Collaborative meaning-making: the emergence of novel languages in humans, machines, and human-machine interactions

Kouwenhoven, T.

Collaborative Meaning-Making

The Emergence of Novel Languages in Humans, Machines, and Human-Machine Interactions

Tom Kouwenhoven

Collaborative Meaning-Making

The Emergence of Novel Languages in Humans, Machines, and Human-Machine Interactions

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op donderdag 30 oktober 2025
klokke 13:00 uur

door Tom Kouwenhoven

geboren te Nijmegen

in 1995

Promotor:
Prof.dr. S.A. Raaijmakers

Co-promotores
Dr. T. Verhoef
Dr.ir. R.E. de Kleijn

Promotiecommissie:

Prof.dr. M.M. Bonsangue	
Prof.dr. K.J. Batenburg	
Prof.dr. G. Lupyan	(University of Wisconsin–Madison)
Prof.dr. R. Verbrugge	(Rijksuniversiteit Groningen)
Dr. L.P.A. Galke	(University of Southern Denmark)
Dr. F.H. Poletiek	

Keywords: Language Evolution, Human-Machine collaboration, Grounding, Large Language Models
Cover by: OpenAI and Alice Mulder.

SIKS Dissertation Series No. 2025-47
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Preface

"Without Practice, No Emergence

'To know' and 'to understand' are different.

Even though we know, without putting that knowledge to practice, we cannot understand."

—Daihonzan Eiheiji temple, Japan 2022

One of the benefits of being a scholar is that you get to travel to all kinds of places around the world. It was during my first-ever conference on language evolution that I had the pleasure of visiting the Daihonzan Eiheiji monastery in Japan. A truly unique and memorable experience which is difficult to put into words really. You must have been there to understand the feeling. It was here that this one excerpt out of the 14 teachings exposed in the corridor caught my eye. These teachings were related to the beliefs important in that monastery, but for me, these words encompass almost perfectly what language evolution is, and so I take the liberty of sharing my interpretation before diving into the scientific part.

What it means for me, in light of this thesis, is that interactions (i.e., practice) cannot be underestimated when it comes to the making of new words (i.e., emergence) and understanding language. Languages obtain meaning when we put them into practice; when we interact and use our knowledge of our language to understand the meanings of each other.

Contents

1 Introduction

This dissertation aims to deepen our understanding of how inductive biases shape the emergence of structured languages across human, machine, and human-machine interactions. It combines experimental and computational approaches to study how processes of language learning and language use are exposed differently in various scenarios. The experiments comprising this dissertation originate from well-known setups in psycholinguistics and are complemented with contemporary artificially intelligent models of language. This enhances our understanding of inductive human and machine biases while promoting the development of natural human-machine interactions, ultimately contributing insights to cognitive science and artificial intelligence research.

The introduction is largely based on our peer-reviewed journal article: Kouwenhoven, T., Verhoef, T., de Kleijn, R.E., Raaijmakers, S.A. (2022). Emerging Grounded Shared Vocabularies between Human and Machine, inspired by Human Language Evolution. In Frontiers in Artificial Intelligence, section: Language and Computation. Volume: 5:886349. doi: 10.3389/frai.2022.886349. Sections 1.1.1, 1.1.3, and 1.1.4 have been updated to account for recent findings.

1

1.1 Background

Our ability to communicate is remarkable. It allows us to collaborate efficiently in large groups, exchange ideas, and build upon knowledge previously acquired by others (Tomasello, 1999). To communicate successfully, the coordinated actions of all participants must adhere to the grounding criterion: that interlocutors agree they have understood what was meant for the current purposes (Clark and Brennan, 1991). In other words, we need a shared language, a vocabulary of mappings between signals, which can be sounds, words, gestures, and so on, and their corresponding meanings. However, it is not at all trivial that we primarily communicate by means of combining words in structured ways to create meaningful sentences. How do signals obtain their specific meaning? And what makes us interpret signals as bearers of communicative intent in the first place?

Questions like these are still relevant today, even though scholars have debated about them for decades. Some argue for the existence of an innate biological component in a language faculty that is shared by all humans (Chomsky, 1965). This entails that acquiring a language is guided by the innate constraints of this faculty. Importantly, it also means that languages can only be acquired when they adhere to a set of grammar rules, thus limiting the number of possible human languages. However, this is in stark contrast to the incredible diversity of languages that can be observed in the world, which exhibit radically different lexical, morphological, and phonological properties (Evans and Levinson, 2009). Moreover, there is considerable evidence suggesting that languages, and their evolution, adapt to social, ecological, and technological factors, indicating that languages adapt to the environments in which they are used (Lupyan and Dale, 2016). Another position that takes such factors into account relies not on a universal biological component, but on the social character of humans. This school argues that language systems evolved as a result of cultural evolution, where behaviours or ideas are learned through social interactions (e.g. Kirby and Christiansen, 2003; Hurford, 2007; Christiansen and Chater, 2008; Tomasello, 2008). These interactions facilitate a moment to negotiate what signals refer to which meanings. In other words, they offer a moment of grounding. According to this line of thought, learning a language is a collaborative process that imposes pressures, such as cognitive or expressive, which play a role during repeated learning and using languages and thereby slowly shape what languages look like (Smith, 2022). This dissertation is situated in this last school, where it is believed that language evolves as a result of cooperative interactions between interlocutors with the goal of

mutual understanding and collaboration.

Although progress has been made over the past decades, research into the evolution of language is still ongoing, and a general consensus on how languages have evolved is still far from present. One complex problem is that spoken languages obviously do not fossilise, i.e., they do not leave traceable records. This necessitates that researchers look beyond written scripts and must instead draw on evidence from many different sources, such as animal communication, sign languages, archaeological evidence, experiments in artificial language learning, and computational simulations (Christiansen and Kirby, 2003). We present work situated in the latter two. In experimental studies of language evolution, participants engage in communication games that often involve creating, learning and producing artificial miniature languages (Kirby et al., 2008; Scott-Phillips et al., 2009; Galantucci, 2005; Verhoef, 2012; Perlman et al., 2015; Raviv et al., 2019a, inter alia). These aim to gain insights into the dynamics affecting how languages evolve by carefully designing experiments that involve learning an artificial language and interacting with it. While useful, participants are often mature language users who have already been exposed to languages, perhaps obfuscating what conclusions can be drawn from these insights. This is why computational simulations play an important role in the endeavour of unravelling the evolution of language. They provide complete control and allow careful investigations into what and how building blocks, biases, or interactions play a role in successfully establishing communicative systems (Steels, 1999; Quinn, 2001; de Boer, 2006; Kirby, 2017, inter alia). In addition, computational simulations are ideal candidates to simulate longer timespans, lending themselves perfectly to mimic evolutionary processes. Such simulations were initially agent-based simulations that viewed language as adaptive dynamic systems where complex solutions could emerge at the population level from simple individual behaviours (e.g. De Boer, 2000; Steels, 2012b). This makes them excellent for simulating large groups of interacting agents, potentially demonstrating the emergence of an apparent design without having an explicit designer, similar to what we observe in bird flocking behaviour. Despite understanding the mechanisms driving these behaviours in simulations, they remain simplified models that cannot fully encompass the rich complexity of human behaviour. Hence, it could be argued that an interdisciplinary approach combining the strengths of both computational models and experiments with real humans is a fruitful direction.

More complex models of language, like Large Language Models (LLMs), emerged as promising tools for studying language acquisition that can enable controlled experiments which model human learning processes (Warstadt, 2022; Contreras Kallens

1

et al., 2023). As such, we now shift positions and briefly discuss these contemporary deep neural networks with a Transformer architecture (Vaswani et al., 2017). LLMs as novel types of Artificial Intelligence (AI) models of human language use are trained with masked language modelling and next-word prediction objectives on increasingly large quantities of (internet) data. As such, they rely on the idea that being exposed to enough textual data models will suddenly result in the capacity to understand language and produce fluent speech, a phenomenon known as emergent behaviour (Wei et al., 2022; Schaeffer et al., 2023). Since their inception, it is difficult to imagine a week without the release of a new model or algorithm; however, arguably the most popular ones are known as GPT-4 (OpenAI, 2024), Gemini (Gemini Team, 2024), and Llama3 (Llama Team, 2024). Even though LLMs are fundamentally different from humans and learn languages primarily through exposure to text, their internal representations effectively simulate cognitive language processing with factors such as data size, model scaling, and alignment training positively relating to fMRI signals of the brain (Ren et al., 2025).

At first glance, it may seem like there is a large gap between the evolution of language and LLMs. These models are, after all, trained on modern natural language that has already evolved into its present form that we use every day. However, while typically seen as a niche field, insights from the field of language evolution are increasingly relevant for computational linguists. For example, methods from language evolution and psycholinguistics can be used to steer the development of LLMs that are more human-like (Zheng et al., 2024; Galke and Raviv, 2025) and can be used to compare (biases in) LLMs directly to humans (Jones et al., 2024). Some even go as far as to argue that the ability of modern LLMs to model language refutes Chomsky's approach to language (Piantadosi, 2024; Kallini et al., 2024). While the representation of meaning in LLMs is not entirely understood, it is argued that they represent the idea of meaning-through-use and capture languages as a culturally evolving, adaptive system that is shaped by learning and communication (Contreras Kallens and Christiansen, 2024). As such, principles that steer and shape languages to become human-like in experiments or simulations have moreover become relevant to developers of LLMs. A prime example is provided by Galke and Raviv (2025), who draw parallels between a well-known pressure for communicative success in emergent communication (Kirby et al., 2015) and the final training stage in reinforcement learning from human feedback (i.e., RLHF). A pressure to be understood seems necessary for the emergence of structure in experiments with humans, and similarly so for computational simulations, where communicative success is encoded in the optimisation objective of the neural

networks (e.g. Lian et al., 2023a). While the linguistic capacity and knowledge of LLMs originate mainly from pre-training (Zhou et al., 2023; Lin et al., 2024), only after language models are fine-tuned to be understood through the process of RLHF, do they become more representative of human communicative behaviours (see Galke and Raviv (2025) for a more complete list of examples).

Vice versa, (large) language models as relatively weakly biased language learners (Wilcox et al., 2023), can also be informative of language acquisition in general (Warstadt and Bowman, 2022; Contreras Kallens et al., 2023; van Dijk et al., 2024). Crucially, we do not claim that artificial LLMs are equivalent to the language mechanisms in the human brain—they are inherently different. Rather, we view them as entirely new forms of understanding that introduce new kinds of problem-solving capabilities that may not be human-like (Mitchell and Krakauer, 2023). This requires evaluation without anthropocentric biases, i.e., without dismissing mechanistic strategies of LLMs or vision-and-language models that differ from those present in humans. Put differently, the way LLMs or other AI models solve a cognitive task cannot serve as evidence against particular cognitive competences or language understanding, as long as the solution generalises (Millière and Rathkopf, 2024). In this regard, we take LLMs as examples that establish a lower bound on what linguistic phenomena *in principle* can be learned from distributional information (van Dijk et al., 2023a). In any case, contemporary language models are interesting models of language that can be used to answer cognitive and typological questions (Warstadt and Bowman, 2022; van Dijk et al., 2023a; Binz et al., 2025) and complement explanations of human cognition resulting from Bayesian modelling (Griffiths et al., 2024). This work presented in this dissertation can be interpreted as an example of how language modelling and psycholinguistic research can complement each other.

The primary focus of this dissertation is on the presence of inductive biases in humans and those present in artificially intelligent systems. We are particularly interested in implicit mechanistic inductive biases that may result in biased language learning, not in behavioural biases observed in humans (e.g., the confirmation bias). This is relevant in the context of language evolution as seemingly arbitrary aspects of linguistic structure may actually result from general learning and processing biases deriving from the structure of thought processes, perceptuo-motor factors, cognitive limitations, and pragmatics (Christiansen and Chater, 2008). At a population level, these biases may *manifest* themselves as preferences for compressibility, simplicity, and efficiency-cognitive tendencies (Kirby et al., 2015; Tamariz and Kirby, 2015; Gibson et al., 2019) that naturally influence language evolution. For example, in the case of

human systems (e.g., language) that are culturally transmitted, a memory constraint can enforce systems to be easy to learn and simple, as hard-to-learn elements are less likely to be transmitted. Furthermore, the sound systems of human languages seem to be optimised for criteria such as acoustic distinctiveness or articulatory ease (Liljencrants and Lindblom, 1972; Lindblom and Maddieson, 1988) through a process of self-organisation (De Boer, 2000). Human constraints like these could well have evolved differently and are inherently different between humans and LLMs. In LLMs, inductive biases are increasingly well understood (Futrell and Mahowald, 2025) and emerge from the Transformer architecture Vaswani et al. (2017), including preferences for simplicity, structural organisation, positional sensitivity, and verbosity (Rende et al., 2024; Chen et al., 2024; Kallini et al., 2024; Liu et al., 2024; Mina et al., 2025; Zheng et al., 2023; Saito et al., 2023). While the underlying mechanisms differ between humans and machines, these inductive biases may produce overlapping behavioural effects since they emerge from the properties of language systems, such as being culturally transmitted and used for successful communication. The behavioural effects thereby provide insights into language acquisition, processing, and development in both natural and artificial systems. We address inductive biases, such as the ones mentioned before, in humans and artificially intelligent systems through emergent communication paradigms. Doing so helps us understand in what respects humans and AI models differ and potentially allows us to alleviate these differences through the process of collaborative meaning-making between humans and machines.

In the remainder of this introduction, we will situate this dissertation between the field of language evolution and computational linguistics, and argue for an interdisciplinary hybrid approach in which humans and artificially intelligent systems collaboratively shape languages (Section 1.1.1). We will then provide more background on the role of interactions during the emergence of meaningful signals in Section 1.1.2. Thereafter, in Section 1.1.3, we lay out prominent processes that influence how languages become structured. We then, given the prominence of language models in our everyday lives, discuss our view on collaborative human-machine language evolution and set the stage for the experiments presented in this dissertation (Section 1.1.4). With this information, we move on to the research questions in Section 1.2.1, methods in Section 1.2.2, and the dissertation outline in Section 1.2.3.

1.1.1 The evolution of Human-AI languages

Building conversational AI systems aims to teach machines to understand human language and respond naturally. The most common way to train language models to produce and interpret natural language is currently by exposing them to large quantities of data, in which models are tasked with infilling masked words given a context (i.e., the cloze task) or predicting the word that follows a given context. These models are hereafter fine-tuned to respond to instructions and align with human preferences (Ouyang et al., 2022). Although this has resulted in advances in many areas, there is fierce debate about the degree to which these systems have an understanding of how language is related to the real world (Mordatch and Abbeel, 2018; Bender et al., 2021; Mitchell and Krakauer, 2023; van Dijk et al., 2023a; Mollo and Millière, 2023), known as the *symbol grounding problem* (Harnad, 1990). It is important to mention that the concept of grounding is heavily conflated, ironically bearing many different meanings. In this dissertation, we refer to grounding as both referential and communicative grounding, as laid out by Mollo and Millière (2023). Here linguistic signals are anchored to a reference in the world (i.e., referentially grounded) or can be seen as a form of coordinated action (i.e., communicative grounding) that involves collaborating to reach a common understanding of what is said (Clark and Brennan, 1991). In addition to the debate on grounding, language models are primarily trained in isolation, while humans are social animals, deeply embedded in culture and surrounded by others. This socio-cultural perspective balances aspects of innovation and imitation, for which Yiu et al. (2024) draw parallels between children and LLM abilities to imitate and innovate, and argue that innovation requires more than large-scale language and image data alone. Complex human behaviours, like language, evolved in socio-cultural contexts and could not exist without a variety of minds using and transmitting these behaviours.

These socio-cultural contexts and mechanisms that influence the emergence of communication and linguistic structure have been studied in the field of language evolution. Although the precise origins of human language are still widely debated, computer simulations (de Boer, 2006; Steels, 2012a; Kirby, 2017) and experiments in which humans use novel communication signals (Scott-Phillips and Kirby, 2010; Galantucci and Garrod, 2010; Kirby et al., 2014), have revealed some key mechanisms that drive the initial emergence of a novel language and the gradual appearance of more complex linguistic structure. Here, we review some of these mechanisms and propose to apply methods that confirm the importance of including micro-societies of

1

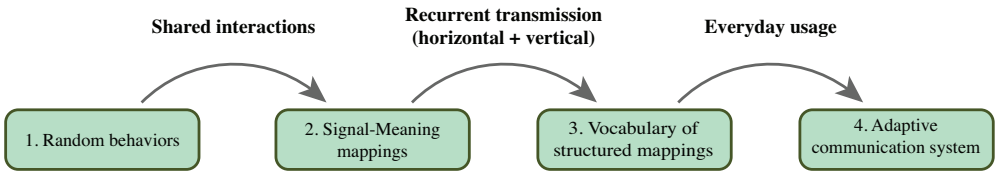


Figure 1.1: The various steps of evolving natural communication systems. First, initially random behaviours obtain meanings and become more structured through recurrent horizontal and vertical transmission. Everyday usage facilitates the continuous evolution of communication systems, which will adapt as a result of the inductive biases of humans and machines.

interacting minds in the emergence of novel human-machine communication systems.

A major insight from these studies is that language *adapts* to *human* inductive biases that influence how it is learned and used (Kirby et al., 2014, 2015; Smith, 2022). Current *language models* based on the Transformer architecture also exhibit inductive biases affecting their language learning abilities (Futrell and Mahowald, 2025). For instance, synthetic free-order case-marking languages are more challenging to model than fixed-order languages (Bisazza et al., 2021). Languages lacking hierarchical structure or having unnatural or irreversible word orders, categorised as 'impossible' by Chomsky (2023), are also more difficult to learn for GPT-2 models than 'possible' languages (Kallini et al., 2024). Provided that languages adapt to their users and that both humans *and* LLMs display inductive biases that play a role in language learning, we suggest that language used in human-machine communication should also evolve more naturally. Concretely, this entails giving a more prominent role to the co-development of shared vocabularies by conversational partners (human or AI-based). This facilitates general processes of language learning and use, which shape languages, which in turn may result in a dynamic grounded communication system that is natural and adapted to inductive biases and constraints of human *and* machine learning. The following sections describe ways to establish such grounded communicative systems and correspond to different chapters in this dissertation. Figure 1.1 shows how the various ways are related to each other. Starting from random behaviours, a signal-meaning mapping emerges from shared interactions (Section 1.1.2, Chapter 2, and Chapter 3) which become more structured through horizontal and vertical transmission (Section 1.1.3, Chapter 5, and Chapter 6) and eventually evolve into an adaptive communication system (Section 1.1.4 and Chapter 7).

1.1.2 Interactive meaning making

Successful communication requires that communicative acts adhere to the grounding criterion: that interlocutors mutually agree on what was meant for the current purposes (Clark and Brennan, 1991). This requires a vocabulary that is (partially) aligned between interlocutors of a conversation (Pickering and Garrod, 2004). The emergence of such a vocabulary starts with agreeing on what kind of (initially random) behaviours should be interpreted as communicative and what they refer to (box 1 & 2 in Figure 1.1).

Experiments with human participants have been conducted to study the emergence of novel communication forms and shared vocabularies (Galantucci, 2005; Steels, 2006; Scott-Phillips et al., 2009; Galantucci and Garrod, 2010). Here, participants need to invent and negotiate novel signals to solve a communicative or cooperative task. Albeit often bound to the starting conditions of the experiment, even when no conventional signalling device is given, actions may gradually become communicative (Scott-Phillips et al., 2009). Typically, humans quickly establish conventions and settle on a shared set of signals. The existence of sufficient common ground, interactions, and social coordination have been identified as crucial to facilitating the emergence of communication systems. With computational agents, Quinn (2001) investigated the emergence of signals and cooperation without dedicated communication channels in a way comparable to the work of Scott-Phillips et al. (2009). Here, robots, equipped only with sensors to observe a shared environment, were tasked with moving away from a starting point while maintaining proximity to each other. Initial random behaviours gradually evolved into an iconic signalling system that could establish the allocation of leader-follower roles (Quinn, 2001; Quinn et al., 2003).

A large body of work in *evolutionary language games*, as reviewed in Steels (2012b), has shown that agents without a pre-programmed language can develop a communication system from scratch. This happens in a self-organising fashion, as alignment between agents arises from repeated interactions between individuals without the existence of a central point of control. In the context of those experiments, Steels already proposed that robots can participate in the ongoing evolution of language and learn from human language users if there are sufficiently rich situated interactions (Steels, 2012a). The former is arguably already the case: scientific English is, for example, changing due to the presence and use of LLMs, with words like ‘delve‘, ‘underscore‘, and ‘intricate‘ appearing increasingly often in publications (Juzek and Ward, 2025). Although building an initially shared vocabulary is well-explored between humans

as well as between agent-based models, to the best of our knowledge, it is rarely applied in human-machine settings. One exception is a large-scale exhibition of Steels' Talking Heads experiment (Steels, 1999), in which both agents and human visitors proposed new words that could become part of an evolving shared vocabulary. We propose revisiting this idea in the context of conversational AI, allowing the process of self-organisation to facilitate the grounding of conventional signal-meaning mappings.

Our proposition should not be seen as a replacement for pre-training language models on data alone, but rather that incorporating interactions that require communicative intent may be a fruitful direction to induce more natural language learning in LLMs. This is more relevant than before, given the recent advancements in LLMs, where interactional aspects of language learning are often overlooked (Beuls and Van Eecke, 2024). To this end, we similarly argue that the role of interactions should be more prominent when developing natural communication between humans and machines. Practically, this pertains to the fundamental question of extending the current training paradigms of LLMs beyond the current practices of pre-training and fine-tuning LLMs. It requires determining how to integrate the meaningful, intentional, situated, communicative, and interactional aspects of human linguistic communication into the training process (Beuls and Van Eecke, 2024). Chapter 2 addresses how such interactional aspects can result in newly formed shared conventions in the case of humans, and Chapter 3 takes an initial step towards modelling this with deep neural networks.

1.1.3 Emergence of structure in language systems

Human language is uniquely structured and exhibits systematicity at multiple levels (Kirby, 2017). For example, words are combined into sentences such that their meaning is a function of the meanings of the parts and the way they are combined, i.e., our language has a compositional structure. The origins of this and other types of structure have been studied using computer models and artificial language learning experiments with humans (for a review see: Kirby, 2017).

Among others, two important processes have been found to contribute to the emergence of structure in languages (boxes 2 & 3 in Figure 1.1). The first is known as *cumulative cultural evolution* where (cultural) information, such as ideas or linguistic signals, is transmitted vertically along generations of users. The seminal experiment by Kirby et al. (2008) investigated *vertical transmission* in an experimental setup known as iterated learning. In their experiment, the first participant was asked to learn

an artificial non-structured language and describe stimuli with the acquired words. Subsequent participants learned the output of the previous participant. Through this process, imitating generations of language learners, the words gradually changed and became more compositional and learnable. Such results consistently show that increases in learnability and structure arise because languages adapt to human inductive biases to be transmitted faithfully (Griffiths and Kalish, 2007a). Words and patterns that are not easily learned or interpreted will not be reproduced by the next generation. Since structured languages are more easily compressible (Tamariz and Kirby, 2015; Kirby et al., 2015), this eventually results in more learnable and structured languages.

