communication and cognition* (2nd ed). Blackwell Publishers.
- Sperber, D., & Wilson, D. (2002). Pragmatics, modularity and mind-reading. *Mind & Language*, 17(1–2), 3–23. <https://doi.org/10.1111/1468-0017.00186>
- 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>
- Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E., & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26), 10587–10592. <https://doi.org/10.1073/pnas.0903616106>
- Sulik, J., & Lupyan, G. (2018). *Perspective taking in a novel signaling task: Effects of world knowledge and contextual constraint*. PsyArXiv. <https://doi.org/10.31234/osf.io/ftz94>

- Sullivan, M. (1989). *The meeting of Eastern and Western art*. University of California Press.
- Tamariz, M., & Kirby, S. (2015). Culture: Copying, compression, and conventionality. *Cognitive Science*, 39(1), 171–183. <https://doi.org/10.1111/cogs.12144>
- Tamariz, M., Kirby, S., & Carr, J. W. (2016). Cultural evolution across domains: Language, technology and art. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2765–2770). Cognitive Science Society.
- 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>
- Tan, R., & Fay, N. (2011). Cultural transmission in the laboratory: Agent interaction improves the intergenerational transfer of information. *Evolution and Human Behavior*, 32(6), 399–406. <https://doi.org/10.1016/j.evolhumbehav.2011.01.001>
- 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>
- Thompson, B., Kirby, S., & Smith, K. (2016). Culture shapes the evolution of cognition. *Proceedings of the National Academy of Sciences*, 113(16), 4530–4535. <https://doi.org/10.1073/pnas.1523631113>
- Tinits, P., Nölle, J., & Hartmann, S. (2017). Usage context influences the evolution of overspecification in iterated learning. *Journal of Language Evolution*, 2(2), 148–159. <https://doi.org/10.1093/jole/lzx011>
- Tomasello, M. (1999). *The cultural origins of human cognition*. Harvard university press.

- Tomasello, M. (2010). *Origins of human communication*. MIT Press.
- Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28(5), 675–691. <https://doi.org/10.1017/S0140525X05220125>
- Tylén, K., Weed, E., Wallentin, M., Roepstorff, A., & Frith, C. D. (2010). Language as a tool for interacting minds. *Mind & Language*, 25(1), 3–29. <https://doi.org/10.1111/j.1468-0017.2009.01379.x>
- Van der Loo, M. P. (2014). The stringdist package for approximate string matching. *The R Journal*, 6(1), 111–122. <https://doi.org/10.32614/RJ-2014-011>
- Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language*, 31(2), 183–194. [https://doi.org/10.1016/0749-596X\(92\)90010-U](https://doi.org/10.1016/0749-596X(92)90010-U)
- 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>
- Witzel, C., & Gegenfurtner, K. R. (2011). Is there a lateralized category effect for color? *Journal of Vision*, 11(16), 1–25. <https://doi.org/10.1167/11.12.16>
- Witzel, C., & Gegenfurtner, K. R. (2013). Categorical sensitivity to color differences. *Journal of Vision*, 13(7), 1–33. <https://doi.org/10.1167/13.7.1>

- Witzel, C. (2018). Misconceptions about colour categories. *Review of Philosophy and Psychology*, 10, 499–540. <https://doi.org/10.1007/s13164-018-0404-5>
- Woensdregt, M., & Smith, K. (2017). Pragmatics and language evolution. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.  
<https://doi.org/10.1093/acrefore/9780199384655.013.321>
- 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>
- Wu, S., & Keysar, B. (2007). The effect of information overlap on communication effectiveness. *Cognitive Science*, 31(1), 169–181. <https://doi.org/10.1080/03640210709336989>
- 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. <https://doi.org/10.1098/rspb.2012.3073>
- Yoon, S. O., Benjamin, A. S., & Brown-Schmidt, S. (2016). The historical context in conversation: Lexical differentiation and memory for the discourse history. *Cognition*, 154, 102–117. <https://doi.org/10.1016/j.cognition.2016.05.011>
- 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>
- Zenil, H., Hernández-Orozco, S., Kiani, N. A., Soler-Toscano, F., & Rueda-Toicen, A. (2018). A decomposition method for global evaluation of Shannon entropy and local estimations of algorithmic complexity. *ArXiv*. <http://arxiv.org/abs/1609.00110>

- 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>
- Zipf, G. K. (1949). *Human behaviour and the principle of least-effort*. Addison-Wesley Press.
- Zobl, H. (1980). The formal and developmental selectivity of L1 influence on L2 acquisition. *Language Learning*, *30*(1), 43–57. <https://doi.org/10.1111/j.1467-1770.1980.tb00150.x>
- 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>

## **Ehrenwörtliche Erklärung**

Ich erkläre hiermit, dass mir die geltende Promotionsordnung der Fakultät für Sozial- und Verhaltenswissenschaften der Friedrich-Schiller-Universität Jena bekannt ist. Ferner erkläre ich, dass ich die vorliegende Dissertation selbstständig ohne die Hilfe Dritter angefertigt habe, sowie alle benutzten Quellen und Hilfsmittel in der Arbeit angegeben habe. Insbesondere habe ich keine Hilfe eines Promotionsberaters in Anspruch genommen. Bei der Auswahl und Auswertung des Materials sowie der Herstellung der Einzelmanuskripte haben mich die angegebenen Koautoren unentgeltlich unterstützt. Darüber hinaus hat kein Dritter unmittelbar oder mittelbar geldwerte Leistungen von mir für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen. Ich erkläre weiterhin, dass ich diese Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung, oder eine gleiche, eine in wesentlichen Teilen ähnliche oder eine andere Abhandlung bei einer anderen Hochschule bzw. anderen Fakultät als Dissertation eingereicht habe. Ich versichere, nach bestem Wissen die reine Wahrheit gesagt und nichts verschwiegen zu haben.

Datum, Ort

Unterschrift

## Publikationen

- Henschel, M., Winters, J., Müller, T. F., & Bräuer, J. (2020). Effect of shared information and owner behavior on showing in dogs (*Canis familiaris*). *Animal cognition*, 23(5), 1019-1034. <https://doi.org/10.1007/s10071-020-01409-9>
- 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>
- Müller, T. F., & Winters, J. (2018). Compression in cultural evolution: Homogeneity and structure in the emergence and evolution of a large-scale online collaborative art project. *PloS one*, 13(9), e0202019. <https://doi.org/10.1371/journal.pone.0202019>
- 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*, 3(2), 91-93. <https://doi.org/10.1093/jole/lzy005>