The second process contributing to the emergence of structure in human language is known as *horizontal transmission*. Here, linguistic structure originates and evolves from social coordination through repeated interactions between individuals in micro-societies. While interactions between dyads can lead to shared vocabularies and initial regularities (Verhoef et al., 2016b; Theisen-White et al., 2011), a community of users seems to be necessary for the emergence of system-wide compositional structure and efficient coding (Fay et al., 2008; Raviv et al., 2019a). An underlying dynamic was recently proposed by Josserand et al. (2024), who demonstrated that repeated dyadic interactions cause languages to evolve in a way that accommodates the specific abilities and preferences of minority individuals at the group level. In these cases, pressures such as the abilities of your interaction partner, the number of interaction partners, and expanding meaning spaces cause initially random languages to become more structured over time.

The effects of horizontal and vertical transmission have also been demonstrated with agent-based computer simulations (Kirby, 2017; Steels and Loetzsch, 2012). Altogether, there is strong evidence suggesting that the transmission of signals (vertical or horizontal) within communities contributes to the emergence of structure in language. In fact, it has been argued that both types of transmission are essential to get a language that is learnable and usable (Kirby et al., 2015). In this dissertation, Chapter 5 shows how horizontal transmission can be applied in simulations with RL agents. Chapter 6 demonstrates how horizontal and vertical transmission affect the learnability of artificial languages in LLMs. We argue that both processes should be projected onto the human-machine language evolution scenario to evolve a vocabulary that shares features with human language and is equally adapted to be learned and used by machines (as shown in Chapter 7).

1

1.1.4 Human-machine evolution and reinforcement learning

Inspired by general mechanisms of language learning and use in humans, the field of computational linguistics started to train machines to understand human language through the emergence of communication systems (e.g. Lazaridou et al., 2017, 2018; Mordatch and Abbeel, 2018; Clark et al., 2019; Manning et al., 2020). A range of work has shown that (multi-agent) Reinforcement Learning (RL; Sutton and Barto, 2018) can converge on communication protocols in various scenarios that require communication (e.g. Lazaridou et al., 2016; Havrylov and Titov, 2017; Chaabouni et al., 2020; Lian et al., 2023a; Ben Zion et al., 2024). Given the resemblance in experimental design, scholars began comparing the resulting computational protocols with those found in human experiments. However, the findings in such models initially did not always match what is typically found in similar experiments with humans, and features found in human language often did not emerge (e.g. Chaabouni et al., 2019a; Lazaridou and Baroni, 2020; Rita et al., 2022b; Galke et al., 2022). Despite having fundamentally different mechanisms, we now know that initially absent linguistic properties can be resolved by artificially inducing human-like biases (Galke and Raviv, 2024) or making the simulations more naturalistic (Lian et al., 2023a). Two give some examples, endowing these agents with a need to be understood (i.e. communicative success), noise, context sensitivity, and incremental sentence processing help induce human-like patterns such as a word-order/case-marking trade-off or dependency length minimisation in RL agents Lian et al. (2023a, 2024); Zhang et al. (2024b). The emergence of anti-efficient languages (i.e. languages that do not follow Zipf's law) found by Chaabouni et al. (2019a) can be mitigated by introducing biases for speaker 'laziness' and 'impatient' listeners (Rita et al., 2020).

Even though communicative systems can emerge that are also human-like, these often suffer from interpretability issues for humans (Mordatch and Abbeel, 2018; Li et al., 2024), making their applicability to human-machine communication less obvious. To this end, Lazaridou et al. (2020) endowed RL agents with a pre-trained language model and used self-play to teach these RL agents to communicate in natural language. Without human intervention, however, this approach suffers from what is commonly referred to as language drift, ultimately causing the initially aligned vocabularies to diverge from human vocabularies and leading to misunderstandings. A similar point is shown by Shumailov et al. (2024), who trained different contemporary types of neural learners on recursively generated data. They found that the data quickly lost relation to the original input and drifted away to accommodate the inductive

preferences of these models. While too much is problematic, we argue that some (language) drift is welcome since it allows data (or languages) to be optimised for entity-specific preferences. When this happens in a collaborative manner, i.e., when the data is optimised for humans *and* machines, it can result in more natural human-machine communication. As such, human feedback should be incorporated directly into the behaviour of an artificially intelligent language user, rather than learning it in isolation. This draws parallels to human interactions, which offer a means to ground communicative signals through recurrent and reciprocal usage (Garrod et al., 2007), provide feedback on the success of a conversational contribution, and alleviate miscommunications resulting from partially aligned vocabularies due to variations or dialects.

In light of contemporary, data-hungry LLMs, the picture is a bit more nuanced as most linguistic knowledge can be obtained during pre-training (Zhou et al., 2023; Lin et al., 2024) while human feedback, e.g., through RLHF, must ensure that the otherwise unwieldy models align with intended human behaviours (Ouyang et al., 2022).[1] Yet, employing RLHF alone is not the same as having collaborative interactions, as it only considers the adaptation of a single entity, rather than both. On this note, Beuls and Van Eecke (2024) argued that modelling the situated, communicative, and interactional environments in which human languages are acquired provides a promising path to overcome the limitations of current LLMs that essentially rely on the distributional hypothesis. A much more collaborative approach that is increasingly often employed is using language games often used in language evolution research. By doing so, the training regime simulates a more natural interactive (vision-)language learning approach and oftentimes results in increased performance on linguistic benchmarks (Lowe et al., 2020; Steinert-Threlkeld et al., 2022; Zheng et al., 2024; Shumailov et al., 2024). As such, instead of learning signal-meaning mappings in a bottom-up approach, we argue that general language learning and use processes important in language evolution should be applied in a top-down manner. We hereby follow Bisk et al. (2020) in that shared experiences make utterances meaningful and that successful linguistic communication relies on a shared experience. This is especially important given the increasing appearance of LLMs in everyday life and their growing influence on human culture (Brinkmann et al., 2023; Yiu et al., 2024).

To unravel the processes involved in creating mutually understood communicative

[1] Recent findings showed that the similarity between LLM representations and brain cognitive language processing fMRI signals increased as a result of RLHF (Ren et al., 2025). Although this is not straightforwardly relevant in the context of human-machine collaborations, it implies that fine-tuning can align representations between humans and machines, which can aid collaboration.

1

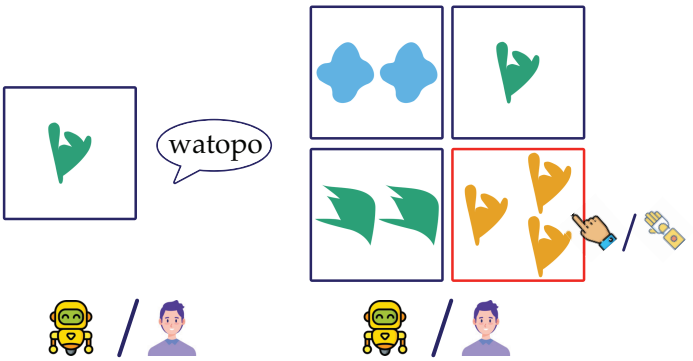


Figure 1.2: An exemplar setup in which humans and machines collaborate in a referential game as used in Chapter 6 and 7. In these games, a speaker (human or machine) utters a non-existing word which the listener uses to guess the target. Repeated interactions offer a means to establish initial conventions and extrapolate simple (grammatical) rules that enable successful interactions. Icons obtained from flaticon.com

systems between humans and machines, we propose to revisit popular methods in language evolution research such as signalling games (Galantucci, 2005; Scott-Phillips et al., 2009), referential games (Steels and Loetzsch, 2012; Chaabouni et al., 2020), and navigation games (Mordatch and Abbeel, 2018; Dubova and Moskvichev, 2020). This enables collaborative interactions between humans and machines, offering a means to ground languages in shared experiences (Figure 1.2). Importantly, the evolved communication systems will not take the same form as human language initially, but through iterations, may come closer towards it and evolve into a form that makes human-machine interactions more natural, with communication systems adapted to biases in both human and machine learning (box 4 in Figure 1.1). Doing so will contribute to our understanding of human and machine intelligence, but at the same time may reveal important mismatches between the two types of learners and thereby inform modelling decisions (Futrell and Mahowald, 2025; Galke and Raviv, 2025). In this dissertation, Chapter 7 is an example of such work.

In summary, we propose to combine insights from psycholinguistics and the field of human language evolution, particularly concerning the influence of vertical and horizontal transmission, with contemporary language models. In doing so, this dissertation will contribute in two ways. First, we address how signals and structure emerge in socio-cultural contexts and discuss how language adapts to the way it is learned and used. We therefore suggest that language used in human-machine communication should also evolve naturally, emphasising the importance of co-development of

shared conventions during communication. To this end, this dissertation addresses several collaborative meaning-making experiments encompassing human, machine, and human-machine pairs. Second, we propose that methods from psycholinguistics can be used to unravel and inform mechanisms in artificial language models. To this end, this dissertation includes empirical studies that originate in psycholinguistics that are adapted with the aim of understanding inductive biases in machines. We believe that this combined approach enhances our understanding of both human and machine cognition while potentially revealing crucial differences between humans and machines that can guide future modelling decisions and enhance human-machine interactions.

1.2 Dissertation design

1.2.1 Research Questions

Now that the background of language evolution and language modelling is in place, we can formulate the following main research question (MRQ):

Main research question

How can human and artificial cognition in emergent communication complement each other?

The main research question inherently asks for an interdisciplinary approach. To this end, this dissertation can be roughly divided into three perspectives: a purely human-oriented approach, a purely computational approach, and finally a hybrid approach. Besides studying different entities, the chapters also differ in terms of the nature of interaction. Some require active cooperation, while other chapters do not involve interactions but rather study whether computational models display human-like biases or study human behaviour through behaviour cloning. Finally, this dissertation is structured such that the linguistic complexity, as a variable of interest, increases throughout the dissertation. Starting with the evolution of elementary core concepts in the process of language evolution, and moving towards experiments that investigate more structured natural language-like expressions.

The first approach (Chapters 2 and 3) focuses on unravelling collaborative human behaviour at a very elementary stage in language evolution where no conventional

communicative medium exists. It comprises an interactive experiment with human participants and training AI models on the behaviour of humans, i.e., behaviour cloning through training deep neural networks. The second approach (Chapters 4 through 7) focuses on computational preferences and behaviours that are believed to influence language learning in a slightly later stage of language evolution, where initial communicative signals are already in place. The computational tools employed comprise vision-and-language models to assess whether novel words are mapped to visual features, and we extract visual features for RL agents who utilise these when learning to communicate. Additionally, we employ a LLM as a novel computational model of language and cognition in an interactive referential game. In Chapter 7, we bring together most of the previous chapters and incorporate key aspects and insights from all previous chapters. In this final chapter, humans and LLMs collaborate so as to shape and develop their own artificial language in a communicative game.

We break down the MRQ into six research questions corresponding to six self-contained chapters. These were originally published as research papers at various peer-reviewed conferences and workshops spanning cognitive science, computational linguistics and AI-oriented research fields. We included the papers in this dissertation mostly as-is, apart from minor edits for consistent terminology use, formatting, and additional clarifications and information. Therefore, each chapter can be read and understood independently, though this approach means some introductory sections contain repeated information across chapters. We ask for the reader's understanding regarding this redundancy. In the remainder of this section, we introduce and motivate each research question.

Research question 1 - Chapter 2

What role does diversity in biases for structure play in developing symbolic communicative systems?

Successful communication requires interlocutors to agree on the meaning of a message (Clark and Brennan, 1991), i.e., they must agree on the meaning of individual signals (semantics) but also on how these signals are composed together (syntax). While this seems very obvious nowadays because we have all sorts of communicative conventions, from an evolutionary point of view, this is not trivial at all (Deacon, 1997). To agree on the meaning of a message, it must first be clear that a message has a specific communicative purpose. That is, one's actions must be understood as having the

intent of communicating something. Only after something is recognised as having a communicative intent, can we negotiate what the intended meaning of this 'message' is supposed to be (Oliphant, 2002). This process can be rendered as a cooperation problem in which individuals must find a common ground that serves as a point of departure for more elaborate signals. The first exploration of the MRQ addresses this problem by observing how this process unfolds precisely. To this end, we employ the Embodied Communication Game (ECG; Scott-Phillips et al., 2009) in which participants must communicate without the existence of an a priori communication method or medium. This means that participants must converge on a shared system of reference through repeated interaction. We use individuals' Personal Need for Structure (PNS; Neuberg and Newsom, 1993) as a measure of human bias for structure since the social coordination of a shared language, which is initially unstructured, can be influenced by an individual's need for structure. Specifically, we investigate whether diversity between participants' PNS influences the process of cooperation. Offering nuance to what is oftentimes argued: that shared experience and overlapping biases may help such processes of cooperation (Tylén et al., 2013; Scott-Phillips and Kirby, 2010).

Research question 2 - Chapter 3

What insights about human sequential processing can be derived from modelling human behaviour in emergent communication?

Modelling human behaviour using computational methods can complement theories about human cognition, but at the same time, it can inform more natural learning mechanisms in machines (Galke and Raviv, 2025). This chapter aims to address both, making it the first interdisciplinary question of this dissertation. First, we set out to instil human communicative behaviour in deep neural networks using the data obtained in the first chapter. Following the methods of previous work (de Kleijn et al., 2018), we perform behaviour cloning to find out whether deep neural networks can learn the sequential behaviours humans exhibit while playing the ECG. We manipulate the processing directionality of our LSTM networks and *approximate* latent cognitive variables, which we relate to PNS metrics. In doing so, we provide results that resonate with the belief that there is a bidirectional sequential processing mechanism in humans and that humans use uncertainty-directed exploration strategies. Second, our findings offer insights into the types of neural networks commonly used in simulations with

1

RL agents.

Research question 3 - Chapter 4

To what extent do vision-and-language models exhibit human-like cross-modal associations such as the bouba-kiki effect?

This chapter marks the point where we transition from relatively simple models to more contemporary large (vision-and-) language models. While LLMs are often criticised for failing to connect linguistic concepts to meanings in the world, i.e, facing the 'symbol grounding problem' (Harnad, 1990), multi-modal vision-and-language models (VLMs) offer a possible solution to this challenge. However, disentangling the inner workings of such models is not at all trivial. Especially so because multi-modal representations are created using various techniques of different complexity levels. Common attempts to understand how VLMs process multi-modal input is through benchmarking them against human performance on specific tasks (e.g. Thrush et al., 2022; Diwan et al., 2022; Kamath et al., 2023), or by using experimental techniques originally designed to probe humans (e.g. Jones et al., 2024). In this chapter, we embrace the latter and focus on one of the most well-known cross-modal associations between speech sounds and visual shapes: the bouba-kiki effect (Ramachandran and Hubbard, 2001). While the link between signals and meanings in language may seem arbitrary, these associations are in fact not arbitrary at all (e.g. Perlman et al., 2015; Davis et al., 2019). A wealth of work was done on why words look, sound, and feel the way they do when they refer to certain meanings (e.g. Perniss et al., 2010; Winter et al., 2017; Dingemanse et al., 2015; Verhoef et al., 2016a; Cuskley and Kirby, 2013). We investigate whether increasingly popular VLMs also display human-like cross-modal preferences by adapting an experimental setup (Nielsen and Rendall, 2013) and probing four contemporary VLMs. Our findings offer nuance to recent claims that VLMs show strong cross-modal associations.

Research question 4 - Chapter 5

What role does representational alignment play in the emergence of compositional language in reinforcement learning?

Representational alignment concerns the degree of agreement between internal representations of two processing systems (Sucholutsky et al., 2023). In simulations with

deep reinforcement learning agents, agents typically project signals and meanings onto agent-specific representations that represent their understanding of the inputs. At its core, this essentially renders the cooperative task of communication to be one where both agents need to align their respective image representations (Rita et al., 2022b). This is similar to what humans do, who perhaps also do not experience perceptions equally (Locke, 1847), but at the same time introduces some problems. Whereas humans use repair mechanisms which can alleviate differing perceptions and meanings to maintain successful cooperation (Garrod et al., 2007), RL agents learn through single-step interactions and typically do not have such repair mechanisms and must rely on redundant explicit messaging (Vital et al., 2025). Furthermore, if we wish machines to have a more natural understanding of human language, they should develop vocabularies that are referentially grounded in concept-level properties that are shared by humans. Preferably, this happens in such a way that the individual characters of signals refer to concepts in the input and such that they are composed in a structured manner (i.e., in a compositional way). In this chapter, we first confirm earlier work revealing that representational alignment hinders the emergence of conceptually grounded languages. To further explore how these artificial languages are affected by representational alignment, we propose an additional loss function and directly test whether RL agents can communicate on a strict compositionality benchmark (Thrush et al., 2022).

Research question 5 - Chapter 6

To what extent can Large Language Models learn and use artificial languages in emergent communication, mirroring human patterns of language evolution?

An important finding in the field of language evolution is that individual learning biases and pressures present during language learning and use continuously shape languages (Smith and Culbertson, 2020). Many iterated learning studies that have contributed to this belief involve a process called cumulative cultural evolution in which information is transmitted across generations (e.g. Kirby et al., 2008, 2015; Verhoef et al., 2015; Arnon and Kirby, 2024). An important aspect of iterated learning experiments is that the information transmitted via iterated learning will ultimately come to mirror the minds of the learners (Griffiths and Kalish, 2007a). In this chapter, we extend earlier work presented by Galke and Raviv (2024) and use LLMs as language learners since scholars are increasingly interested in testing LLMs as if they were

subjects with cognitive abilities (Binz and Schulz, 2023; Pellert et al., 2024; Binz and Schulz, 2024; Löhn et al., 2024). We subject an LLM to a referential game—an artificial language learning task commonly used in language evolution experiments with humans. These experiments involve learning an artificial language and then using it during a communicative task, which shapes initially holistic unstructured languages such that they can be reliably used during communication. Additionally, we simulate cumulative cultural evolution by creating transmission chains of language learners and users who learn from the output of previous generations. In doing so, we shed light on the question of whether LLMs can be used in artificial language learning experiments and show that, just like how this happens in humans, the underlying mechanistic inductive biases in these models influence the resulting languages.

Research question 6 - Chapter 7

Can humans and Large Language Models develop shared vocabularies through collaborative communication?

Instead of only learning from humans, scholars sometimes argue that current technologies are actively *contributing* to our culture (e.g. Yiu et al., 2024; Brinkmann et al., 2023). The incorporation of technologies in everyday life also increasingly demands maintaining alignment between humans and machines (Beuls and Van Eecke, 2024) and arguably requires referential grounding. One way to do so is to incorporate repeated interaction between humans and machines (Bisk et al., 2020; Shumailov et al., 2024; Beuls and Van Eecke, 2024) and optimise for communicative success (Smith et al., 2024). The integration of a pressure for communicative success in the training procedure of machines has successfully promoted more natural (vision-)language learning through the referential game (Lowe et al., 2020; Steinert-Threlkeld et al., 2022; Zheng et al., 2024). However, this game also offers a fitting test bed for cooperative alignment between humans and machines. Hence, we test whether general processes of language learning and use can result in referentially grounded languages that are mutually understood and expressive when humans *and* LLMs, each with their own set of inductive biases, engage in the referential game. In particular, our work extends the previous chapter by comparing the languages evolved for LLMs only with those resulting from Human-Human and Human-LLM interactions. In doing so, we additionally provide suggestions that promote more natural language learning in current training processes of LLMs.

Method	RQ1	RQ2	RQ3	RQ4	RQ5	RQ6
Human experiments	•	•				•
Computational experiments		•		•	•	•
Statistical modelling	•	•	•	•		•
NLP			•	•	•	•
Qualitative analysis	•				•	•
Questionnaires	•	•				•
Chapter	2	3	4	5	6	7

Table 1.1: Overview of the main methods employed for each RQ/chapter.

1.2.2 Methods

The evolution of language is studied in many ways, among which various methods inspire our combinations of computational and experimental methods. They are described below.

Lab experiments in which humans play *communication games* are used to test what processes are important during the creation of signals (e.g. Galantucci, 2005; Scott-Phillips et al., 2009) or how these signals become gradually more structured (e.g. Kirby et al., 2008; Verhoef et al., 2014; Raviv et al., 2019a). *Computational simulations* may complement such theories by showing that processes of self-organisation can result in elaborate behaviours (e.g. De Boer, 2000; Quinn, 2001; Verhoef et al., 2011) and are used to test hypotheses and findings in human experiments (e.g. Kirby et al., 2015; de Kleijn et al., 2022; Lian et al., 2023a, 2024). Finally, in AI research, it is common practice to *benchmark* model competencies against human performance on cognitive tasks (e.g. Thrush et al., 2022; Kamath et al., 2023; van Duijn et al., 2023).

In this dissertation, we adopt the methods mentioned above to address the research questions. For example, we test LLMs in referential games, comparing languages optimised for LLM biases (Chapter 6) against those developed when humans participate in artificial language learning experiments (Chapter 7). An overview of the methods employed per chapter is given in Table 1.1. To guide the reader, we elaborate on these methods in the remainder of this section.

- **Human experiments –** Refers to conducting controlled language evolution experiments carried out in the lab in which participants engage in artificial language learning games (RQ1, RQ2, and RQ6).

- **Computational experiments –** Refers to the involvement of artificially intelligent

models in emergent communication setups through training them on human data (RQ2), via reinforcement learning (RQ4), or via in-context learning (RQ5, RQ6).

- **Statistical modelling –** Refers to statistical models such as (Bayesian) linear models used for testing hypotheses regarding PNS (RQ1), human latent variables (RQ2), cross-modal associations (RQ3), and collaborative language evolution (RQ6).
- **NLP –** Refers to the extraction of information from images or texts using pre-trained computational tools, for probing cross-modal effects (RQ3), further use in RL simulations (RQ4), or evolving signal-meaning mappings (RQ5, RQ6).
- **Qualitative analysis –** Refers to the process of manually going through the behaviours and languages that evolve during experiments. This allowed the discovery of behaviours associated with establishing sequential signals (RQ1) and the structure patterns in languages (RQ5, RQ6).
- **Questionnaires –** Refers to administering the Personal Need for Structure questionnaire (RQ1, RQ2) or to inquiring participants communicative strategies (RQ1,RQ2, and RQ6).

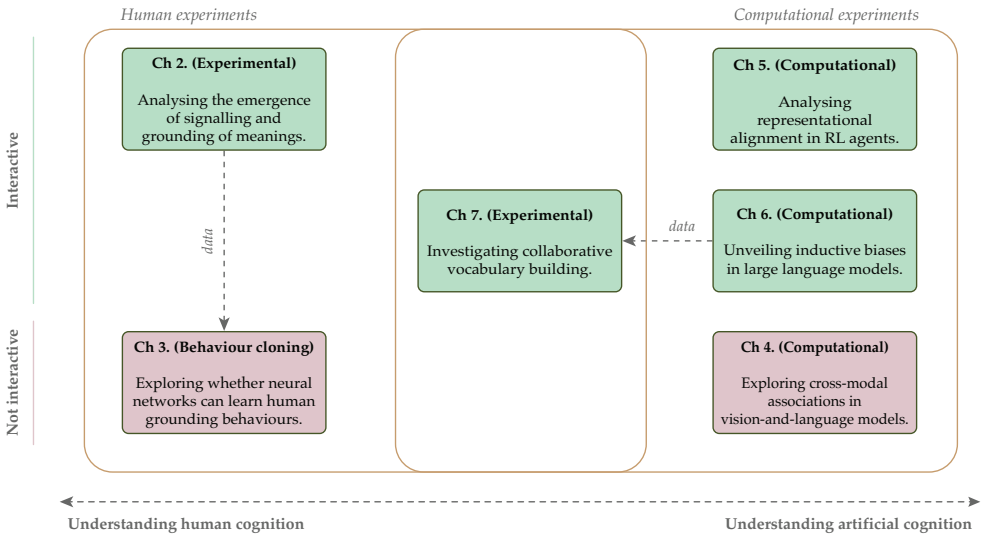


Figure 1.3: The structure of this dissertation as constituted by six chapters and their themes. Block colours indicate whether the Chapter incorporated interactions (green) or constituted a non-interactive approach (pink).

1.2.3 Outline

This section is intended as a brief guide for the reader explaining how this dissertation is organised, which is best read alongside the structure laid out in Figure 1.3.

Chapter 2 – Grounding and the Need for Structure

This interactive *experimental* chapter is arguably situated at the very early stages of language evolution with very elementary 'linguistic' communicative interactions. It introduces the notion of grounding, our questionnaire data, and serves as training data for Chapter 3.

Published as: Kouwenhoven, T., de Kleijn, R.E., Raaijmakers, S.A., Verhoef, T.(2023). Need for Structure and the Emergence of Communication. In J. Culbertson, A. Perfors., H. Rabagliati. & V. Ramenzoni., editors, *Proceedings of the Annual Meeting of the Cognitive Science Society*, Volume 44, pages 549-555. Toronto, Canada. Cognitive Science Society.

1

Chapter 3 – Computationally Modelling Human Emergent Communication

This non-interactive *behaviour cloning* chapter is the first interdisciplinary chapter of this dissertation. It employs deep learning tools to unravel the rudimentary sequential human behaviours that resulted in grounded vocabularies in the previous chapter.

> **Published as:** Kouwenhoven, T., Verhoef, T., Raaijmakers, S.A., de Kleijn, R.E. (2023). Modelling Human Sequential Behavior with Deep Learning Neural Networks in Emergent Communication. In M. Goldwater., F. K. Anggoro., B. K. Hayes., & D. C. Ong., editors, *Proceedings of the Annual Meeting of the Cognitive Science Society*, Volume 44, pages 549-555. Sydney, Australia. Cognitive Science Society.

Chapter 4 – Kiki or Bouba?

This *computational* chapter draws on theories regarding cross-modal associations in humans. These associations are still pre-linguistic in nature but compared to Chapter 2 and Chapter 3, their role in language evolution is important after communicative mediums are established. This non-interactive chapter compares associations between humans and VLMs by probing the latter directly, enabling fine-grained analyses that bring nuance to claims made in other work.

> **Published as:** Tessa Verhoef*, Kiana Shahrasbi, and Tom Kouwenhoven*. 2024. What does Kiki look like? Cross-modal associations between speech sounds and visual shapes in vision-and-language models. In Kuribayashi, T., Rambelli, G., Takmaz, E., Wicke, P., Oseki, Y., editors, *In Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 199–213, Bangkok, Thailand. Association for Computational Linguistics. (*denotes equal contribution.)

Chapter 5 – The Curious Case of Representational Alignment

This *computational* chapter moves towards more linguistically complex evaluations as opposed to the earlier chapters. It adopts a Reinforcement Learning setup that enables training deep neural networks to interactively develop languages. We assess how they fare on a strict computational benchmark proven difficult for contemporary VLMs.

Published as: Tom Kouwenhoven, Max Peeperkorn, Bram van Dijk, and Tessa Verhoef. 2024. The Curious Case of Representational Alignment: Unravelling Visio-Linguistic Tasks in Emergent Communication. In Kuribayashi, T., Rambelli, G., Takmaz, E., Wicke, P., Oseki, Y., editors, *In Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 57–71, Bangkok, Thailand. Association for Computational Linguistics.

Chapter 6 – Searching for Structure

This *computational* chapter assesses the role of inductive biases in a contemporary LLM on emergent languages. Hence, this chapter shifts the perspective to using LLMs as psychological subjects. Drawing on methods from the field of language evolution, we employ LLM-augmented agents in interactive referential games known to reveal inductive biases. We assess compositional language use and their ability to generalise to novel inputs. Finally, the data generated here are complemented in Chapter 7.

Published as: Tom Kouwenhoven, Max Peeperkorn, Tessa Verhoef. 2025. Searching for Structure: Investigating Emergent Communication with Large Language Models. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Di Eugenio, B., Schockaert. S., editors, *In Proceedings of the 31st International Conference on Computational Linguistics*, pages 9977–9991, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chapter 7 – Shaping Shared Languages

This *experimental* chapter encompasses a hybrid experiment combining most aspects important to this dissertation. It explores our proposition that language used in human-machine communication should emphasise the importance of co-development of shared conventions. To this end, we assess languages optimised for human, LLM, and Human-LLM pairs that interactively shape languages in the communication game developed in the previous chapter. Enabling direct comparisons between the languages, and investigating what languages look like when they are optimised for humans and machines.

1

Published as: Kouwenhoven, T., Peeperkorn, M., de Kleijn, R.E. and Verhoef, T. (2025). Shaping Shared Languages: Human and Large Language Models' Inductive Biases in Emergent Communication. In Kwok, J., editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI-25, International Joint Conferences on Artificial Intelligence Organization. Human-Centred AI

Chapter 8 – Conclusions

The *conclusion* chapter presents answers to all research questions and provides discussions regarding the limitations of this dissertation and directions for future research.

2

Grounding and the Need for Structure

Language is a unique hallmark of humans, it is both learned and symbolic, which poses the problem of emergence: if neither form nor meaning is known, how can individuals communicate in the first place? The current study replicates work that investigates the emergence of signal forms and meanings and explores how Personal Need for Structure (PNS) of interacting partners can aid or hinder the emergence of communicative systems. We use an existing measure of personal need for structure to investigate its relationship with the emergence of such systems while participants play the embodied communication game (ECG). Similar to the original study, our work shows that a bootstrapping process and sufficient common ground are integral to the recognition of signalhood. Moreover, this process appears to be more successful for individuals who respond differently to a lack of structure compared to their interaction partners. Contrary to what is usually assumed, our results indicate that not only shared expectations and biases seem to matter in communicative tasks, but that diversity in biases of communication partners can also be beneficial for the emergence of new communication systems.

Originally published as: Kouwenhoven, T., de Kleijn, R.E., Raaijmakers, S.A., Verhoef, T.(2023). Need for Structure and the Emergence of Communication. In J. Culbertson, A. Perfors., H. Rabagliati. & V. Ramenzoni., editors, *Proceedings of the Annual Meeting of the Cognitive Science Society*, Volume 44, pages 549-555. Cognitive Science Society.

2.1 Introduction

Humans can share and accumulate knowledge through language, enabling them to pass this knowledge on to future generations. Communication through language can be formulated as the joint action that emerges when speakers and listeners perform actions in coordination (Clark, 1996), and uses signals that are both symbolic and learned. The emergence of signals is therefore a defining event in human cognitive evolution. However, the exact dynamics of language emergence—the settling of two individuals on an effective interchange through discrete, grounded symbols—is complex and not yet fully understood (Tylén et al., 2013; Scott-Phillips and Kirby, 2010). If form and meaning are unknown, one fundamental question concerns the cooperative process of agreeing on what form should refer to what meaning (Oliphant, 2002). This process has been studied quite extensively through laboratory experiments in which participants need to invent and negotiate novel signals to solve a communicative or cooperative task (Steels, 2006; Scott-Phillips and Kirby, 2010; Tylén et al., 2013). A general finding from such studies is that participants are able, through social coordination, to establish conventions and gradually develop a communication system. Consistently, researchers report on the importance of common ground and the reliance on shared biases and expectations between interacting partners on the road to success. However, building an entirely novel system of signals from scratch is not easy, and in such experiments, it is often the case that not all pairs manage to solve the game. Analyses tend to focus on the conventions established in successful games, which have generated many insights, but we propose that a focus on differences in coordination outcomes and properties of the individuals involved can help to understand these dynamics better. In this chapter, we show how sometimes diversity rather than alignment of initial cognitive biases and preferences of individuals might positively influence success in the social coordination of a shared language.

In essence, the emergence of signals can be formulated as a cooperation problem, where individuals have a common goal and need to figure out how to influence each other in an *initially unstructured* environment. It has been proposed that the emergence of language is influenced by human biases to prefer compressible, simple systems (Kemp and Regier, 2012; Kirby et al., 2015). Such a bias can, for example, drive the emergence of systematic structure over generations of transmissions (Kirby et al., 2015). Individuals have been found to differ in their *personal need for structure* (Neuberg and Newsom, 1993) which can affect problem-solving capabilities such as solving maths problems (Svecova and Pavlovicova, 2016) and learning a foreign language or text comprehension (Eva et al., 2014). As such, the social coordination of a shared language, which is initially unstructured, can potentially also be influenced by an individual's personal need for structure. We expect that PNS might also affect how individuals act in language emergence tasks, and investigate how a personal need for structure affects the evolution of a communication system that is created de novo.

Specifically, the experiment presented in this chapter was designed to study the relationship

between personal need for structure as measured by the PNS questionnaire (Neuberg and Newsom, 1993), its F1 and F2 sub-factors and the emergence of a communicative system while playing the Embodied Communication Game (Scott-Phillips et al., 2009), which is described in detail in the next section.

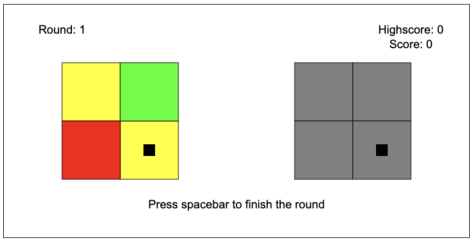
2.2 Background

The current study is based on an experiment designed by Scott-Phillips et al. (2009), who investigated the emergence of newly created communication systems when humans are not able to communicate verbally, or in any other conventional way. Participants played the ECG, a cooperative game, and the results revealed how signals acquire informative meaning without pre-defining a communication channel, roles of signaler and receiver, or a form space.

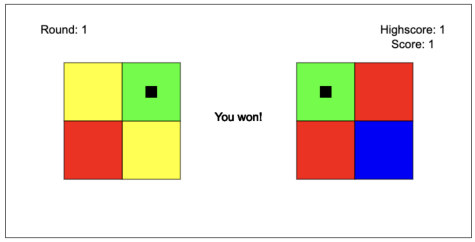
2.2.1 The Embodied Communication Game (ECG)

The ECG is a cooperative two-player game that consists of two 2×2 grid worlds, where players are embodied in the sense that they are given a physical form (a black square) to move around with. Each quadrant has one of four colours (red, green, yellow, blue), which is determined at random. The goal of the participants is to end on identically-coloured quadrants and, if they do, score a point. Players can move within their own grid and see movements in both grids, but can only see the colours of their own quadrants, showing the others' quadrants as grey (Figure 2.1a). Once finished moving, the colours of all quadrants are revealed to both players (Figure 2.1b) as a means of feedback. The colours of the quadrants and starting positions of both players are randomly chosen with the proviso that there is always one overlapping colour between both worlds, such that it is always possible to score a point. Players are informed that their goal is to score as many consecutive points as possible, meaning that players cannot win by playing many games but must instead find a way to communicate reliably and coordinate behaviours with each other (see Scott-Phillips et al. (2009) for a more elaborate explanation).

The setup of this experiment required participants to coordinate their behaviours by agreeing on what behaviours correspond to what meaning, and they had to find a way to signal that these behaviours were of communicative intent. Crucially, this problem can be solved when movements between the quadrants eventually come to be understood as communicative. It turned out to be a non-trivial task since only 7 out of 12 pairs managed to co-opt one's movements for the purpose of communication. Scott-Phillips et al. (2009) conclude that the problem of mapping form onto meaning is solved by finding sufficient common ground and bootstrapping new meanings upon that. As such, the authors suggest that the latter significantly increases the likelihood that a symbolic communication system emerges and that the emergence of dialogue is a crucial step in the development of a system that can be employed to achieve shared goals.



(a) The view while participants are playing.



(b) The view after both players ended the round.

Figure 2.1: The game environment, figure (a) shows the view while players are moving, where movements from both, but only the colours from the participants' own world are visible. Figure (b) shows the environment after both players are done with their movements. The colours of all quadrants are revealed to both players as a means of feedback.

2.2.2 Successful interactions and shared expectations

Many studies have involved the experimental emergence of artificial languages, where participants are not permitted to use conventional language systems (e.g. Steels, 2006; Scott-Phillips and Kirby, 2010; Tylén et al., 2013). A task that is somewhat related to the ECG was studied in an experiment by Galantucci (2005). Here participants played a collaborative computer game and were required to develop new semiotic conventions, which map signals and meanings, to communicate information regarding their location using a novel communicative channel. Similar to the findings of Scott-Phillips et al. (2009), not all pairs succeeded in this task. Moreover, pairs who did succeed differed widely in the manner and rate at which they managed to solve the game. Success in such tasks is typically attributed to feedback, alignment, shared biases, and similarities between pairs; however, a specific focus on the underlying mechanisms that allow some pairs to converge on a system while others can not achieve this is lacking. We are interested in precisely these dynamics and investigate how the diversity of preferences and biases in pairs influences collaborative tasks.

2.2.3 Personal need for structure

Individual differences in the desire for structure may influence how people understand and interact with their worlds. This desire can be measured by means of the Personal Need for Structure Scale (Neuberg and Newsom, 1993). It consists of 12 statements (e.g. "I enjoy having a clear and structured mode of life") that are answered on a 6-point Likert scale, which measure the tendency to seek structure in chaotic environments. It is characterised by a representation of simplified information and generalisation of previous experience into fewer complex categories that an individual uses in new and ambiguous situations (Svecova and Pavlovicova, 2016). Two conceptually different sub-factors are identified: the desire for structure in unstructured

environments (*F1*) and an individual's response to the lack of structure (*F2*).

2.3 Current study

As mentioned before, reports on cooperative games and the emergence of communication often emphasise the importance of common ground and the reliance on shared biases and expectations between interacting partners. However, we expect that differences can also play a role as interacting partners that differ might complement each other's shortcomings, which possibly aids cooperation. Arguably, the initial states of the ECG can be considered as an unstructured environment and thus may evoke different responses in humans that differ in PNS. We investigate precisely how PNS might affect the evolution of a communication system that is created de novo.

2.4 Methods

Participants ($N = 40$: 31 females, 9 males; $M_{\text{age}} = 22.12$, $SD_{\text{age}} = 3.56$) were recruited via two methods: the participant recruitment website from the Psychology department of Leiden University, and by the experimenters during lectures or other events. As a result, 20 pairs played the ECG. Upon arrival, they were given instructions about the experimental procedure and seated behind a computer in two separate rooms. The entire experiment took place on two connected computers via a web application. Participants then read instructions explaining the goal of the game, its mechanics, and were allowed to ask clarifying questions solely concerning the mechanics. This setup ensured that no conventional communication was possible and that the problem of signalling signalhood had to be solved by the participants themselves. The pairs then played the game for 40 minutes uninterrupted for, on average, 255 rounds. Both players could move between the centres of each of their own quadrants using the arrow keys and finalised their movements with the spacebar, after which both players received feedback on their performance (Figure 2.1b) and continued to the next round. The game was stopped after 40 minutes. Participants then completed the PNS questionnaire and reported whether they thought that any communication had occurred. If any, they described the communication systems they developed or attempted to develop. Finally, they were debriefed and given the opportunity to discuss their experience. This study was approved by the Psychology Research Ethics Committee of Leiden University.[1]

[1] All code, materials, and data are available on OSF: `https://osf.io/n3uj6/`.

2.4.1 Measures

Game performance was measured using a *score* and *high score*. The *score* was increased by one point when both players ended on a quadrant with identical colours. The *high score* represents the number of consecutive successful rounds. *PNS* and its sub-factors were measured using a survey of 12 statements (see Neuberg and Newsom (1993) to see all statements), where the sum of all answers defines PNS; a higher sum corresponds to a higher *need for structure*. Here, items 3, 4, 6, and 10 correspond to sub-factor *F1* (i.e., the desire for structure in unstructured environments) and items 1, 2, 4, 7, 8, 9, 11, 12 sum to *F2* (i.e., the response to the lack of structure). Finally, participants described the communication system they developed via three open questions. We manually cross-checked the post-game descriptions, in which the participants described their communication systems, with the corresponding game data to validate whether both players reported identical systems, and to identify emerging patterns.

2.5 Results

Statistical analyses were performed using *R* 4.1.0 (R Core Team, 2023) and the *BayesFactor* 0.9.12-4.2 package (Morey et al., 2018). Our results align with those of Scott-Phillips et al. (2009), who found that out of 20 pairs, only 11 pairs managed to create a robust communicative system, confirming that this is not a trivial task. Participants perform on average 6.87 moves ($SD = 5.86$) per round and obtain a mean high score of 29.9 ($SD = 31.4$).

2.5.1 Emergence

The emergence of communicative systems happened in a similar manner to what was reported by Scott-Phillips et al. (2009); hence, we refer the reader to their work for a more elaborate description. Successful pairs typically converged on a default colour, allowing them to score above chance levels. This happened for 12 out of 20 pairs (note that one pair was not able to further develop a communication system beyond a default colour). Logically, this strategy failed when, in some of the following rounds, the default colour was not present in the quadrants. In such cases, a new convention had to be formed. Players typically did so by moving between quadrants in initially random directions. An initial convention was formed if these behaviours were recognised as communicative signals. Specifically, these random movements between quadrants could be recognised by the interlocutor as a communicative signal (e.g. “No, not the standard colour”). If the interlocutor did recognise this and, by mere chance, both players would finish on identically coloured quadrants, the initially random movements could be recognised as communicative signals and, henceforth, mapped to the finishing colour. From here, players could bootstrap their signalling behaviour when there were no colours available for which a signal exists and establish new signal-meaning mappings. These elaborate behaviours quickly

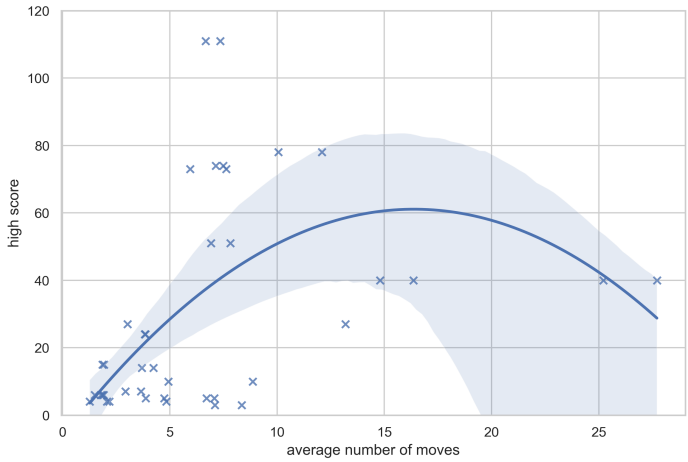


Figure 2.2: The quadratic relationship between the average number of steps over all rounds and high score.

became symbolic signals that participants explicitly recalled in their reports. The timing of convergence on a default colour was crucial towards a high score; pairs that quickly settled on a default colour typically evolved more elaborate and robust systems. Since the action space was rather limited, we observed patterns that are similar to the study by Scott-Phillips et al. (2009), namely oscillating up and down between quadrants, moving in (anti-)clockwise circles, forming U-shapes, or L-shapes. Despite having much overlap between communicative signals across experiments, the mapping to different meanings, i.e., colours, was specific to each experiment. Hence, the evolved systems were idiosyncratic to the pairs that evolved them and consequently would not be useful to immediately communicate successfully with new unseen partners. An example system of a successful pair is as follows: red was the standard colour, move there and wait for other signals. Moving in anti-clockwise circles indicated green, yellow was signalled by clockwise circles, and horizontal oscillations indicated blue.

Successful pairs agreed on a colour through dialogue. In a typical dialogue, one player initiated a signal after which the other copied it to confirm that colour. However, when that colour was not available, the recipient became the signaller and suggested another colour by using its corresponding signal. Such behaviour continued until both players agreed on a certain colour and finished the round. This robust system enabled participants to communicate successfully and gain high scores. We found that this is also reflected in the average number of moves participants made, where dialogue, quantified by the mean number of moves, has a quadratic relationship to higher scores, $F(2, 37) = 7.29, p = .002, R^2 = .28, R^2_{adjusted} = .24$ (see Figure 2.2). We also tested a linear relation between dialogue and *high score*, but found that this resulted in a lower fit ($F(1, 38) = 5.24, p = .02, R^2 = .12, R^2_{adjusted} = .09$). Moreover,

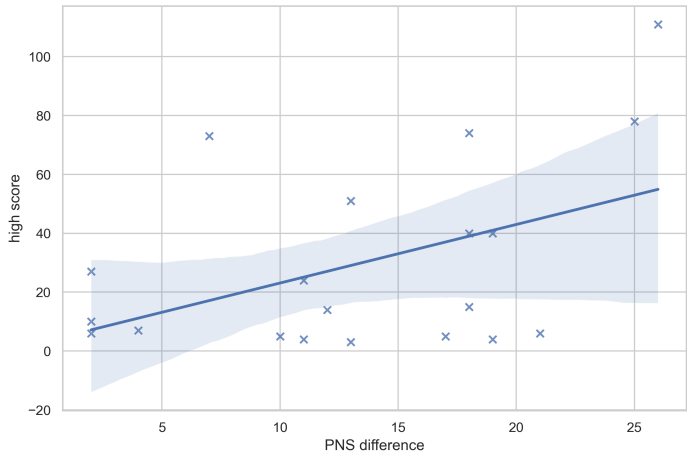


Figure 2.3: Pairs' difference in PNS score positively influences high score.

the quadratic relation remains best when outliers, the two points larger than 25 moves, are removed $R^2_{adjusted} = .28$ for quadratic regression and $R^2_{adjusted} = .19$ for linear regression. Taken together, this suggests that there appears to be an optimum number of moves: too few movements cannot convey communicative content, while too many movements can become confusing.

The reports of non-successful pairs typically describe that at least one participant tries to stick to their own system, not paying attention to the behaviours of the other. In some cases, participants even report having actively tried to communicate, whilst realising that their teammate did not notice and thus decided to unsuccessfully submit to their dominance. This is not trivial and often fails. This again shows that settling on conventions and the emergence of a communicative system requires *all* members to cooperate and interact actively.

2.5.2 Need for structure

Simple linear regression showed no relation between *PNS* ($M = 41.8, SD = 8.78$), *F1* ($M = 15.4, SD = 3.48$), *F2* ($M = 26.4, SD = 6.53$), and *high score* or the average number of moves on an individual level. However, the ECG enforces team cooperation of both players; we therefore combined individual scores to calculate team scores and assess team performance. We computed the difference in PNS between the two participants, and Figure 2.3 reveals that pairs with individuals that have a large difference in PNS score higher, $F(1, 18) = 4.869, p = .041, R^2 = .21, R^2_{adjusted} = .17$. This means that partners that respond differently to chaotic environments perform better in the ECG than those that have both either a high or low personal need for structure.

2.5.3 Comparing teams

As mentioned earlier, not all pairs were able to form a robust communication system and successfully convey their intentions. To further investigate why some are successful and some are not, we labelled games based on self-reports that describe the communication system that was used. After playing the game, participants individually reported on the communication system they thought was present, and the answers to these questions were cross-checked between pairs and used to split the pairs into groups. Teams were labelled as *good* ($n = 11$) when both participants individually reported identical signals for the same colours. They are labelled *medium* ($n = 3$) when there was partial overlap or when there was only a default colour, and *bad* ($n = 6$) otherwise. An analysis of variance (one-way ANOVA) showed that the mean high scores of these groups were significantly different, $F(2, 17) = 7.91, p = .004$. When we combined *medium* and *bad* performing pairs to have roughly equal sample sizes, the mean high scores were again significantly different, $t(18) = 4.07, p < .001$. This is expected because when two players can both recall the same systems, communication was probably successful in many consecutive rounds.

Although Figure 2.2 shows that pairs which use more movements do not necessarily reach higher scores, when comparing the two groups we do see that teams that performed well in the ECG, on average, moved more than those who performed worse ($M_{\text{good}} = 9.50, SD_{\text{good}} = 6.58, M_{\text{bad}} = 3.65, SD_{\text{bad}} = 2.29, t(27) = 3.89, p < .001$). This supports the assumption that well-performing teams have sufficient dialogue, which indicates that a pair can be considered to have a robust system (Scott-Phillips et al., 2009).

Figure 2.4 shows the correlations between team measures for pairs labelled as *good* and *medium* or *bad* performing pairs. A significant relationship between the difference in PNS scores and the high score is present for *good* teams ($r = .693, p = .018$). Since *PNS* is the sum of *F1* and *F2*, it allows us to investigate the main contributor to this effect. Differences in desire for structure (*F1*) do not explain higher scores ($r = -.107, p = .754$), yet differences in the response to the lack of structure (*F2*) do ($r = .78, p = .004$).

A Bayesian test for correlation between *PNS* difference and *high score* on good performing pairs yielded $\text{BF}_{10;\kappa=2} = 3.69$, indicating *PNS* difference positively influences the high score. For *F2* difference and high score, $\text{BF}_{10;\kappa=2} = 13.25$, confirming that a greater difference in the response to a lack of structure predicts higher high scores. We did not find these relationships in the group of medium and bad-performing pairs. This could be expected since high scores, in general, were lower for these pairs. Figure 2.3 shows that, although pairs with the largest differences in *PNS* or *F2* tend to score the highest, a relatively large difference in *PNS* or *F2* does not necessarily lead to a higher *high score*. We also observe pairs with a medium difference in *PNS* or *F2* that do not perform better than the lowest-scoring pairs in general. This indicates that diverse reactions to chaotic environments may be beneficial in establishing communication systems, but it does not guarantee success.

Figure 2.4: Comparison of the relations between all measures for good performing pairs (left) and medium or bad performing pairs (right). For good pairs, there is a positive correlation between the difference in Personal Need for Structure (*PNS_dif*), difference in the response to the lack of structure (*F2_dif*) and the obtained *high score*. These relations are not present for medium or bad pairs. In both groups *F2_dif* correlates with *PNS_dif*, while *F1_dif* does not, indicating that *F2_dif* is the main contributor of the relation between *PNS_dif* and *high score*. Note: the colour represents the correlation coefficient and the annotations correspond to p-values.

2.6 Discussion

In this chapter, we describe an experiment in which participants played the Embodied Communication Game from Scott-Phillips et al. (2009) and replicated their findings, while also introducing a novel way of comparing differences in game success. Paired participants had a shared goal without having access to conventional means of communication. This meant that they had to create a novel communication system that allowed them to coordinate their intentions. This non-trivial cooperation problem was typically solved through the formation of initial conventions (common ground) and a bootstrapping process. We extended the original work by incorporating a measure that allowed us to compare the cognitive traits of cooperating individuals. Results showed that a difference in personal need for structure between partners influenced the emergence of the communication systems in this game.

It is important to note that the current sample size limits the possibility of making far-reaching generalisations. Still, the results reveal intriguing relationships that provide insight into the working mechanisms of the emergence of communication systems and may inspire future work. When examining individual participants, no measure of personal need for structure, *PNS*, *F1*, and *F2* correlated with high scores. However, when comparing partners in a team, we found that team measures—defined as the difference in pairs' individual scores—influenced performance. Greater differences in *PNS* and *F2* positively correlated with a team's *high score*.

Situated in the ECG, this entails that pairs of individuals who respond differently to unstructured situations were more successful in building a communication system together. A split of pairs into *good*, *medium*, and *bad* teams revealed that this relation is only present for well-performing teams. We, therefore, concluded that, while our results indicate that diverse reactions to a lack of structure may be beneficial in creating a communication system together, this difference does not necessarily *guarantee* better performance in the ECG. Many other factors, of course, influence the complex process of social coordination, and here we have identified one of them. Therefore, we suggest also to study other factors and interactions between people. We propose not only to further investigate the relation of *PNS* to the creation of novel communication systems but also to include analyses of other personality traits, such as the Big Five personality inventory (McCrae et al., 2005) or other questionnaires that assess personality traits (e.g. leadership, submissiveness). This would allow us to investigate further how various combinations of traits influence the creation of novel communication systems and create a deeper understanding of what might lead to success in collaborative tasks.

Human language is highly structured. It is suggested that systematic patterns emerged in language because humans are naturally biased towards compressible systems, through a general preference for simplicity (Kemp and Regier, 2012; Kirby et al., 2015). Here, we investigated the influence of such a *bias for structure* in a task where participants had to cooperate and coordinate their signals. These biases also significantly affect the emergence of structure in language as languages are learned and transmitted across generations (Kirby et al., 2008; Theisen-White et al., 2011; Verhoef, 2012; Kirby et al., 2015). Such experiments of iterated transmission often also expose participants to initially unstructured systems, which then gradually become structured over generations of transmission. Yet, diversity in the bias for structure has never been used as a factor in these studies, as such we propose there is an opportunity to further investigate this by assessing how differences in PNS may affect the emergence of patterns in transmission chain experiments like those of Kirby et al. (2008); Theisen-White et al. (2011); Verhoef (2012). This could reveal whether, besides the processes of transmission and interaction (Kirby, 2017), a direct individual need for structure, or differences therein, indeed affect the evolution of signals. If the latter is true, this would provide more evidence for the benefits of diverse members in collaborative tasks. The effect of diverse members in groups on the emergence of signalling systems can also be investigated when the ECG is adapted to accommodate groups instead of pairs. It has been found that communicating with multiple interaction partners introduces pressures that result in more stable shared vocabularies (Raviv et al., 2019a). In combination with our findings (i.e. that the ECG is a non-trivial task for pairs), we speculate that establishing common ground and emerging signals in an adapted ECG will be more difficult for groups, but that once these are in place, they will be more robust. We, moreover, expect that groups consisting of diverse members that score differently on PNS will benefit from this and obtain higher high scores.

It seems obvious why alignment in expectations may aid cooperation; it makes it easier to

coordinate and predict the moves of another player. The reason why diversity in expectations may be beneficial in cooperation tasks may be less intuitive, but we suggest that differences between interacting partners might complement each other's weaknesses, possibly aiding cooperation. In light of the ECG, this happens when one partner actively tries to create structure, while the other is looking for structure. This simultaneously raises questions.

For instance, what is the relationship between an individual's need for structure and their willingness to accommodate during moments of misunderstanding or ambiguity? While a high personal need for structure might imply a preference for clarity and predictability, this could manifest either as rigid insistence on one's own expectations or as a willingness to accommodate the other's framework to re-establish clarity. Conversely, a low need for structure might afford greater flexibility in interpretation, but also less urgency to coordinate or accommodate under pressure. These nuances, moreover, suggest that establishing common ground may benefit from specific combinations of individuals' preferences. A dyad composed of two high-PNS individuals might appear aligned in their desire for structure, but diverge when their preferred structures conflict. Slightly mismatched levels of structure preference—where one individual seeks guidance and the other provides it—may, in such cases, result in smoother coordination. This functional complementarity could enable teams to balance the need for structure with adaptive responsiveness. We suggest targeted team formation based on different cognitive preferences as a fruitful research direction.

2.7 Conclusion

In general, we argue that novel insights can be obtained if we do not only focus on the systems invented by successful pairs in communication game studies but also investigate what might separate those who score high from those who perform worse. Contrary to what is usually assumed, namely that overlap in cognitive biases and similarities in expectations drives the emergence of shared systems (Tylén et al., 2013; Scott-Phillips and Kirby, 2010), we found that differences in personal need for structure also matter in cooperative tasks and that diversity of communication partners might be beneficial for the emergence of new communication systems. While more evidence is needed to support this benefit, we speculate that differences in biases or personalities can aid by complementing the weaknesses of partners in unfamiliar collaborative situations, such as language evolution. We propose that novel insights can be obtained by focusing on targeted differences between interacting pairs that have been unable to communicate successfully. Finally, we suggest including other personality traits and investigating the exact workings of the dynamics between mixed prior expectations, personality traits and the emergence of novel communication systems.

3

Computationally Modelling Human Emergent Communication

In this chapter, we study human sequential behaviour by integrating cognitive, evolutionary, and computational approaches. Our work revolves around the emergence of shared vocabularies in the Embodied Communication Game (ECG). Here, participant pairs solve a shared task without access to conventional means of communication, enforcing the emergence of a new communication system. This problem is typically solved by negotiating a shared set of sequential signals that acquire meaning through interactions. Individual differences in Personal Need for Structure (PNS) have been found to influence how this process develops. We trained deep neural networks to mimic the emergence of new communicative systems in humans and used hyperparameter optimisation to approximate latent human cognitive variables in an attempt to explain human behaviour. We demonstrate that models based on bidirectional LSTM networks are better at capturing human behaviour than unidirectional LSTM networks. Suggesting that, in the ECG, human sequence processing is influenced by expected future states. The approximated variables cannot explain the differences in PNS, but we do provide evidence suggesting that random and uncertainty-directed exploration strategies are combined to develop optimal behaviour.

Originally published as: Kouwenhoven, T., Verhoef, T., Raaijmakers, S.A., de Kleijn, R.E. (2023). Modelling Human Sequential Behavior with Deep Learning Neural Networks in Emergent Communication. In M. Goldwater., F. K. Anggoro., B. K. Hayes., & D. C. Ong., editors, *Proceedings of the Annual Meeting of the Cognitive Science Society*, Volume 44, pages 549-555. Cognitive Science Society.

3.1 Introduction

For communication—between humans or between humans and machines—to be successful, the coordinated actions of all interlocutors must adhere to the *grounding criterion*. Accordingly, interlocutors have to agree on the meaning of the current communicative purposes (Clark and Brennan, 1991). The fulfilment of this criterion relies extensively on the availability of a (partially) shared vocabulary between interlocutors of a conversation (Pickering and Garrod, 2004). Yet, the exact dynamics of how humans or agents settle on an effective grounded shared vocabulary are still unclear (Tylén et al., 2013; Mordatch and Abbeel, 2018). Recent work in computational linguistics started modelling emergent communication setups using multi-agent simulations to understand this process better (e.g. Lazaridou et al., 2018; Chaabouni et al., 2019a, 2020, 2022). However, the findings from these simulations often do not align with the outcomes of similar experiments with humans (Lazaridou et al., 2020; Galke et al., 2022). As such, literature proposes to instil human language patterns in machines by including human feedback in the learning loop instead of only learning from large quantities of data (ter Hoeve et al., 2022; Brandizzi and Iocchi, 2022), or by inducing additional artificial human-like biases into machines (Galke and Raviv, 2025).

The interdisciplinary research presented here attempts to instil such human communicative behaviour in machines, using an experimental setup that allows studying the initial emergence of simple signals where no communication existed before. As such, we explore the grounding problem from an evolutionary perspective, where humans must collaboratively create a novel, shared communication system to play the ECG successfully (Scott-Phillips et al., 2009). This two-player game addresses two fundamental questions in the emergence of languages: how does a signal obtain its communicative intent, and how does this signal obtain its meaning? Most human participants can solve this non-trivial task by establishing an initial convention (i.e., settling on a default behaviour) and collaboratively bootstrapping new signals onwards (Scott-Phillips et al. (2009), Chapter 2). These meaningful signals are subsequently used to play the ECG successfully, creating sequences of communicative behaviour.

Once a communicative system exists, it must be processed by the brain for comprehension and production. However, it is not entirely clear how this happens for human languages. Traditional views see the human brain as a forward-looking prediction machine (e.g. Clark, 2013), but recent findings indicate the importance of backward-looking processes for language comprehension in two self-paced reading and eye-tracking tasks (Onnis et al., 2022). Specifically, context, in the form of preceding words, can be informative for integrating current words. As such, Onnis et al. concluded that both forward *and* backward-looking appear to be important characteristics of language processing. A similar debate exists regarding the processing of everyday sequential actions (De Kleijn et al., 2014). Early accounts suggested that sequential actions are triggered by the perception of motor execution of the previous action (Washburn, 1916). Yet, there is also evidence that anticipated future states also influence subsequent actions

and that planning mechanisms play a role in sequential tasks (e.g. Lashley et al., 1951; Cohen and Rosenbaum, 2004; de Kleijn et al., 2018); however, how this happens exactly is hitherto not well understood.

Context, in the form of preceding behaviour or incoming signals, and intended future states also play a role in the ECG. Incoming and produced signals (i.e., context) are informative of future behaviour, and anticipated future states can be thought of as desired behaviours by the other (i.e., ending on a specific colour). The behaviours in the ECG are moreover sequential but less complex than everyday actions and can therefore be studied in a relatively controlled manner. As such, investigating this through computational modelling may reveal how sequential processing possibly played a role in shaping human language, what types of agent architectures are required to facilitate natural communication between humans and machines, and contribute to the debate on sequential action processing in humans.

From a computational view, we use behaviour cloning to 1) investigate whether deep learning models can learn the expressed human behaviours during the development of signal–meaning mappings in the ECG; 2) approximate latent human cognitive variables by optimising model parameters that influence learning and exploration (for an overview of similar work, see Schulz and Gershman, 2019); 3) identify the applicability of networks with different processing directions to model human behaviour. We then relate the model parameters with a cognitive measure of Personal Need for Structure (Thompson et al., 1989) and compare the ability to learn human behaviour for models with different processing directions and mechanistic learning preferences. Doing so has the potential to facilitate more natural human-machine interactions through the development of (language) models that possess shared biases, resulting in a more human-like quality. Vice versa, deviations between human and computational biases provide a better understanding of why outcomes of computational simulations might not be as desired. Lastly, a better understanding of the influence of such biases on the emergence of language could steer learning mechanisms in computational simulations of emergent communication and close the gap between evolved human and computational behaviour.

3.2 Background

The origin of language is extensively studied, but the exact dynamics of language emergence remain unknown. One question concerns the origins of the initial signal–meaning mappings in case no prior communication system exists. If neither form nor meaning is known, a possible way to establish this concerns the cooperative process of agreement on the relations between communicative signals and meanings. This process has been studied extensively through laboratory experiments in which participants invent and negotiate novel signals to solve a cooperative task (Steels, 2006; Scott-Phillips and Kirby, 2010; Tylén et al., 2013). These studies show that humans can establish shared conventions and develop communication systems through social coordination. It is, moreover, suggested that in addition to language use, human

learning and the transmission of a language affect the emergence of patterns (Kirby et al., 2015; Smith, 2022). A paramount explanation for the highly structured nature of human language is that it emerges due to a human bias for compressible systems, driven by a preference for simplicity (Kemp and Regier, 2012; Kirby et al., 2015; Kirby and Tamariz, 2022).

The Personal Need for Structure Scale is a measure that assesses the presence and degree of a human bias for simplicity (Thompson et al., 1989). This questionnaire quantifies individuals' need for structure (*PNS*), desire for structure and cognitive simplicity (*F1*), and the aim of restructuring an environment into a more manageable and simplified form (*F2*) (Neuberg and Newsom, 1993). Differences in the desire for structure influence how individuals understand and interact with the world (Neuberg and Newsom, 1993) and also affect problem-solving capabilities (Eva et al., 2014; Svecova and Pavlovicova, 2016). Furthermore, *PNS* affects the task progression of participants playing the ECG in that participant pairs who respond differently to a lack of structure are more successful (Chapter 2).

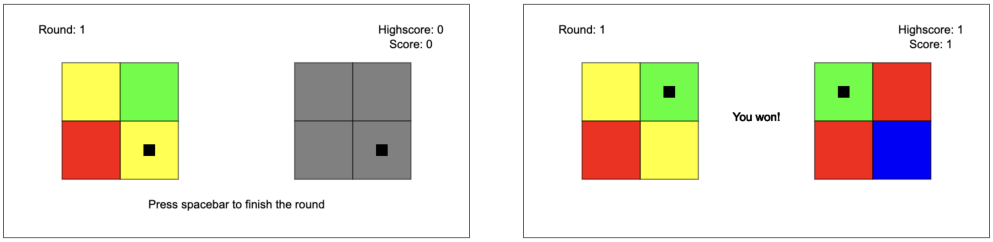
3.2.1 Embodied Communication Game

The ECG is a cooperative two-player game consisting of two 2×2 grid worlds. Each quadrant of the grid has one of four colours. Both players move between the quadrants, using the arrow keys, and share the goal of ending on identically coloured quadrants. When they manage to do so, they score a point. For both grids, the colours and starting positions are determined randomly for each round, with the proviso that there is one overlapping colour such that it is always possible to score a point, i.e., communicate successfully. Players see their own movements and the movements made by their partner, but only see the colours of their quadrants (Figure 3.1a). The colours of both worlds are revealed to both players (Figure 3.1b) when both finish moving. Their goal is to score as many consecutive points as possible, meaning that pairs must find a way to communicate reliably and coordinate behaviours (see Scott-Phillips et al., 2009, for an in-depth explanation).

3.2.2 Modelling Human Behaviours

Our work attempts to model human (sequential) behaviour using computational methods. Similar work by de Kleijn et al. (2018), for example, used reinforcement learning (RL) models to fit human behaviour in a serial reaction time (SRT) task and found that good human performance requires a high learning rate and a low discount factor. Suggesting that low-scoring individuals do not update their action-value function or the expected utility of their actions. Curricularised learning for RL agents in the SRT task showed that similar to infants' curiosity-based learning, exploration can promote robust later learning in virtual agents (de Kleijn et al., 2022).

For textual data, Nikolaus and Fourtassi (2021) evaluated the ability of neural networks to acquire meanings of words and sentences through laboratory tasks that involve cross-situational learning used with children. They showed that neural networks mirror learning patterns of



(a) The view while participants are playing.

(b) The view after both players ended the round.

Figure 3.1: An overview of the two possible game states. While the players are moving, only the participants' own grid is coloured (3.1a). When both players are done, the colours of all quadrants are revealed to both players and feedback is provided (3.1b).

acquiring semantic knowledge in early childhood and suggested that children might use partial representations of sentence structure to guide semantic interpretation. Additionally, language models seem to rely more on word frequency than children, but like children, learn words more slowly when these are part of longer utterances (Chang and Bergen, 2022). These models notably differed from children in the effects of word length, lexical class, and concreteness on learning, emphasising the importance of social, cognitive, and sensorimotor experience in child language development.

3.3 Methods

In this chapter, we attempt to investigate the relationships between computational hyperparameters and cognitive measures through training deep neural networks on human behaviours in the ECG. Specifically, algorithmic hyperparameters are used as a *proxy* of human preferences. We do not claim the existence of exactly these representations in the human brain, but merely use them as another measuring device of human behaviour.[1]

3.3.1 Data

The data used in this chapter was collected for the study described in Chapter 2. Here, we conducted three additional experiments ($N = 46$: 36 females, 10 males; $M_{\text{age}} = 22.2$, $SD_{\text{age}} = 3.53$). Participants received instructions after which they were separated and placed behind two connected computers. This setup ensured that conventional communication was impossible and that the problem of emerging signal–meaning mappings had to be solved by the participants. The game was played for 40 minutes, for an average of 256 rounds, after which participants completed the PNS questionnaire and described the communication systems they attempted to

[1] All code, materials, and data are available on OSF: `https://osf.io/n3uj6/`.

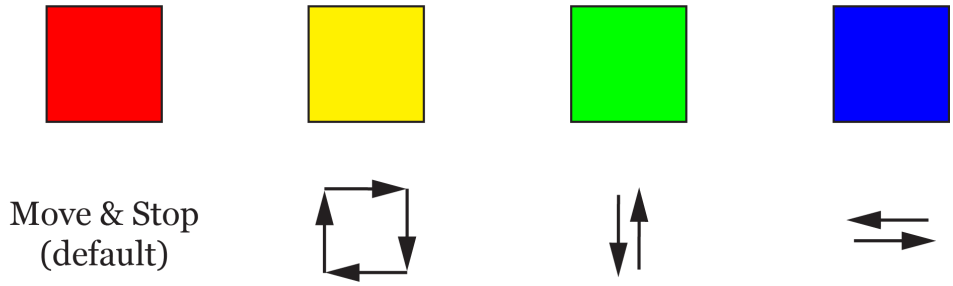


Figure 3.2: An example communication system established by participants. In this system, participants would default to a red quadrant or signal another colour through repetitive movements (displayed by the arrows).

develop. Finally, they were debriefed and allowed to discuss their experience. The Psychology Research Ethics Committee of Leiden University approved this study.

Out of 23 pairs, only 14 managed to create (i.e., reported and demonstrated) a robust communicative system. A Bayesian *t*-test showed that these pairs achieved higher scores than pairs that did not establish a system ($BF_{10} = 26.73$). A typical system contains sequences of movements (i.e., signals) to indicate different colours (i.e., meanings), an exemplary system is displayed in Figure 3.2. Once established, pairs negotiate which colour is available to both by repeating the sequential moves associated with this colour. We refer the reader to Scott-Phillips et al. and Chapter 2 for a detailed description of the emergence of such communicative behaviour.

A sequence of game states, produced by the movements of each participant, is stored for each round. These game states are a digital representation of the visual environment participants see and are used to train our neural networks. A single state contains the players' position, the position of the other player, the colour of the currently occupied quadrant, and the entire colour layout of the players' grid. This representation reflects the information that a participant sees during the game. A target label—corresponding to arrow keys and the spacebar—is stored for each game state, creating a sequence of state-actions pairs. The target label serves as a class label that is predicted by our deep learning model and is used to compute the prediction loss required to update the model.

3.3.2 The model

We trained a deep neural network—implemented with Long Short Term Memory (LSTM, Hochreiter and Schmidhuber, 1997) cells—on the state–action sequences of each participant. The input data, therefore, differs for each model, but its architecture is generic and fixed (Figure 3.3). The objective of the model is to predict a participant's subsequent move given a particular sequence of states. For unidirectional processing, each state of a sequence is processed

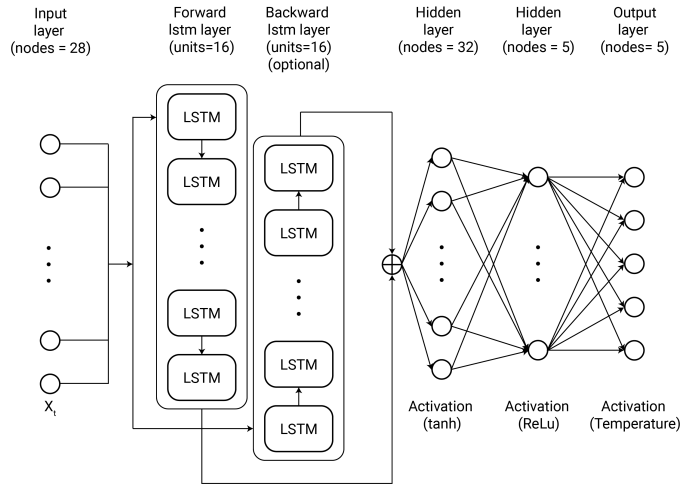


Figure 3.3: The neural network architecture used to model human behaviours. The model input X_t is the state at time t. The output layer uses temperature scaling as an activation function.

chronologically, beginning with the first and ending with the last state[2]. For bidirectional processing, the states are additionally processed in reverse order, thus incorporating (i.e., anticipating) future behaviour to predict a subsequent move. The model output layer computes probabilities for subsequent moves using temperature (τ) scaling. Here, high values of τ cause actions to be approximately equiprobable, and therefore lead to exploratory behaviour. Low values of τ result in greater differences between the probabilities, with higher probabilities for actions with higher expected rewards, and lead to deterministic behaviour. The model learning rate (lr) influences how quickly it updates its predictions, where a high learning rate means quick changes. The Adam optimisation algorithm (Kingma and Ba, 2015) is used to minimise categorical cross-entropy loss.

3.3.3 Measures

Game performance was measured by the number of consecutive successful rounds (*high score*). *PNS* and its sub-factors were collected using a 12-statement questionnaire (see Neuberg and Newsom, 1993), here, high values for *PNS*, *F1*, and *F2* correspond to a high need for structure. To obtain participant-specific τ and lr, we performed hyperparameter optimisation on the game data of each participant, resulting in 46 independently trained models. Put differently, an exhaustive grid search was used to optimise model performance using $lr \in \{0.0001, ..., 0.075\}$ and $\tau \in \{0.001, ..., 3.00\}$, with 10 equally spaced steps per parameter, resulting in 100 param-

[2]The backward processing layer is not used for unidirectional networks.

eter settings per participant. Each model was trained independently for five epochs on each parameter combination. We take the learning rate as a *proxy* of the extent to which individuals weigh feedback when updating their estimates and use temperature as an *approximation* of how explorative their behaviour was. The ability of the model to predict human sequential behaviours is reflected in its accuracy (*acc*). Lastly, the categorical cross-entropy loss (*cce*, i.e., negative log-likelihood) explains how likely the model and human would perform the same action in a particular game state. For each model, we used three-fold cross-validation to ensure that the model was not learning the data explicitly but captured the underlying structures of that participant. The cross-validation score (i.e., the average over all folds) described model performance. The parameter combination that resulted in the highest cross-validation score was used as a *proxy* for the latent human cognitive variables.

3.4 Modelling human sequential behaviour

Behaviour cloning was used to explain human behaviour in the ECG on two accounts. Firstly, by comparing PNS measures with the computational parameters. Since Neuberg and Newsom (1993) showed that differences in the need for simple structure influence how individuals understand and interact with the world, the inferred computational parameters, such as learning rate and temperature, may capture these effects as well. Therefore, we sought correspondence between these parameters and the PNS scores of each participant. We hypothesised that learning rate relates to the desire for cognitive simplicity (*F1*) and *high scores* since a desire for structure implies active searching for patterns, which seems crucial to learning signal–meaning mappings in the ECG. Learning these patterns more quickly (i.e., high lr) might result in faster emergence of communicative patterns. Individuals who feel uncomfortable in unstructured environments (i.e., high *F2*) show lower adaptability and flexibility in new environments, preferring to respond with familiar behavioural patterns to counter the uncomfortable feeling (Steinmetz et al., 2011). Since lower values of τ correspond to less exploratory behaviour and a high lr corresponds with high adaptability, it was expected for lr and τ to correlate negatively with *F2*.

Secondly, we manipulated the sequential processing cells of the models. As argued before, the next move of a signal and the intended finishing colour influence immediate action selection and can therefore be thought of as an anticipated future state. As such, optimisation as described in the previous section is done for the unidirectional (LSTM) and bidirectional LSTM (biLSTM) models. Whereas unidirectional cells process time steps of sequences in a chronological forward manner, bidirectional cells compute inputs forward *and* backwards to make predictions (Schuster and Paliwal, 1997). Note that although the LSTM layer in our model differs for both types, the remaining architecture is identical.

Table 3.1: The average model performance (*acc*) over the cross-validation scores for each participant and the average optimal learning rate and temperature across participants. *Uni* and *bi* correspond to the model types LSTM and biLSTM respectively.

	acc		cce		lr		τ	
Type	M	SD	M	SD	M	SD	M	SD
Uni	.831	.112	.355	.241	.019	.019	.356	.745
Bi	.972	.055	.084	.153	.039	.020	2.28	.716

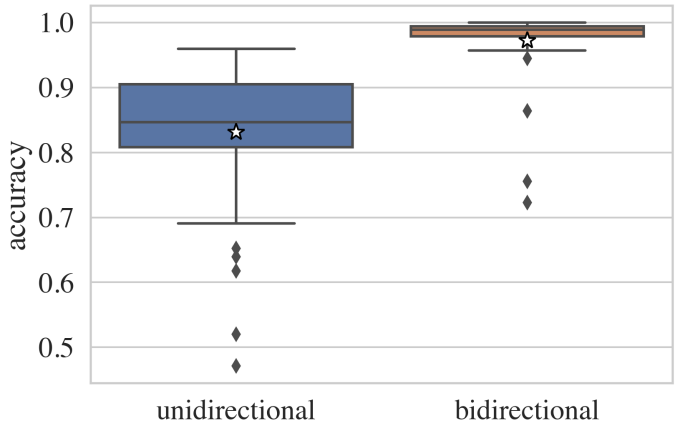
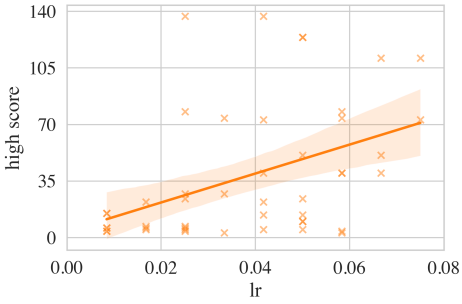


Figure 3.4: BiLSTM models show greater and more robust accuracy than LSTM models. Stars indicate mean accuracy.

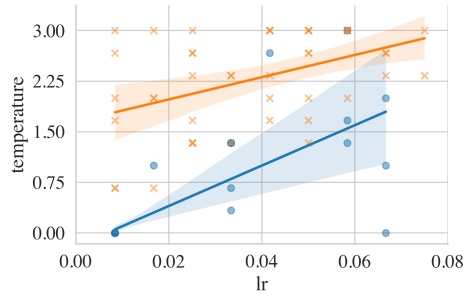
3.4.1 Results

Statistical analyses were done using *R* 4.0.5 (R Core Team, 2023) and the BayesFactor 0.9.12-4.3 package (Morey et al., 2018). First, we consider the overall performance of both network types. The mean accuracy (*acc*) over all independently trained models shows that both network types can learn to predict subsequent moves relatively well (Table 3.1).

Comparison between the two network types with a Bayesian *t*-test on *acc* and *cce* with network type as a predictor revealed a large performance difference ($BF_{10}acc = 6.63e+11, d = 1.66$ and $BF_{10}cce = 1.50e11, d = -1.59$). Indicating that bidirectional sequence processing can better capture the human behaviour in the ECG than unidirectional sequence processing (Figure 3.4). This result is robust when controlled for the number of parameters between the two network architectures. Optimal learning rate and temperature were higher for biLSTM networks when compared to LSTM networks ($BF_{10}lr = 5.85e3, d = .790$ and $BF_{10}\tau = 3.46e+14, d = 2.00$). Since the learning rate was taken as a proxy for the extent to which individuals update their estimates, a higher learning rate implies flexible behaviour. Therefore, this result



(a) Relationship between learning rate and high score.



(b) Relationship between learning rate and temperature.

Figure 3.5: Relationships between learning rate, high score, and temperature. Each point corresponds to one participant. Note: darker marks denote overlapping data points, and the shaded area is the 95% confidence interval. Blue is used for *unidirectional* networks and orange is used for *bidirectional networks.*

suggests that bidirectional processing requires more flexibility toward updating behaviour policies. Additionally, it implies that explorative behaviour might complement updating these policies. We can assume that a higher learning rate translates to better learning in humans since learning is required to play the ECG successfully and learning rates were significantly higher for pairs that managed to establish a communicative system compared to those that did not ($M_{successful} = .047, M_{unsuccessful} = .025, BF_{10} = 556, d = 1.39$).

We now consider the relationships between model parameters, cognitive measures, and high scores as described earlier. Successful participants (i.e., those with a high score) performed complex and structured sequences in order to communicate. Nevertheless, we find that for LSTM networks, but not for biLSTM networks, *high score* negatively influences *acc* ($BF_{10} = 3.07, r = -.346, r^2 = .120$). This suggests that unidirectional processing is able to learn simpler human behaviour relatively well but has difficulties capturing more elaborate behaviours. This finding may explain the difference observed in Figure 3.4.

Bayesian regression showed that for biLSTM networks, there is a positive linear relationship between learning rate and high score (Figure 3.5a $BF_{10} = 12.8, r^2 = .183$), confirming our hypothesis and suggesting that participants who adopt new behaviours faster are more successful in creating new signal–meaning mappings in the ECG. We moreover find that regardless of processing directionality, temperature, and learning rate are related (Figure 3.5b, $BF_{10} biLSTM = 28.1, r = .452, r^2 = .204$ and $BF_{10} LSTM = 1.40e7, r = .772, r^2 = .597$), suggesting that participants who explored more also adapted new behaviours faster. Surprisingly, we did not find a relation between exploration and *high score*. A relationship was expected since explorative behaviour may lead to new conventions in the ECG. Lastly, learning rate or temperature cannot explain *PNS*, *F1*, or *F2* for LSTM and biLSTM networks. Thereby also

rejecting the remaining hypotheses.

3.5 Discussion

In this chapter, we modelled human sequential behaviour in the Embodied Communication Game with deep neural networks and investigated possible relationships between human cognitive preferences and computational parameters. Specifically, we looked at relationships between participants' personal need for structure, learning rate, and temperature parameters. Though we showed that current deep neural networks can learn the behaviour associated with creating signal–meaning mappings, we did not find any correspondences between cognitive and computational variables. As such, *PNS*, used here as a human bias for structure (Kirby and Tamariz, 2022), cannot be captured with this setup. Further research should investigate how parameters of various network architectures may correspond to cognitive measures or look at different games that investigate emergent communication (e.g. Galantucci, 2005; Steels and Loetzsch, 2012; Mordatch and Abbeel, 2018). The ability to capture human biases, such as the human bias for compressible and simple systems (Kemp and Regier, 2012; Kirby et al., 2015), in computational systems is insightful for simulations of emergent communication as they are then closer to human experiments. Furthermore, playing these collaborative games between humans and machines might also result in shared grounded vocabularies that are adapted to the biases of humans *and* computers, ultimately resulting in better conversational AI (Chapter 1).

Manipulation of the processing directionality of action sequences showed that participants' behaviour was explained better by biLSTM models than by LSTM models. This thereby provides additional arguments for the bidirectional processing of sequential actions in humans (Lashley et al., 1951; Cohen and Rosenbaum, 2004; Onnis et al., 2022). For communicative purposes in the ECG, integrating current actions is dependent on the preceding shared context (i.e., the negotiations of signals and intended final colours), and must be taken into account when deciding what moves to take next. The difficulties for LSTM networks to learn more complex behaviours performed by more successful participants also indicate that unidirectional processing is insufficient to capture more elaborate human behaviour. Although additional analysis is needed to support this, these findings suggest that the effect of a backward-looking mechanism found by Onnis et al. (2022) in a self-paced reading task might originate in the very early stage of forming signalling conventions. To verify this, simulations of emergent communication with deep learning agents should look at the effect of processing directionality of network architectures on the structure of emergent communicative protocols. Integrating bidirectional networks may close the current gap between human experiments and simulations.

We demonstrated that for biLSTM networks, the learning rate has a positive influence on high scores and is positively correlated with temperature (Figure 3.5b). This seems to support the recent view which suggests that humans combine random and uncertainty-directed exploration strategies to develop optimal behaviour (Jepma et al., 2016; Schulz and Gershman, 2019). An

explanation for this could be that explorative behaviour in the ECG led to the emergence of new signals, which need to be learned quickly (i.e., require a high learning rate) to be useful. In other words, the correlation between learning rate and temperature likely reflects the fact that participants who are more explorative benefit from higher learning rates (i.e., there is no benefit to explorative behaviour if you do not use the explored options to update expected values). However, a more in-depth analysis is required to strengthen this link further. For optimal behaviour, learning rate and explorative behaviour would be expected to decrease over time as strategies are learned and exploration becomes less necessary, instead exploiting the knowledge gathered thus far. However, literature on how learning rate and temperature parameters develop with age and experience has yielded conflicting results (Nussenbaum and Hartley, 2019). Games like the ECG could be extended over time to investigate the dynamic nature of the temperature and learning rate parameters.

Lastly, we acknowledge that the ECG is a highly simplified setup, thereby limiting the generalisability to real-world processing (Nastase et al., 2020). It also goes without saying that these models are mere approximations of the human brain and do not capture its breadth, but we can nevertheless use them as a proxy to mimic human processes. These findings must therefore be replicated in more ecological settings.

3.6 Conclusion

In this chapter, we modelled sequential human behaviour captured in the Embodied Communication Game with deep neural networks. Here, participants establish a communication system from scratch to solve a collaborative task. We demonstrate that neural networks can learn the human behaviours associated with the creation of a new communication system. Manipulation of network types shows that bidirectional processing of sequential actions better explains human behaviour than unidirectional processing, hereby providing additional arguments for the existence of a planning mechanism for sequential signal production in humans. No relationship was found between Personal Need for Structure and participant-specific computational parameters, but our results suggest that humans combine random and uncertainty-directed exploration strategies to develop optimal behaviour in the ECG. Future research should attempt to extrapolate our results to communicative settings with complex linguistic signal exchange (e.g., between chatbots and humans). Additionally, experiments on the emergence of a more complex human–AI language will deepen the understanding of the relationship between natural and artificial biases that play a role during the emergence of communicative systems.

4

Kiki or Bouba?

Humans have clear cross-modal preferences when matching certain novel words to visual shapes. Evidence suggests that these preferences play a prominent role in our linguistic processing, language learning, and the origins of signal-meaning mappings. With the rise of multimodal models in AI, such as vision-and-language (VLM) models, it becomes increasingly important to uncover the kinds of visio-linguistic associations these models encode and whether they align with human representations. Informed by experiments with humans, we probe and compare four VLMs for a well-known human cross-modal preference, the bouba-kiki effect. We do not find conclusive evidence for this effect, but suggest that results may depend on features of the models, such as architecture design, model size, and training details. Our findings inform discussions on the origins of the bouba-kiki effect in human cognition and future developments of VLMs that align well with human cross-modal associations.

Originally published as: Tessa Verhoef*, Kiana Shahrasbi, and Tom Kouwenhoven*. 2024. What does Kiki look like? Cross-modal associations between speech sounds and visual shapes in vision-and-language models. In Kuribayashi, T., Rambelli, G., Takmaz, E., Wicke, P., Oseki, Y., editors, *In Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 199–213, Bangkok, Thailand. Association for Computational Linguistics. (*denotes equal contribution.)

4.1 Introduction

The development of machine understanding and generation of natural language has benefited immensely from the introduction of transformer-based architectures (Vaswani et al., 2017). These architectures have since then been adapted and extended to handle multimodal data, leading to the creation of various types of multimodal models, including vision-and-language models. These models can potentially revolutionise how AI systems understand the world and interact with humans. However, we lack direct access to the exact representations and associations they encode. How VLMs integrate representations in the two modalities and whether associations between modalities are made in a human-like way is still being actively investigated (Alper et al., 2023; Kamath et al., 2023; Zhang et al., 2024c; Karamcheti et al., 2024; Jones et al., 2024).

Here, we use a well-known paradigm from the field of cognitive science to probe into a specific cross-modal association between speech sounds and visual shapes: the bouba-kiki effect. When humans see two figures, one with jagged and one with smooth edges, and are told one is a Kiki and the other a Bouba, 95% will name the jagged figure Kiki (Ramachandran and Hubbard, 2001). This effect was initially discovered and described anecdotally by Wolfgang Kóhler (Köhler, 1929, 1947), using the two images shown in Figure 4.1 with the labels *maluma* and *takete*. Since then, it has been widely studied (as reviewed in Section 4.2), and expanded with many other cross-modal preferences in human processing of (speech) sounds and visual imagery. Moreover, a wealth of evidence suggests that such preferences widely influence patterns we see in human languages (e.g., Ramachandran and Hubbard, 2001; Cuskley and Kirby, 2013; Imai and Kita, 2014; Verhoef et al., 2015, 2016a; Tamariz et al., 2018). Even though non-arbitrariness in language is often still regarded as an exception in some disciplines, in fields such as language evolution and sign language linguistics, iconic form-meaning mappings are considered omnipresent (Perniss et al., 2010). Given the central role cross-modal preferences play in human visio-linguistic representations and their effects on language, it is pertinent to investigate whether VLMs associate non-words and visual stimuli in a human-like way.

Examining universal human cross-modal preferences in VLMs can help us gain key insights across disciplines. First, it may reveal whether VLMs process multimodal information in a human-like way and whether similar biases drive their understanding of visual-auditory form-meaning mappings. Overlap in cognitive biases can potentially increase mutual understanding and improve interactions between humans and machines (Chapter 1). Second, it may help pinpoint what is missing to make VLMs more suitable for realistic simulations of human language emergence. Increasingly, VLMs are used in emergent communication settings, where agents communicate with each other and develop a novel language (Bouchacourt and Baroni, 2018; Mahaut et al., 2025). These models are used to improve machine understanding of human language (Lazaridou and Baroni, 2020; Lowe et al., 2020; Steinert-Threlkeld et al., 2022; Zheng et al., 2024), but also to simulate and study human language evolution processes (Galke et al.,

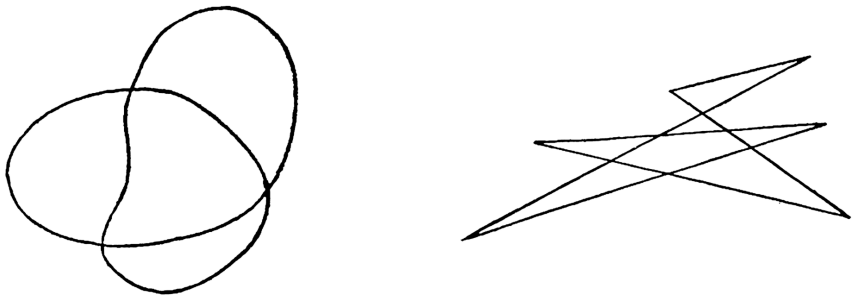


Figure 4.1: Which of these two shapes is Kiki? Images from Köhler (1929, 1947).

2022; Lian et al., 2023b). While the influence of cross-modal associations on the emergence of language has been studied extensively in language evolution experiments with humans (Verhoef et al., 2015, 2016a; Tamariz et al., 2018; Little et al., 2017), this phenomenon remains absent from current emergent communication paradigms. Evidently, cognitively plausible VLMs are more suitable for simulating aspects of the evolution of meaning in language. Finally, the actual origin of the bouba-kiki effect is still being debated within cognitive science and linguistics, with proposed explanations ranging from attributing it to similarities between shape features and features of either orthography (Cuskley et al., 2017), acoustics and articulation (Ramachandran and Hubbard, 2001; Maurer et al., 2006; Westbury, 2005), affective–semantic properties of human and non-human vocal communication (Nielsen and Rendall, 2011), or physical properties relating to audiovisual regularities in the environment (Fort and Schwartz, 2022). If the bouba-kiki effect can be reproduced in a VLM, it can help reveal the crucial ingredients for this effect, potentially leading to models better aligned with human representations.

To the best of our knowledge, only one previous paper discussed the bouba-kiki effect in VLMs. Alper and Averbuch-Elor (2023) tested two models, CLIP (Radford et al., 2021) and Stable Diffusion (Rombach et al., 2022), and reported finding strong evidence for the effect in these models. This is somewhat surprising given the way these models are trained and the absence of relevant data sources, such as auditory information and experience with physical object properties. Therefore, we introduce nuance in this discussion and show, contrary to the previous finding, that the bouba-kiki effect does not occur consistently in VLMs and that the presence of this cross-modal preference may depend on the way it is tested, as well as properties such as model architecture, attention mechanism, and training details.

4.2 Background

4.2.1 Sound-symbolism and cross-modal associations in language and cognition

When Hockett (1960) listed a set of design features deemed essential to natural human language, 'arbitrariness' was included. This feature refers to the arbitrary/unmotivated mapping between words and their meanings. However, when exploring beyond Indo-European languages, non-arbitrary form-meaning mappings appear to play a significant role in many languages (Imai et al., 2008; Perniss et al., 2010; Dingemanse, 2012). Most obviously, perhaps, sign languages are rich in non-arbitrary 'iconic' mappings, with articulators that lend themselves particularly well to representing meanings by mimicking, for example, shapes or actions. However, some spoken languages also have specific classes of words where characteristics of the meaning are mimicked or iconically represented in the word. Examples have been identified as 'ideophones,' 'mimetics', or 'expressives,' and this phenomenon is often referred to as sound-symbolism (Imai et al., 2008; Imai and Kita, 2014; Dingemanse, 2012). Even in languages not typically considered rich in sound symbolism, such as English and Spanish, vocabulary items from specific lexical categories, like adjectives, are also rated high in iconicity (Perry et al., 2015). Perhaps the most overwhelming evidence for the widespread importance of sound-symbolism in human languages comes from a study by Blasi et al. (2016), who analysed vocabularies of two-thirds of the world's languages and found evidence for strong associations between speech sounds and particular meanings across geographical locations and linguistic lineages. Consequently, non-arbitrariness is an important property of all languages.

In addition, human language learning, processing, and evolution are affected by cross-modal associations. Sound-symbolic mappings help young children acquire new words (Imai et al., 2008), and iconic words are learned earlier in child language development (Perry et al., 2015). Furthermore, parents use sound-symbolic words in their infant-directed speech more often than in adult-to-adult conversations (Imai et al., 2008). In a novel word learning task, participants trained on a mapping congruent with a known cross-modal association performed better than participants in an incongruent condition (Nielsen and Rendall, 2012). Sound-symbolic mappings in language have been connected to cross-modal mappings in the human brain (Ramachandran and Hubbard, 2001; Simner et al., 2010; Lockwood and Dingemanse, 2015) and processing of sound-symbolic words is less affected by aphasia (language-affecting brain damage after left-hemisphere stroke), than arbitrary words (Meteyard et al., 2015). It is also argued that universally shared cross-modal biases play an essential role in the evolution of language by bridging the gap between sensory input and meaning by providing a basis for linguistic conventions (Ramachandran and Hubbard, 2001; Cuskley and Kirby, 2013; Imai and Kita, 2014). Shared biases can help to create mutual understanding because communicative partners will automatically understand what is meant when a word like 'kiki' is used for the

first time in a context like the one shown in Figure 4.1.

While the bouba-kiki effect may be the most famous example of a universal cross-modal association, numerous other cognitive biases in cross-modal perception have been reported. For instance, non-arbitrary associations exist in human processing between high pitch sounds and light shades (Marks, 1974; Melara, 1989; Ward et al., 2006), light shades with rising intonation (Hubbard, 1996), graphemes and colours (Cuskley et al., 2019), vowel height and lightness (Cuskley et al., 2019), small size and high pitch (Evans and Treisman, 2010; Parise and Spence, 2009) and vowel openness and visual size (Schmidtke et al., 2014). Therefore, the findings presented in this chapter only scratch the surface of what is possible in this domain.

4.2.2 Testing the bouba-kiki effect in humans

After its initial discovery, the bouba-kiki effect has been studied increasingly rigorously, extending the initial pair of two images with more possible pairs (Maurer et al., 2006; Westbury, 2005) and even randomly generated ones to control for biases related to deliberate selection by the researchers (Nielsen and Rendall, 2011, 2013). In addition, various sets of labels and pseudowords have been contrasted and compared to study the relative importance of vowels versus consonants in the labels (Westbury, 2005; Nielsen and Rendall, 2011, 2013). The role of orthography, in addition to the auditory properties of speech sounds, has also been studied (Cuskley et al., 2017; Bottini et al., 2019). Across setups, non-arbitrary preferences are found to be robust across varying cultures and writing systems (Ćwiek et al., 2022). Remarkably, to some extent, this can even be found in blind individuals who undergo a haptic version of the bouba-kiki task (Bottini et al., 2019).

Most experiments in this domain are conducted using a two-alternative forced choice design, where two contrasting images are shown side by side (one jagged and the other curved), and two possible labels are offered, asking participants to make the "correct' mapping. However, it has been argued that this is an anti-conservative method in the sense that the concurrent presentation of two images that differ along one dimension and two labels that also differ along one dimension strongly primes participants to match the two, noticing their similarities. Nielsen and Rendall (2013) therefore introduced a different method, in which images are presented independently, and participants are asked to generate novel pseudowords to match the images. Here, we adopt their approach as a stringent method for probing VLMs for the bouba-kiki effect.

4.2.3 Vision-and-language models

Despite recent advances in multi-modal models (Zhang et al., 2024a) using transformer architectures, they remain poorly understood and often show unwanted behaviours such as poor visio-compositional reasoning (Thrush et al., 2022; Diwan et al., 2022) or spatial reasoning skills (Kamath et al., 2023). In addition, in the visual question-answering domain, it is a well-known

problem that models often lack visual grounding and struggle to integrate textual and visual data (Goyal et al., 2017; Jabri et al., 2016; Agrawal et al., 2018). This makes it perhaps even more puzzling that Alper and Averbuch-Elor (2023) found strong evidence for a bouba-kiki effect in CLIP and Stable Diffusion. Even if these models are able to extract sound-symbolic information in the absence of auditory data, they will likely struggle to actually associate that information with visual properties.

The approach taken by Alper and Averbuch-Elor (2023) involved generating two large sets of pseudowords, where one set was more likely associated with round shapes (examples: *bodubo, gunogu, momomo*) and the other set would evoke associations with jagged shapes (examples: *kitaki, hipehi, texete*). The CLIP embedding vector space was used to define a visual semantic dimension that best separates two sets of pre-selected adjectives (various synonyms of round and jagged). Within this space, pseudoword properties could reliably predict adjective type (round or jagged), and geometric properties associated with those adjectives could predict the category of pseudowords. With Stable Diffusion, novel images were generated based on pseudowords and analysed by embedding them using CLIP and through human evaluation. Both methods revealed evidence for the presence of sound symbolic mappings in these models (Alper and Averbuch-Elor, 2023).

While their methods mainly involved generating images from text (with Stable Diffusion) or investigating text-to-text mapping (with CLIP embeddings), we focus on image-to-text classification. We use images previously used in experiments with humans, as well as novel images generated following a procedure previously used to generate items for human experimentation. This approach provides an additional way of testing for cross-modal associations in VLMs and yields data that can be more directly compared to human data from studies into the bouba-kiki effect. If VLMs indeed learned human-like associations between visual and textual modalities, these should show robustly across multiple experiments that test the same associations differently. Moreover, Alper and Averbuch-Elor (2023) did not explicitly compare different VLMs (Stable Diffusion also uses CLIP). However, it would not be surprising if properties relating to the architecture, for example, affect the presence of this effect since these properties directly determine how the modality gap is bridged. Previous findings also suggest that dataset diversity and scale are the primary drivers of alignment to human representations (Conwell et al., 2023; Muttenthaler et al., 2023). As such, we compare four models here, each with a different architecture, attention mechanism, and training objective.

While many different architectures exist, they typically use single or dual-stream architectures. Either combining the inputs from two modalities and encoding them jointly (single-stream) or encoding them by two separate modality-specific encoders (dual-stream). Single-stream architectures typically use merged attention, where the language and visual input attend to both themselves and the other modality. Dual-stream architectures often use some form of cross-model attention, like co-attention and modality-specific attention, in addition to merged attention. Recently, Li et al. (2023) introduced a lightweight Querying Transformer (Q-Former) to

Model	Objective	Architecture	Attention	#Params	#imgs,#caps
CLIP	CON	Dual-stream	Mod-spec	151.3M	400, 400M
ViLT	ITM, MLM	Single-stream	Merged	87.4M	4.10, 9.85M
BLIP2	CON, IGTG, ITM	Dual-stream	Q-Former	∼3.8B	129, 258M
GPT-4o	Unknown	Unknown	Unknown	Unknown	Unknown

Table 4.1: Overview of the models used in this chapter. The training objectives are Image Text Matching (ITM), Masked Language Modelling (MLM), Image-grounded Text Generation (IGTG), or Contrastive Learning (CON). Mod-spec refers to modality-specific attention. Numbers are millions (M) or billions (B).

bridge the modality gap between any arbitrary pre-trained frozen vision model and a language model, resulting in BLIP2. Frequently, image text matching and masked language modelling are used as learning objectives (e.g., ViLT; Kim et al., 2021), but some methods use a contrastive learning objective (e.g., CLIP) or use image-grounded text generation loss (e.g., BLIP, BLIP2). The models used in this chapter are shown in Table 4.1. They are different in the above aspects, allowing investigation into the effect of their designs and training data on the cross-modal associations that are potentially learned. In addition, we include GPT-4o; even though no information is available for this model, its generative performance is unprecedented.

4.3 Methods

To test for the presence of a bouba-kiki effect in VLMs, we employ previously used as well as newly generated images (Section 4.3.1) and use a method for constructing pseudowords (4.3.2) that is directly borrowed from Nielsen and Rendall (2013). Probing (Section 4.3.3) was used to obtain image-text scores, and responses were analysed in two ways (Section 4.3.4).[1]

4.3.1 Image selection and generation

The original set of images used by Köhler (1929, 1947), as shown in figure 4.1, has been expanded in subsequent experiments. Maurer et al. (2006) for example introduced additional line drawings and Westbury (2005) used images with white shapes on a black background. Here we use the original pair and the two sets of four image pairs by Maurer et al. (2006); Westbury (2005). In

[1] All code, materials, and data are available on OSF: `https://osf.io/3w7k9/`.

addition, we generated new random curved and jagged images using a method inspired by Nielsen and Rendall (2013). We generated 10 uniformly distributed points within a circle with a radius of 1. These points were connected with either smooth curves or straight lines. For curved images, we generated curves that pass through the given points such that they form a closed path. Jagged images were generated by connecting the ordered points with straight lines, also forming a closed path. All images are displayed in Section A.1.

4.3.2 Pseudoword generation

Following the experiment with human participants conducted by Nielsen and Rendall (2013), we present the VLMs with a constrained set of syllables that can be used to construct novel pseudowords. Based on previously established cross-modal association patterns, Nielsen and Rendall (2013) selected sets of vowels and consonants that were expected to evoke a sense of correspondence with either jagged or curved visual shapes. We adopt exactly their set here, consisting of sonorant consonants M, N and L and rounded vowels OO, OH and AH, expected to match to curved shapes, and plosive consonants T, K and P and non-rounded vowels EE, AY and UH, expected to match to jagged shapes. Syllables were created by making consonant-vowel combinations. In total, 36 different syllables (e.g., loo, nah, kee, puh) can be constructed in this way, with nine different versions of each syllable type: sonorant-rounded (S-R), plosive-rounded (P-R), sonorant-non-rounded (S-NR) and plosive-non-rounded (P-NR).

In addition to single syllables, we generated pseudowords by concatenating two syllables, as this was exactly the task human participants were asked to complete in the experiment (Nielsen and Rendall, 2013). However, since we are not primarily interested here in distinguishing the separate roles played by consonants versus vowels in the bouba-kiki effect, and Nielsen and Rendall (2013) demonstrated that both have an impact, we limit the set of possible syllables in two-syllable probing to combinations of S-R syllables and P-NR syllables.

An important difference between the human setup and our work is that their participants also listened to a spoken version of the pseudowords, whereas our models are only exposed to the written form. Since the bouba-kiki effect is most often assumed to integrate vision and sound, this may influence the result. However, the relation between orthographic shapes and the sounds they represent is not arbitrary either and has presumably been shaped by human iconic strategies in their development and evolution (Turoman and Styles, 2017). This perhaps also explains why a role for English orthography has been demonstrated in the bouba-kiki effect for humans (Cuskley et al., 2017), while at the same time it is robust across different writing systems (Ćwiek et al., 2022).

4.3.3 VLM probing

To assess the preferences of BLIP2, CLIP, and ViLT, in each trial, we extract probabilities for all possible labels (i.e., syllables and pseudowords) conditioned on an image. Instead of only

embedding the label, each label is fed in a sentence ("The label for this image is {label}") such that embedding the textual input is closer to the models' natural objective[2]. Importantly, only the labels differ between inferences such that variance in the probability given an image is only caused by the label of interest. Where Alper and Averbuch-Elor (2023) use an *indirect* metric by embedding the inputs in CLIP space, our method uses the model probabilities as a more *direct* measure of how well a given syllable or pseudoword matches a novel image. For GPT-4o, we prompt the model to generate a label and use its probability directly (Appendix A.2).

4.3.4 Analysis

All findings were analysed for statistical significance using Bayesian models with the *brms* package (Bürkner, 2021) in R (R Core Team, 2023). To analyse VLM probability scores, we fitted Bayesian multilevel linear models (4 chains of 4000 iterations and a warmup of 2000, family = *gaussian*) to predict probability with image shape (Jagged versus Curved), consonant (plosive or sonorant) and vowel (rounded or non-rounded) categories ($Probability \sim shape * (consonant + vowel)$). For all models of this type, the random effects structure consists of varying intercepts for image and label with by-label random slopes for shape. When comparing proportions of vowels, consonants, or selected pseudoword types, we fitted Bayesian logistic models (4 chains of 1000 iterations and a warmup of 500, family = *binomial*) to test whether shape predicts the occurrence of particular vowels, consonants or pseudoword types ($Occurrence|trials(SampleSize) \sim Shape$). Effects are considered significant when the computed 95% Credible Interval does not include 0, i.e., the lower and upper bounds of the CI must be either both positive or both negative. All plots were created in ggplot2 (Wickham, 2016).

4.4 Results

The findings are analysed in two ways. First, we compare the results of VLM probing to the performance of human participants (Nielsen and Rendall, 2013). For BLIP2, CLIP and ViLT this means we first only consider the syllable or pseudoword with the highest probability for each image. These are then analysed similarly to those selected by humans or generated by GPT-4o. Second, we examine the probabilities for *each* possible syllable or pseudoword from BLIP2, CLIP and ViLT, to obtain a more comprehensive measure of cross-modal associations. For the GPT-4o results reported below, one image in the Jagged shape condition is consistently missing since it (top right image in Figure A.2 in Section A.1) was flagged as 'content that is not allowed by our safety system'.

[2]Additional analysis revealed that the overall results remain consistent even when only the label is provided.

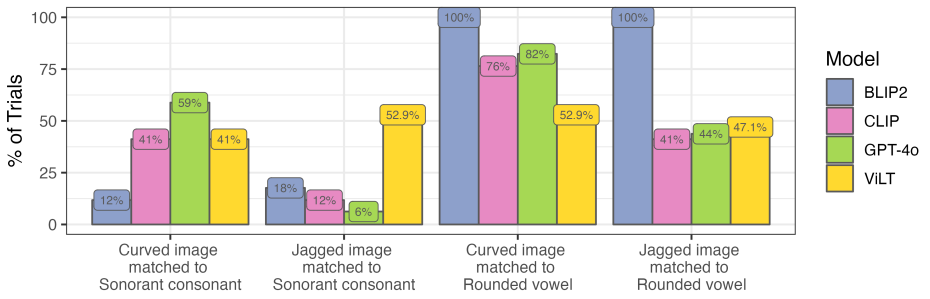


Figure 4.2: Percentages of trials in which selected syllables contain sonorant consonants or rounded vowels, separated by image shape (Jagged or Curved) for all four VLMs. Human percentages as reported by Nielsen and Rendall (2013) are (from left to right): 52.4%, 45.1%, 56,9%, and 48.3%.

4

4.4.1 Single syllable selection

VLMs were first probed using single syllables; here, we are interested in seeing if the models predominantly pair Jagged images with P-NR and Curved images with S-R syllables, as was found with humans. Figure 4.2 shows these results as the percentage of trials (where each individual image of the set of 17 pairs forms a trial) in which model probabilities were highest for sonorant consonants or rounded vowels with either Curved or Jagged shapes. A result that fits the expected human pattern would show higher bars for the Curved than for the Jagged shapes in both sets. The only models where this seems to go in the right direction are CLIP and GPT-4o. BLIP2 mostly displays a general preference for P-R syllables, without considering the shape, and ViLT does not display any clear preference. To test whether the differences in percentages for CLIP and GPT-4o are significant, we use Bayesian logistic models (as described in Section 4.3.4). For both models, Jagged images are paired with sonorant consonants significantly less often than Curved images (CLIP: b = -1.79, Bayesian 95 % Credible Interval $[-3.86, -0.05]$, GPT-4o: b = -3.51, 95 % CI $[-6.69, -1.37]$) and Jagged images are paired with rounded vowels significantly less often than Curved images (CLIP: b = -1.62, 95 % CI $[-3.06, -0.19]$, GPT-4o: b = -1.97, 95 % CI $[-3.66, -0.36]$).

4.4.2 Probability scores for novel syllables

While GPT-4o only selects the best-fitting syllable out of all options for each image, CLIP, BLIP2, and ViLT provide probability scores for each possible syllable, yielding more comprehensive data. Here, we therefore also analyse the probability scores for these three models to investigate whether higher scores occur when pairing S-R syllables with Curved images than with Jagged images and vice versa for P-NR syllables. Figure 4.3 shows the probabilities for the pseudoword

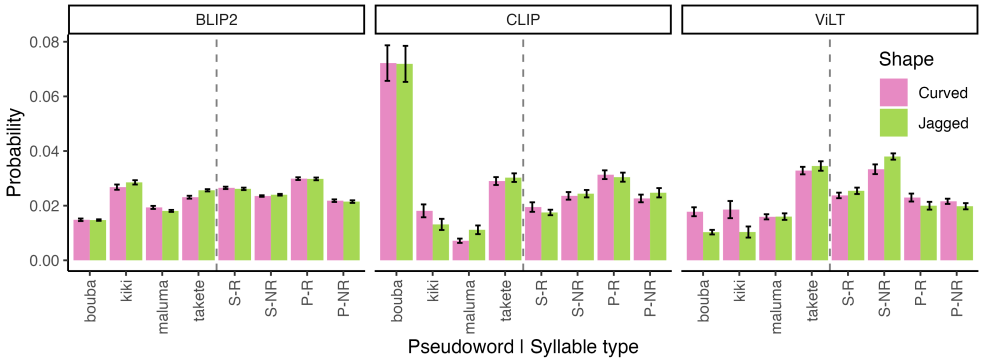


Figure 4.3: Probability scores for the original pseudowords (bouba, kiki, takete and maluma), as well as for the four different generated syllable types: Sonorant-Rounded (S-R), Sonorant-Non-Rounded (S-NR), Plosive-Rounded (P-R) and Plosive-Non-Rounded (P-NR), paired with two types of shapes (Jagged or Curved) for three VLMs.

pairs that were used in the classic experiments with humans (bouba & kiki, takete & maluma) and the four different syllable types (S-R, S-NR, P-R, P-NR).

Looking at the original pseudowords, none of the models display a clear bouba-kiki or takete-maluma effect. Probabilities for the different words differ overall (with a curiously high probability for 'bouba' in CLIP), but this does not seem modulated by the visual shape. For the syllables, BLIP2 shows no shape-modulated variation at all, and ViLT displays contradictory patterns (e.g. higher probability scores for S-NR than S-R syllables with Curved shapes and higher scores for S-NR with Jagged than with both P-R and P-NR). Only CLIP gets close to the expected pattern, with equal scores for the ambiguous syllable types (S-NR and P-R) but slightly higher scores for P-NR with Jagged and S-R with Curved. Yet, no significant effects are found when testing whether CLIP shows a pattern of preferring the expected consonants and vowels with their associated shapes using a Bayesian multilevel linear model (as described in Section 4.3.4). For ViLT, we find one (tiny) interaction between shape and consonants in the opposite direction of what is expected, where scores for Jagged shapes are significantly higher when paired with sonorant versus plosive consonants (b = .0056, 95 % CI $[.0001, .0112]$). For BLIP2, we find a significant overall preference for rounded vowels (b = 0.0055, 95 % CI $[.0019, .0091]$), but no other effects.

4.4.3 Two-syllable pseudoword selection

Although the results in Nielsen and Rendall (2013) were analysed by looking at single syllables, the actual task human participants performed involved creating novel pseudowords consisting of two syllables. We therefore also used our VLMs to generate (GPT-4o) or provide probability scores (CLIP, BLIP2 and ViLT) for two-syllable pseudowords that were created by concatenating

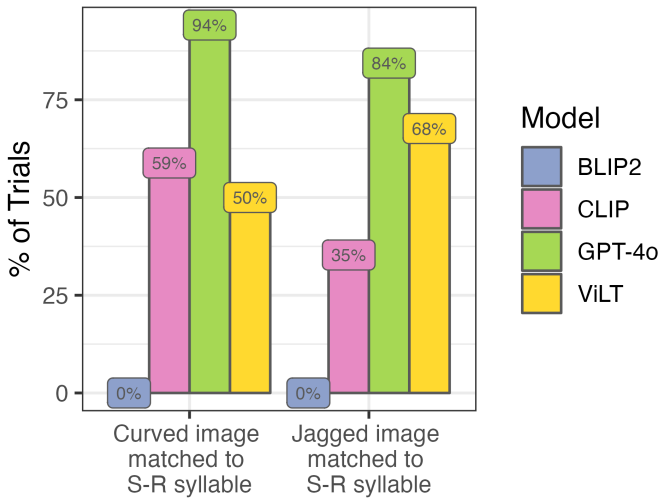


Figure 4.4: Percentages of trials in which Jagged or Curved visual shapes were matched to Sonorant-Rounded (S-R) syllables embedded in two-syllable pseudowords for all VLMs. Here 0% for S-R syllables implies a 100% preference for P-NR syllables.

two of the possible syllables from the set of S-R (most Curved) and P-NR (most Jagged) syllables resulting in 324 words. For CLIP, BLIP2 and ViLT, we first look at the 'preferred' pseudowords by only considering the option with the highest probability score for each image. Figure 4.4 shows the percentages of trials in which S-R syllables were matched to either Curved or Jagged images, counting each one of the two syllables in a word separately. BLIP2 never used S-R syllables and only selected pseudowords that contained two P-NR syllables, regardless of which image was shown. Both CLIP and GPT-4o show a higher percentage of Curved matched to S-R compared to Jagged, but GPT-4o seems to mostly just prefer S-R syllables overall. A manual inspection of GPT-4o's generated pseudowords revealed that in 25 out of 33 trials, the word 'nohmoh' was used, 12 times for Jagged and 13 times for Curved images. For ViLT, if a preference is present, it is in the wrong direction. In the case of CLIP, we find that Jagged images are indeed paired with S-R syllables significantly less often than Curved images (b = -1.00, 95 % CI $[-2.04, -0.04]$).

4.4.4 Probability scores for novel two-syllable pseudowords

We obtained probability scores for all possible two-syllable pseudowords when paired with each image for CLIP, BLIP2 and ViLT. Figure 4.5 shows these results by plotting probabilities for four different pseudoword types. The pseudoword on the left combines two P-NR syllables and is therefore expected to result in higher probabilities for Jagged shapes. Conversely, the

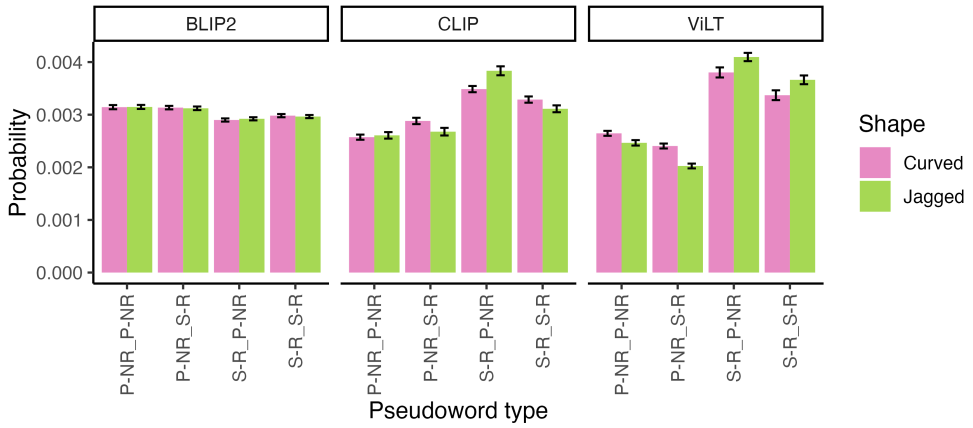


Figure 4.5: Probability scores for four pseudoword types, combining Sonorant-Rounded (S-R) and Plosive-Non-Rounded (P-NR) syllables, paired with two types of shapes (Jagged or Curved) for three VLMs.

most right pseudoword combines two S-R syllables and should evoke higher probabilities for Curved shapes. In the latter case, a pattern in which pink (Curved) bars rise while green (Jagged) bars fall would therefore reflect evidence for the bouba-kiki effect. None of the tested VLMs fit this pattern. Since GPT-4o generated 'nohmoh' (and similar variants like 'moomoh') almost exclusively when given the freedom to select two syllables from the full set of Jagged-associated and Curved-associated syllables, we also independently obtained probabilities for both syllable types. For this, we asked GPT-4o to generate a pseudoword for each image twice, once when given only the set of Jagged-associated syllable options, and once with only the Curved-associated syllables as options. Yet again, no significant effect of shape on probability scores for different syllable types was found. Figure A.4 in Section A.3 shows this result.

4.4.5 Summary

In summary, the bouba-kiki effect appeared absent for BLIP2 and ViLT, while for CLIP and GPT-4o, the results varied depending on how the effect was tested and the results were analysed. When asking the model to select one best-fitting syllable, CLIP and GPT-4o both display the effect in the expected direction. However, this pattern disappears when looking at a richer dataset of probability scores (from CLIP, BLIP2, and ViLT) for each possible syllable. In the case of two-syllable words, GPT-4o results no longer display significant evidence for a bouba-kiki effect.

4.5 Discussion

Our findings partly contradict previous work, which found that sound-symbolic associations are present in CLIP and Stable Diffusion (Alper and Averbuch-Elor, 2023). A possible reason for this could be that we use a different method, focusing on image-to-text probabilities, which is more similar to how the effect has been tested with humans. If VLMs indeed learned human-like cross-modal associations, we should be able to observe them in various experimental setups, i.e., the results should be robust. Given the contradicting findings, we suggest that it is too early to conclude that VLMs understand sound-symbolism or map visio-linguistic representations in a human-like manner, as the results depend heavily on which specific model is tested and how the task is formulated.

The asymmetry between the results coming from our method and those of Alper and Averbuch-Elor (2023) implies that performance is influenced by the method used. But perhaps more urgently, there is also contradicting evidence *within* the same method. In a replication of Alper and Averbuch-Elor (2023)'s experiment for Japanese, Iida and Funakura (2024) found that Japanese VLMs did not exhibit the expected bouba-kiki effect, despite Japanese being a language rich in sound-symbolism (Dingemanse, 2012; Ćwiek et al., 2022). Hence, Kouwenhoven et al. (2025) suggest that the method used to disambiguate sharp and round pseudowords and images may pick up on relationships between semantic concepts and word forms—being heavily entangled with the choice of ground-truth adjectives—rather than capturing true sensory mappings in languages. This is unsurprising given that CNN-based models often classify based on superficial textural rather than shape features (Baker et al., 2018; Geirhos et al., 2019; Hermann et al., 2020) and, albeit less so, this texture bias is also present in vision transformers (Geirhos et al., 2021). Moreover, Darcet et al. (2024) identified that, during inference, ViT networks create artefacts at low-informative background areas of images that are used for computations rather than describing visual information. Both findings are in stark contrast with what, at its core, is required for sound symbolism. However, the fact that some evidence for a bouba-kiki effect could be found in two of the four models tentatively suggests that real-world physical experience with different object properties may not be needed to develop this cross-modal preference but that it can, to some extent, be learned from statistical regularities in data containing text and images.

Human language on its own already contains many non-arbitrary regularities between speech sounds and meaning (Blasi et al., 2016), and these regularities, like phonesthemes (Bergen, 2004), can be detected and interpreted by models such as word embeddings (Abramova and Fernández, 2016) and LSTM-based language models (Pimentel et al., 2019). No visual input is needed for this, and perhaps this is also what caused the appearance of the observed bouba-kiki effect in the work by Alper and Averbuch-Elor (2023). In our work, we gave more prominence to the visual input and found much less convincing evidence for the effect. With two complementary methods closely modelled after human experiments, Kouwenhoven et al.

(2025) attempted to interpret the visual attention patterns of two variants of CLIP in a shape-word matching task. Neither of the models showed performance in line with the expected associations. Direct comparison with prior human data on the same task, additionally, showed that the models' responses fall markedly short of the robust, modality-integrated behaviour characteristic of human cognition. Finally, qualitatively, they showed that both CLIP variants do not focus on sharp edges or round attributes of images, but instead mostly focus on the centres of shapes or background areas. Both observations are in contrast with what, at its core, is required for a bouba-kiki-like effect.

Regarding the design features of the models we tested, we observe that the model with the best bouba-kiki alignment to human preferences, CLIP, is also trained on the largest amount of data (comparing the three models we have information about, not including GPT-4o). This finding aligns with previous work showing that dataset properties affect alignment with human representations (Conwell et al., 2023; Muttenthaler et al., 2023). However, despite having many more parameters than CLIP, BLIP2 does not show the effect. In addition, while both BLIP2 and CLIP use dual-stream architectures, only CLIP, which uses modality-specific attention mechanisms, displays some evidence of a bouba-kiki effect. Despite impressive performance on vision-language tasks, the Q-Former in BLIP2 apparently does not promote sound-symbolic associations. This is important knowledge for developing models with vision-language representations that align with those of humans. Especially since more aligned models show more robust few-shot learning (Sucholutsky and Griffiths, 2023) and promote more natural interactions between humans and machines (Chapter 1). Although we find modest evidence for a bouba-kiki effect in GPT-4o, we cannot know the origin of this effect as model details are unknown.

4.6 Conclusion

Given the pervasive role that cross-modal associations play in human linguistic processing, learning, and evolution, we tested for the presence of a bouba-kiki effect in four VLMs that differ along various dimensions such as architecture design, training objective, number of parameters, and input data. Evidence for this effect is limited, but not entirely absent, in the tested VLMs. These findings inform discussions on the origins of the bouba-kiki effect in human cognition and future developments of VLMs that align well with human cross-modal associations.

4.7 Limitations

Our work has a few notable limitations. First, we used synthetic images that had been previously used in experiments with humans. Even though this makes our results easily comparable to those of human studies, there is a potential risk that these images are out-of-domain for models

that are predominantly trained on realistic images. In future extensions of this work, we therefore plan to include more naturalistic images.

A second limitation manifests itself in the tokenisation of the textual input. While humans in the experiment evaluate pseudowords as a whole, the tokenisation process in language models may split our syllables or pseudowords into tokens that would not necessarily evoke the expected cross-modal associations in humans either (e.g., a separate evaluation of H in OH may invite a jagged association instead of a curved one). Despite being a fundamental difference, the primary goal of this chapter was to assess the preferences of VLMs in their most basic form. Further work should investigate whether tokenisation affects results and identify whether there may be model-specific cross-modal associations on a token instead of a word level.

Third, the pseudowords we used were based on an experiment with humans but were different from those used by Alper and Averbuch-Elor (2023), who did find a strong bouba-kiki effect in CLIP embeddings. To allow for a better comparison with their findings, future work should also test our image-to-text approach with their set of pseudowords.

Finally, our experiments included a relatively small number of trials, limited by the availability of experimental stimuli from human studies. However, by combining images from several previous studies and augmenting this set with additional newly generated images, we used more trials than most studies conducted with humans. The set of generated images can easily be expanded in future work. But then again, given the current pattern of results, this is not expected to lead to a more robust bouba-kiki effect in most models.

5

The Curious Case of Representational Alignment

Natural language has the universal properties of being compositional and grounded in reality. The emergence of linguistic properties is often investigated through simulations of emergent communication using referential games. However, these computational experiments have yielded mixed results compared to similar experiments that address the linguistic properties of human language. Here we address representational alignment as a potential contributing factor to these results. Specifically, we assess the representational alignment between the image representations agents have and between agent representations and input images. By doing so, we confirm that the emergent language does not appear to encode human-like conceptual visual features, as the image representations of agents drift away from their inputs while inter-agent alignment increases. We moreover identify a strong relationship between inter-agent alignment and topographic similarity, a common metric for compositionality, and discuss its consequences. To address these issues, we introduce an alignment penalty that prevents representational drift but interestingly does not improve performance on a compositional discrimination task. Together, our findings emphasise the key role representational alignment plays in simulations of language emergence.

Originally published as: Tom Kouwenhoven, Max Peeperkorn, Bram van Dijk, and Tessa Verhoef. 2024. The Curious Case of Representational Alignment: Unravelling Visio-Linguistic Tasks in Emergent Communication. In Kuribayashi, T., Rambelli, G., Takmaz, E., Wicke, P., Oseki, Y., editors, *In Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 57–71, Bangkok, Thailand. Association for Computational Linguistics.

5.1 Introduction

Human language bears unique properties that make it a powerful tool for communication. A well-known property is compositionality: the ability to combine meaningful words into more complex meanings (Hockett, 1959). The emergence of compositionality is studied extensively in the field of language evolution through human experiments (Selten and Warglien, 2007; Kirby et al., 2008, 2015; Raviv et al., 2019a, inter alia). A key finding in this field is that the unique nature of human language can be explained as a consequence of a general preference for simplicity and a pressure to be expressive, both of which are imposed during continuous language learning and use (Smith, 2022). Computational simulations of language emergence have also been used to study the emergence of linguistic properties (e.g. de Boer, 2006; Steels and Loetzsch, 2012), and have seen a rising interest in the field of computational linguistics (Lazaridou and Baroni, 2020). Here, compositionality in the emergent communication protocols is commonly measured through a quantitative proxy for compositionality known as topographic similarity (*TopSim*; Brighton and Kirby, 2006). This metric was first introduced to contemporary computational simulations by Lazaridou et al. (2018) and has been used in a large body of work since. Conceptually, this metric gauges whether similar meanings map to similar messages (see Section 5.4.4). Yet, the interpretation of linguistic properties emerging in simulations remains challenging, since language protocols used among artificial agents often show critical mismatches with known properties of human languages (Galke et al., 2022; Lian et al., 2023b) such as efficiency, word-order vs. case-marking biases, or compositional generalisation (see Section 5.2). Only when human-like biases are introduced artificially, do languages with human-like properties emerge (Galke and Raviv, 2025). Consequently, it is evident that the biases of artificial agents in recent simulations and the signal-meaning mappings they make differ from those of humans. This underscores the critical need to obtain deeper insight into referential games in the language learning setting (Rita et al., 2022b).

A possible explanation for these mismatches could stem from representational alignment—the degree of agreement between the internal representations of two information processing systems (Sucholutsky et al., 2023). To the best of our knowledge, representational alignment in emergent communication was first reported by Bouchacourt and Baroni (2018), who measured the degree to which agents aligned their internal image interpretations (inter-agent alignment) by performing Representational Similarity Analysis (*RSA*; Kriegeskorte et al., 2008). Using *RSA* (see Section 5.3), they showed that agents establish successful communication in an artificial manner by aligning their internal image representations while *losing* any relation to the images presented (image-agent alignment). This enabled them to communicate about noise input even though they were trained on real images. As such, their communication protocol captured not conceptual properties of the objects depicted in pictures, but most likely focused on non-human-like spurious image features (e.g., pixel intensities). While inter-agent alignment is not a problem per se, the loss of image-agent alignment is problematic for two reasons. First, for

emergent communication simulations to provide meaningful insights into the emergence of natural human language, agent image representations must be grounded in the content of the images. Only then can we deduce *what* the agents communicate about and assess linguistic properties or their ability to generalise to novel concepts. Second, emergent communication setups have been proposed to fine-tune pre-trained (vision-)language models, aiming to enhance machine understanding of natural human language (Lazaridou and Baroni, 2020; Lowe et al., 2020; Steinert-Threlkeld et al., 2022; Zheng et al., 2024). In this context, maintaining substantial alignment between representations and images is crucial for preserving mutual understanding between machines and humans.

Representational alignment, however, did not receive the necessary attention since a host of papers appeared *after* Bouchacourt and Baroni shared their findings. In these papers, results on referential games were reported without taking *RSA* into account (e.g. Lazaridou et al., 2018; Guo et al., 2019; Li and Bowling, 2019; Ren et al., 2020; Chaabouni et al., 2020; Dagan et al., 2021; Mu and Goodman, 2021; Chaabouni et al., 2022). Admittedly, some use attribute-value objects instead of real images as input. But *importantly*, in nearly all cases, neural agents must map inputs—whether attribute-value objects or image representations—onto agent-specific representations. The problem of inter-agent alignment can, therefore, *always* occur and is *agnostic* to the input type. Although this warrants further analysis of earlier results, the field is already employing referential games in more complex simulations with real images (e.g. Dessi et al., 2021; Chaabouni et al., 2022; Mahaut et al., 2025).

This chapter addresses the understudied alignment problem in standard referential game setups used in emergent communication. We train Reinforcement Learning (RL) agents equipped with a recent vision module (DinoV2; Oquab et al., 2024) to communicate about images. In addition to evaluating the agents on MS COCO (Lin et al., 2014) image pairs, we assess them on noise pairs and image pairs sourced from the Winoground dataset (Thrush et al., 2022). The latter is explicitly created to gauge the visio-linguistic compositional reasoning abilities of vision and language models. We first confirm that effective communication in the referential game relies on inter-agent alignment and then continue with our contributions. First, we find a strong correlation between the degree of inter-agent alignment and the *TopSim* metric. Our second contribution involves a solution to the alignment problem by incorporating an alignment penalty term to the loss, resulting in equivalent communicative success and higher *TopSim* whilst ensuring that the agents communicate about images instead of spurious features. We then argue to start evaluating emergent communication protocols on more stringent tasks that directly target the intuition behind popular metrics to obtain a better understanding of the protocols used. Overall, our results highlight the importance of representational alignment in simulations of language emergence and underscore the need to better understand the divergence in human and artificial language emergence.

5.2 Background

Most research in simulating emergent communication is modelled after the Lewis signalling game (Lewis, 1969) with a speaker and a listener agent. The speaker observes a state (e.g., an image) and sends a signal to the listener, who acts based on this signal. In the case of the referential game, this means selecting a target among a set of distractors. Both agents are rewarded for successful communication, meaning the listener points to the target object. The solution to this game requires the agents to have a shared protocol (i.e., an artificial language), which typically emerges when the agents learn based on trial and error over multiple games. This resembles how, for humans, language learning and use impose constraints such as pressures for learnability and compression that shape our language design (Kirby et al., 2014, 2015). Importantly, the emergent language in the case of simulations with artificial agents is also shaped by biases resulting from, for example, the agent architecture, loss function, and learning protocol (Rita et al., 2022b). The current work uses the referential game: a variant of the Lewis signalling game extensively used to explore language evolution (e.g. Steels and Loetzsch, 2012; Kirby et al., 2015; Lazaridou et al., 2017; Kottur et al., 2017; Lazaridou et al., 2018; Kharitonov et al., 2020; Chaabouni et al., 2022).

An important challenge in emergent communication is that artificial learners often do not behave the same manner as human learners in experimental settings. Some emergent protocols do not follow Zipf's law and thus are anti-efficient unless pressures for brevity are introduced (Chaabouni et al., 2019a), others do not show the word-order vs. case-marking trade-off found in human languages (Chaabouni et al., 2019b; Lian et al., 2021). Additionally, there is an ongoing debate on the degree to which the emergent languages allow for compositional generalisation (Lazaridou and Baroni, 2020; Conklin and Smith, 2023). As such, it has been suggested to introduce communicative (e.g., alternating speaker/listener roles) and cognitive (e.g., memory) constraints (Galke et al., 2022) and use more natural settings to promote more human-like patterns of language emergence with neural agents (Chapter 1). Doing so changes the learning pressures to which the agents need to adapt and can recover initially absent linguistic phenomena of natural language in emergent languages (for a review see Galke and Raviv, 2024). An example of such work, investigating the word-order vs. case-marking trade-off, has successfully replicated this trade-off for neural learners (Lian et al., 2023b). Their setup differs from other work in that agents first learn a miniature language via supervised learning, and then optimise it for communicative success via RL, resulting in emergent languages that share linguistic universals with human language.

To enhance understanding of emergent communication in the Lewis game, Rita et al. (2022b) decomposed the standard objective in Lewis games into two key components: a co-adaptation loss and an information loss. In doing so, they shed light on potential sources of overfitting and how they might hinder the emergence of structured communication protocols. They demonstrated that desired linguistic properties (e.g., compositionality and generalisability)

emerge when they control the listener's ability to converge to the speaker agent (i.e., control for overfitting on the co-adaptation loss). While the co-adaptation loss has parallels to inter-agent alignment, their work does not address the alignment between the agents' image representation and the input features, which we deem crucial in developing grounded communication protocols.

Another challenge in emergent communication is the disentanglement of the underlying meanings of emergent languages. Earlier studies by Lazaridou et al. (2017) suggested that agents assign symbols to general conceptual properties of objects in images, rather than low-level visual features. However, as previously mentioned, follow-up work from Bouchacourt and Baroni (2018) showed this is not always the case. They found that agents align their agent-specific image representations without developing a language that captures conceptual properties depicted in the images. Moreover, agents lost any sense of meaningful within-category variation where two similar objects in human perception (e.g., two avocados) were observed as maximally dissimilar for the agents. In response to these findings, recent studies have implemented sanity checks testing whether trained agents can communicate about noise (Dessi et al., 2021; Mahaut et al., 2025). However, to the best of our knowledge, there has been little attention to what we consider to be their main result: the alignment problem.

5.3 Representational alignment

Representational alignment is the degree of agreement between the internal representations of two information processing systems, whether biological or artificial. Even though widely recognised in cognitive science, neuroscience, and machine learning (Sucholutsky et al., 2023), representational alignment has not seen much interest in the field of emergent communication, except for the work by Bouchacourt and Baroni who analysed the referential game using *RSA*. This metric measures the alignment between two sets of numerical vectors, for example, image embeddings and agents' representations thereof. In practice, it is calculated by taking the pairwise (cosine) distances between vectors of a set and calculating the Spearman rank correlation between these distances.

In this chapter, we also use *RSA* to operationalise representational alignment. Given the speaker image representations r_s of the DinoV2 input embeddings i and r_l as the same images represented in the listener image representation space, we compute the pairwise cosine similarity between the representations for the speaker s_s and for the listener s_l and calculate Spearman's ρ between s_s and s_l. As such, *RSA* measures the degree of inter-agent alignment (RSA_{sl}) between image representations s_s and s_l, relative to their input. Additionally, we use *RSA* to measure image-agent alignment between the speaker and listener image representations and the DinoV2 embeddings (RSA_{si} and RSA_{li} respectively). It is important to stress that representational alignment is *agnostic to the type of input*—being either images or attribute-value objects—and can always happen when inputs are projected onto agent-specific representations.

Now that representational alignment is formalised, we turn to the question of what it means if agents align their representations. Intuitively, a high inter-agent RSA_{sl} value can be interpreted as agents with *similar* representations for similar images. Importantly, this can have two causes: both agents' image representations either *maintain* a relation to the image input (i.e., have a high RSA_{si} and RSA_{li}), or *lose* this relation (i.e., they have a low RSA_{si} and RSA_{li}). While the former is desirable, the latter means that the agents' image representations diverged from their input, but did so in a similar way. Since the agents' image representations are used to compose a message, low image-agent alignment means that they are not communicating about the same high-level image features that are captured by DinoV2, but are likely communicating about non-human-like spurious features. In the case of a low inter-agent alignment (RSA_{sl}) value, something similar happens. This entails that the agents have developed *different* interpretations for the same image, e.g., the speaker maintains a close relation to the input image while the representation of the listener drifts away. While this may be similar to the question of whether people have different perceptual experiences of colour (Locke, 1847), in the case of emergent communication, agents should develop a referentially grounded vocabulary with overlapping concept-level properties since we wish machines to have a more natural understanding of human language. To unravel how representational alignment plays a role in emergent communication, we use *RSA* 1) as a metric to re-assess findings from Bouchacourt and Baroni and 2) implement it as an auxiliary loss to mitigate the alignment problem and ensure that the agents communicate about image features.

5.4 Methods

The standard Lewis referential game is used as provided by the commonly used EGG framework (Kharitonov et al., 2021). This ensures that our findings are representative of this setup, rather than being influenced by specific design decisions. The game implementation is a multi-agent cooperative RL problem where a speaker and a listener communicate to discriminate a target image from two shuffled distractor images. The speaker receives a target image t and generates a message m of at most length L, using vocabulary V. Using message m, the listener guesses which of the two images is the target image ($\hat{t}$). Communicative success is defined as $\hat{t} = t$, meaning that the listener correctly identified the target image among the candidate images. The speaker, crucially, *only* observes the target image and does not see the distractor images. As such, the speaker constructs messages about the target image only and *cannot* construct messages that entail information about differences or similarities between the target and the distractors. Messages and symbols have no a priori meaning but are assumed to obtain meaning and become grounded during the game. Once meaningful, the symbols are ideally combined in a structured manner to create compositional messages that express more complex meanings.[1]

[1] All code, materials, and data are available on OSF: `https://osf.io/9drb5/`.

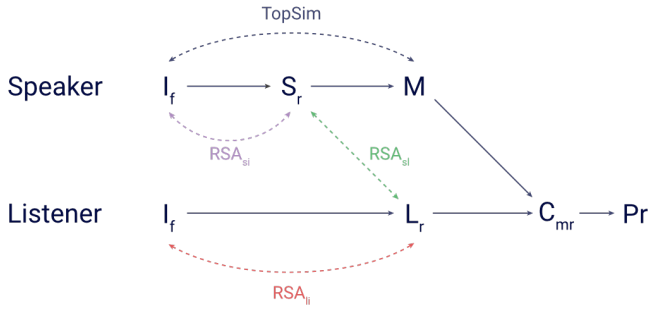


Figure 5.1: An overview of the setup used and the components that are used to calculate our metrics. I_f denotes the image features of DinoV2. S_r and L_r denote the speaker and listener representations of I_f. M is the message, C_{mr} the multimodal representation, and Pr is the probability of an image I_f belonging to message M.

5.4.1 Agents

Agents contain a language and a vision module. The latter consists of a frozen pre-trained visual network (DinoV2) and a learned agent-specific representation layer. While it is difficult to know what conceptual image features are present in DinoV2 embeddings, they have demonstrated capability in semantic segmentation tasks (Oquab et al., 2024), which is similar to the agents' objective. In contrast to the hybrid structure of the vision module, the language module is entirely trained from scratch.

The speaker agent processes images by applying a linear transformation to the image embeddings i_f, followed by batch normalisation, to create its agent-specific image representation r_s (S_r in Figure 5.1). Its language module embeds this representation and passes it through a single-layer Gated Recurrent Unit (GRU; Cho et al., 2014) that spells out messages to describe the target image.

The listener receives the message and the distractor images. It encodes the message into an embedding using another single-cell GRU layer. To obtain an image representation r_l (L_r in Figure 5.1) for each image, the listener agent, like the speaker, applies a linear transformation and batch normalisation on the image embeddings. Finally, temperature-weighted (with a default temperature of 0.1) cosine scores construct a multi-modal representation C_{mr} between the image and message representation (Dessi et al., 2021), where a higher probability (Pr) should be assigned to the target image. The listeners' target distribution comprises the probability for each possible image. Figure 5.1 illustrates the communicative setup and the components used to calculate our metrics.

5.4.2 Optimisation

Communicative success ($\hat{t} = t$) is used to optimise the trainable parameters of both agents. The listener minimises cross-entropy (ce) loss using stochastic gradient descent, amounting to supervised learning. The ce loss is calculated over the listeners' target distribution and thereby provides a direct pressure for communicative success. During inference, the candidate image with the highest probability is chosen as the target $\hat{t}$. The gradients required to optimise the speaker are calculated using the `REINFORCE` (Williams, 1992) update rule as each generated symbol must be assigned a loss. Following standard practice (Rita et al., 2024), entropy regularisation (Mnih et al., 2016) is added to the loss to maintain exploration in message generation.

In addition to the conventional ce loss, we introduce an alignment loss ($ce + RSA$) that includes an alignment penalty term to enforce high inter-agent and image-agent alignment. The term

$$L_{RSA} = (1 - RSA_{sl}) + (1 - RSA_{si}) + (1 - RSA_{li})$$

is added to the ce loss with equal importance. We use TorchSort (Blondel et al., 2020) to calculate L_{RSA}, ensuring that the entire loss term is differentiable. Importantly, L_{RSA} is not influenced by communicative success and does not interact with the ce loss (Section B.2). Only adding RSA_{sl} to the ce loss is not sufficient as high inter-agent alignment can be achieved while *losing* image-agent alignment (see Section 5.3). As such, we also include RSA_{si} and RSA_{li} to ensure that the agents communicate about the content displayed in the images. Including RSA_{sl} entails that representational information is shared between the agents, thus differing from how humans interact. Yet, ranking the speaker and listener representations in calculating RSA_{sl} bears *some* resemblance to projecting beliefs upon the interpretations of the other communicative partner. The current solution should be seen as a step towards more grounded vocabularies prone to refinements such as cognitive plausibility. We train for 30 epochs regardless of the loss used. The hyperparameters (Table B.1) that yielded the best validation accuracy across 42 different communication channel capacities (Section B.1) were used for our findings.

5.4.3 Data

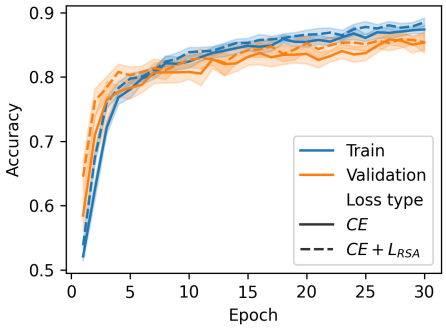
Agents are trained to discriminate MS COCO images but tested on three different datasets (Figure 5.2) to assess out-of-distribution (o.o.d.) performance.

MS COCO – We use a subset of 1200 images from the MS COCO 2017 validation set to train and test the agents using an 80/20 split. To obtain this subset, we first select the categories that contain more than 100 images (resulting in 12 categories) and subsequently sample 100 images for each supercategory present in the resulting set of images. Distractor images are sampled from the same category to ensure that there is *some* relevance to the target image. Sampling these images is done for each batch, meaning targets have different distractors at each epoch.

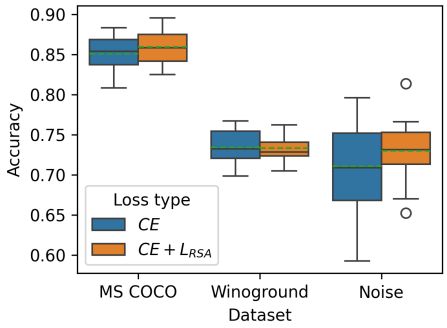
Figure 5.2: Exemplar pairs of each dataset used for evaluation. Left column: an image pair from MS COCO. Middle column: A Winoground example. Right column: A Gaussian noise pair. All images are cropped for display purposes.

Winoground – The Winoground dataset (Thrush et al., 2022) was created to assess the visio-linguistic compositional reasoning abilities of vision and language models. Here, we repurpose it as a proxy for the agents' ability to endow in compositional reasoning for image-based settings. The dataset contains 800 image-caption tuples, comprising 400 Winoground pairs. Image-caption pairs were included when the captions share the same words but are of different *compositions*, implying completely different semantics (e.g., "a tree smashed into a car" versus "a car smashed into a tree" in Figure 5.2 (middle)). As such, discriminating image-caption pairs requires the ability of vision and language models to use compositional language and to understand how language is manifested in the visual modality. Hence, it is posited that to successfully address this task, grounding in images and comprehension of compositional language is imperative. Here, we only use the image pairs, not the captions, and thus test whether RL agents can establish a communicative system that can describe concepts and their compositions. Crucially, this task differs from MS COCO in that the image pairs are *fixed*, *conceptually similar* and meant to be discriminative if the agents' language allows for compositional reasoning and is grounded in the visual modality.

Noise – Following Bouchacourt and Baroni (2018), we test whether agents can communicate about Gaussian noise ($\mu = 0, \sigma = 1$) image pairs when they are trained on real images. If this is the case, it would imply that messages relate to spurious instead of high-level concept features.



(a) The learning curves for the MS COCO dataset on train and validation data.



(b) Communicative performance (Accuracy) during inference on discriminating between two images of different datasets.

Figure 5.3: In (a) we see that the agents learn to communicate successfully without overfitting on the training data. In (b) we see that agents can discriminate MS COCO images but struggle with discriminating Winoground images. Line style indicates the loss type. Results are averaged over 15 seeds, areas indicate the 95% confidence intervals. Green dashed lines indicate averages.

5

5.4.4 Metrics

The performance of our agents is assessed through communicative success (*accuracy*) and the degree of representational alignment is measured using *RSA* (Section 5.3). The degree of compositionality in the emergent language is assessed through the commonly used *TopSim* metric. Formally, *TopSim* is the Spearman correlation between pairwise input distances and the corresponding message distances. As such, it is agnostic to which distance function is used. Input distance can, for example, be computed as attribute-value overlap (when the input space contains categorical attribute-value pairs), or as cosine distance (for continuous input vectors, as is the case in this chapter). The distance between messages is typically calculated as the minimum edit distance. The correlation between these sets of distances is taken as a tendency for messages with similar meanings to have a similar form. However, *TopSim* is relatively agnostic about how these messages are similar, as long as a minimum edit distance captures it. Other metrics for compositionality, such as positional disentanglement and bag-of-symbols disentanglement (Chaabouni et al., 2020), are not straightforward in this chapter due to the continuous nature of the input, i.e., the image embeddings.

5.5 Results

We now present our results, starting with the performance on three datasets, after which we revisit the alignment problem and investigate the relationship between alignment and *TopSim*.

We then show how the alignment penalty term affects communicative success, alignment, and *TopSim*.

5.5.1 Communicative success

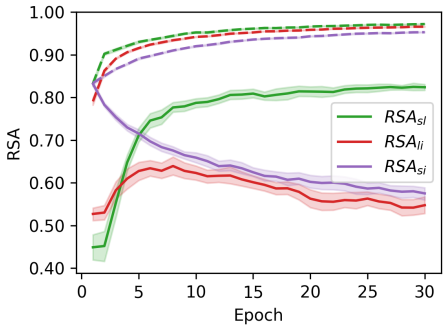
Unsurprisingly, results show that agents can successfully disambiguate between image pairs from MS COCO using an emergent language (Figure 5.3a). Notably, we also confirm previous observations by (Bouchacourt and Baroni, 2018) that agents trained on real images can communicate relatively well about Gaussian noise (Figure 5.3b). Since the speaker must construct its messages purely based on the target image, this suggests that the speaker uses spurious image features to do so. This finding, therefore, again suggests that the emerged languages convey information about spurious features rather than concept-level information. Interestingly, their performance on Gaussian noise is comparable to the performance on Winoground pairs, which requires the messages to capture concept-level properties. This reveals the difficulty of discriminating between strict pairs of conceptually similar images. The observed decrease in out-of-distribution performance aligns with findings from other studies, such as those presented by Lazaridou et al. (2018) and Conklin and Smith (2023) and highlights that generalisation to novel meanings is still difficult for our agents.

5.5.2 The alignment problem

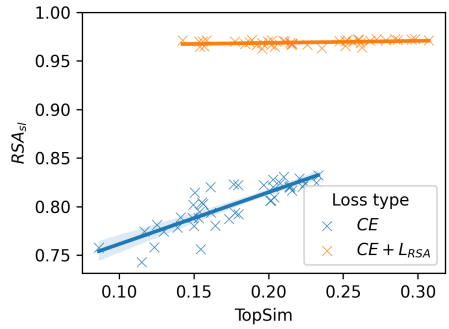
Considering the metrics used to assess representational alignment, the solid lines in Figure 5.4a clearly show that inter-agent alignment increases while alignment sensitivity to image features decreases for both agents. Again, it is in principle not a problem that the agents' image representations align, but it becomes problematic when the alignment between the image embeddings and the image representations declines. Ablations across different channel capacities (Section B.1) and with different pre-trained vision modules (Section B.3) showed that these trends appear consistently and are not influenced by the capacity or type of vision model. In addition to the communicative success on Gaussian noise, this re-confirms that the agents do not learn to extract concept-level information from the image embeddings but instead use the embeddings to solve this task differently.

5.5.3 *TopSim* and representational alignment

Earlier findings show mixed results on the relationship between *TopSim* and generalisation in image-based settings, *TopSim* was either positively related to generalisation (Chaabouni et al., 2022) or not (Rita et al., 2022b). Our results indicate that generalisation and *TopSim* are correlated with both ce ($r = .856$, $p < .001$) and $ce + RSA$ ($r = .767$, $p < .001$) losses. This suggests that more structured languages, as measured using *TopSim*, enable better communication on unseen validation pairs. Moreover, we find a strong positive relationship between RSA_{sl} and *TopSim*



(a) The representational alignment curves for inter-agent image representations (green) and between the image and the sender/listener representations (purple, red).



(b) The relationship between *TopSim* and inter-agent alignment (RSA_{sl}) for both loss types.

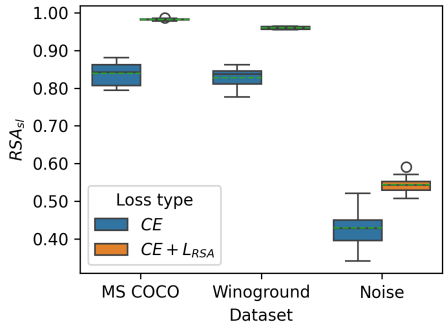
Figure 5.4: In (a) we see that the alignment problem occurs with the ce (solid lines) but not the $ce + RSA$ (dashed lines) loss. In (b) we see that *TopSim* and RSA_{sl} are correlated when the ce loss is used ($r = .838$, $p < .001$). This is also the case with the $ce + RSA$ loss ($r = .408$, $p = .001$) but the effect is decoupled from *TopSim*. Results are averaged over 15 seeds, areas indicate the 95% confidence intervals.

($r = .838$, $p < .001$) in the ce setup (Figure 5.4b). While this relation is also present in the ce+RSA setup ($r = .408$, $p = .001$), it is decoupled from *TopSim* given the (very) small spread ($\sigma = .003$) of RSA_{sl}. Although representational alignment may alleviate the need for discriminative messages, we do not observe an influence of inter-agent alignment on the number of uniquely produced messages.

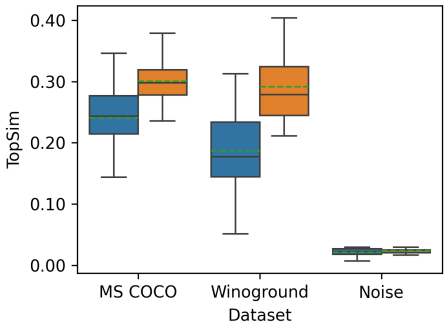
5.5.4 Mitigating the alignment problem

We now focus on the $ce + RSA$ setup, which was introduced to ensure that the agents maintain alignment with the image embeddings. Figure 5.4a and Figure 5.5a show that this indeed happens: inter-agent alignment *and* agent-image alignment increase during training and remain high during inference. Yet, there does not seem to be a benefit for communicative success at inference time as *accuracy* across the datasets remains relatively similar (Figure 5.3b). This is likely because the alignment penalty only forces agents to represent images similarly to the image embeddings and acts independently from the cross-entropy loss used to assess the success of communication (Section B.2). In the case of images containing Gaussian noise, we still observe above-chance performance, which suggests that communication between the agents still occurs in an artificial manner.

In addition to increased representational alignment between agents, the alignment penalty also leads to increased *TopSim*, which suggests that the messages used during communica-



(a) Inter-agent representational alignment (RSA_{sl}) across different datasets.



(b) Topographic similarity (*TopSim*) between the images and the messages displayed for different datasets.

Figure 5.5: In (a) we see the effect of the loss function on the degree of inter-agent representational alignment. In (b) we see that *TopSim* increases as a result of the $ce + RSA$ loss.

tion have a higher degree of structure (Figure 5.5b). Given the higher values of RSA_{sl}, this strengthens our finding that *TopSim* and inter-agent alignment are related. This suggests that the observed variations in *TopSim*, whether higher or lower, as noted in previous studies (e.g. Kottur et al., 2017; Chaabouni et al., 2020), should not be interpreted without considering representational alignment since they may be attributable to this underlying artefact rather than alterations to the original setup.

When tested on more strict Winoground pairs, communicative success does not improve as a result of using the alignment penalty (Figure 5.3b). Given the correlation between *TopSim* and generalisation that was observed earlier, this is surprising since the higher degree of *TopSim* should imply that the language is more structured. Moreover, both, RSA_{si} and RSA_{li} have not drifted away from the image features (Figure 5.4a). This combination, *in theory*, should be ideal for discriminating image pairs from the Winoground dataset since it was designed to be discriminative with compositional visio-linguistic reasoning. However, in *practice* this is not the case.

5.6 Discussion

In this chapter, we revisited the representational alignment problem in a common setup used in emergent communication and proposed a solution to this underrepresented problem. We corroborated earlier findings by demonstrating that agents align their image representations and rely on spurious image features instead of human-like concept-level information (Bouchacourt and Baroni, 2018). We then showed that inter-agent alignment strongly correlated with the

commonly used *TopSim* metric. Our solution to the alignment problem involves an alignment penalty that forces the agents to remain aligned with the input features, thereby mitigating the alignment problem without compromising communicative success. Finally, when agents are tested on more challenging Winoground pairs, we observed reasonable but lower performance regardless of whether image representations were similar to the image embeddings or not. With this work, we hope that the alignment problem will receive more attention in the field of emergent communication, as is already the case in adjacent fields (Sucholutsky et al., 2023).

5.6.1 Importance of representational alignment

It is common practice in simulations of emergent communication to process (visual) inputs into an agent-specific hidden representation and update their weights simultaneously (e.g. Lazaridou et al., 2017; Bouchacourt and Baroni, 2018; Chaabouni et al., 2019a, 2020; Rita et al., 2022b). As such, inter-agent alignment, *irrespective of the input form*, likely happens in other simulations too. This phenomenon is therefore potentially widespread and can perhaps be the cause for findings that are at odds with experimental findings. This bears much similarity to a concept known as shortcut learning: a form of understanding that is in many ways not human-like, but introduces a new "alien" kind of problem-solving (Schwartz and Stanovsky, 2022; Mitchell and Krakauer, 2023). While it is not always the case that the representation structure we *expect* to help solve a task will do so (e.g. Montero et al., 2021; Xu et al., 2022), such discrepancies may hinder the use of emergent communication models in developing a more natural understanding of human languages and leave them less suitable for directly simulating language evolution phenomena. Especially so if we want machine representations of natural language to align with human representations (Sucholutsky et al., 2023). *RSA* should therefore be used to rule out, or at the bare minimum report about, representational alignment in the future.

5.6.2 Relating *TopSim* and representational alignment

Measuring representational alignment using *RSA* is similar to how *TopSim* measures the structure in messages. While they differ in their inputs, they both calculate the Spearman-ranked correlation between metric-agnostic pairwise distances. Crucially, the input makes all the difference; the inputs for *RSA* are from both agents and are trained independently, whilst *TopSim* only assesses the relation between the fixed inputs and learned output (Figure 5.1). Despite the similarities, the metrics thus describe different phenomena and are rarely reported simultaneously.

We hypothesise that the relationship between *TopSim* and inter-agent representational alignment is a by-product of the setup, which in essence implies that the listener has to align its representation r_l to the speaker representation r_s (Rita et al., 2022b). It has to do so using only the speakers' messages, being a compressed abstraction of r_s. A possible solution to

this problem is to align representations, which eases the listeners' training objective. If the speaker consistently produces structured messages during training, aligning r_l with r_s becomes easier, thereby leading to higher inter-agent alignment. Essentially, this renders *TopSim* to be an *indirect* metric for the rate of alignment, for which RSA_{sl} is a *direct* metric. In the context of learnability, the relationship between *TopSim* and inter-agent alignment and the fact that alignment always occurs can be seen as reasons for why languages with higher *TopSim* are easier to learn (Li and Bowling, 2019; Cheng et al., 2023). This underscores the need to report inter-agent representational alignment to avoid conclusions drawn about the effect of specific interventions on *TopSim* which may be attributable to inter-agent alignment.

5.6.3 Targeted o.o.d. evaluations

An important implication of our findings concerns the standard practice of reporting o.o.d. accuracy where the agents are tested on unseen input after training (e.g. Auersperger and Pecina, 2022; Conklin and Smith, 2023). In essence, doing so should inform us about the agents' ability to generalise from one dataset (e.g., MS COCO) to another dataset (e.g., the Winoground pairs), much like human language allows us to talk about an infinite number of situations. Crucially, this overlooks the representational alignment problem in that we do not know *what* the agents are precisely generalising about. This problem can be mitigated using the alignment penalty term to assess generalisation more directly, or at least should be taken into consideration.

We assessed o.o.d. performance on the more challenging Winoground pairs as a proxy for the agents' ability to endow in compositional reasoning for image-based settings. Good performance on the Winoground dataset requires a grounded language that can be used to create compositional messages since the objects and their underlying relations need to be described. In general, we suggest starting to evaluate simulations of referential games on targeted, strict tasks, such as probing state-of-the-art vision language models on, for example, visio-compositional (Thrush et al., 2022; Diwan et al., 2022; Hsieh et al., 2023; Ray et al., 2023) or spatial (Kamath et al., 2023) reasoning tasks. Re-purposing such datasets can reveal more directly whether agents develop the attested communicative abilities that are trivial to humans without having to rely solely on metrics. Our results illustrate this through a shortcoming of the *TopSim* metric. We observed that agents still struggle with distinguishing pairs of *conceptually similar* Winoground images, even though *TopSim* is higher with the alignment penalty. If the language protocol were to communicate concept-level information *and* compositional messages were created, we should not observe this struggle, meaning that the emerged protocols do not enable human-like communicative success.

Interestingly, the o.o.d. performance remains substantially above chance in the $ce + RSA$ setting. Given that MS COCO is not a dataset for learning to model compositionality, this delineates the limits of what can be achieved qua performance based on MS COCO image features in the Winoground context. Nevertheless, this leaves open the question of the above-

chance performance on Gaussian noise with the $ce + RSA$ loss. A tentative explanation is that the higher inter-agent alignment on noise input ($M_{ce} = .428$, $M_{ce+RSA_{sl}} = .543$, $t = -8.71$, $p < .001$) alleviates part of the problem (Figure 5.5a). To validate this, future experiments should involve controlling the prior distributions of the agents' image encoders by training their vision modules on different data. Doing so ensures that they have to communicate about novel objects and cannot rely on similar representations.

5.7 Conclusion

This chapter revisited the underrepresented alignment problem present in the referential game often used in simulations of emergent communication. Specifically, we focused on the problem of increasing alignment between agent-image representations in combination with a decreasing alignment between the input and agent representations. We first confirmed that agents align their image representations while losing connection to their input, meaning that the emergent languages do not appear to encode human-like visual features. We then showed that, in the common setup, inter-agent alignment is related to topographic similarity, and argued that this renders *TopSim* an *indirect* metric of the rate of inter-agent alignment. To further investigate the effects of alignment, we introduced an alignment penalty to mitigate the alignment problem. We showed that the communicative ability on a strict compositionality benchmark did not improve, leaving the question of inducing compositional generalisation in emergent communication for images unsolved. Our findings underscore the need to better understand the divergence between human and artificial language emergence within the prevalent referential setup and highlight the importance and potential impact of representational alignment. We hope that future work rules out or at least reports about representational alignment.

5.8 Limitations

Our work has a few notable limitations. First, it only involves the referential game. Another popular variant, the reconstruction game (e.g. Chaabouni et al., 2019a, 2020; Lian et al., 2021; Conklin and Smith, 2023), requires the listener to reconstruct the input object based on the speaker's message. Since this setup has a different objective and presents different learning biases, it may have different results. We still expect the results to be similar as there is no pressure to retain alignment between the image input and agent representation. It would, however, be interesting to investigate whether the language protocol in this scenario is more structured than in the referential game.

Another limitation in our setup is that we only consider the scenario with two agents, which may be a requirement for alignment to be possible. Since experiments with human participants show that larger communities create more systematic languages (Raviv et al., 2019b), simulations

on emergent multi-agent communication with populations of agents are also conducted, but these yield mixed results. The emergent communication protocols oftentimes do not evolve to be more structured unless explicit pressures such as population diversity or emulation mechanisms are introduced (Rita et al., 2022b; Chaabouni et al., 2022). However, Michel et al. (2023) showed that population setups can result in more compositional languages if agent pairs are trained in a partitioned manner to prevent co-adaptation. Despite the mixed results, we believe that emergent communication with populations of agents is ecologically more valid and could result in different alignment effects. Much like how Tieleman et al. (2019) showed that autoencoders encode better concept category representations when they learn representations in a community-based setting with multiple encoders and decoders collectively.

The final limitation of our study regards its scale. While simulations of emergent communication are typically conducted on relatively small-scale datasets, human language emergence is accompanied by rich and diverse multimodal experiences. Recent results in the field of computer vision suggest that dataset diversity and scale are the primary drivers of alignment to human representations (Conwell et al., 2023; Muttenthaler et al., 2023). As such, this key difference between the setting of artificial emergent communication and human language emergence can drive the observed differences in representations. Due to the difficulty of interpreting these representations, we see this as another reason to evaluate emergent protocols on more strict datasets with clear pragmatic value for humans.

6

Searching for Structure

Human languages have evolved to be structured through repeated language learning and use. These processes introduce biases that operate during language acquisition and shape linguistic systems toward communicative efficiency. In this chapter, we investigate whether the same happens if artificial languages are optimised for the implicit biases of Large Language Models (LLMs). To this end, we simulate a classical referential game in which LLMs learn and use artificial languages. Our results show that initially unstructured holistic languages are indeed shaped to have some structural properties that allow two LLM agents to communicate successfully. Similar to observations in human experiments, generational transmission increases the learnability of languages, but can at the same time result in non-humanlike degenerate vocabularies. Taken together, this work extends experimental findings, shows that LLMs can be used as tools in simulations of language evolution, and opens possibilities for future human-machine experiments in this field.

Originally published as: Tom Kouwenhoven, Max Peeperkorn, Tessa Verhoef. 2025. Searching for Structure: Investigating Emergent Communication with Large Language Models. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Di Eugenio, B., Schockaert. S., editors, *In Proceedings of the 31st International Conference on Computational Linguistics*, pages 9977–9991, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

6.1 Introduction

Vocabularies of signals enable us to communicate about meanings, but to express an arbitrary number of meanings, vocabularies would require an equally large set of words as there are meanings, and learning such holistic vocabularies is cognitively challenging. Human languages, therefore, typically show some form of compositional structure, where meaningful signal-meaning mappings can be composed such that the combination of individual meaningful signals can express more than the meaning of the individual components alone (Hockett, 1960). An important finding in the field of language evolution is that such structural properties can emerge at the population level as a result of individual learning biases and pressures that continuously shape the languages on a longer timescale, often eventually resulting in languages that are easier to learn and exhibit some degree of structure (Smith, 2022).

The processes involved in the evolution of language have been extensively investigated through experiments and simulations. The latter typically use hard-coded agents with inductive biases (de Boer, 2006), Bayesian learners (e.g. Griffiths and Kalish, 2007b; Culbertson and Smolensky, 2012; Kirby et al., 2015), or reinforcement learning agents (Lazaridou and Baroni, 2020) to investigate the evolution of structured languages. In contrast, we investigate whether more flexible LLMs as relatively unbiased language learners (Wilcox et al., 2023) are appropriate tools to study how languages evolve. While their internal mechanisms are fundamentally different from those of humans, they still are the first close flexible comparators of human language users, which can be used as tools to answer cognitive and typological investigations (Warstadt and Bowman, 2022; van Dijk et al., 2023a). Given that languages are shaped by the biases and pressures of individual language learners, which differ for LLMs (e.g., fewer memory constraints), we are interested in identifying similarities and differences between humans and LLMs on specific language evolution-oriented tasks.

Our work largely follows the experimental design by Kirby et al. (2015) in which Bayesian learners and humans learn an artificial language to communicate in a referential game. They find that linguistic structure arises from a trade-off between pressures for compressibility and expressivity. This chapter extends their work by using LLMs as objects of investigation. Specifically, we investigate how artificial languages evolve when two LLMs communicate in a referential game and what the effects of generational transmission on these languages are. We compare the properties of these languages to those that are found in experiments involving humans. Results show that 1) LLMs can learn artificial languages and use them to communicate successfully, 2) the languages exhibit higher degrees of structure after multiple communication rounds, 3) LLMs generalise in more systematic ways when the evolved language is more structured, and 4) languages adapt, although not necessarily in a human-like way, and become easier to learn by the LLMs as a result of generational transmission.

6.2 Background & Related work

6.2.1 The evolution of structure

Learning novel signal-meaning mappings, and the emergence of rules that can combine these signals into structured languages have been abundantly investigated in the field of language evolution using human experiments (Kirby et al., 2008; Galantucci, 2005; Scott-Phillips et al., 2009; Verhoef, 2012; Raviv et al., 2019a,b) and computational simulations (de Boer, 2006; Steels and Loetzsch, 2012; Lazaridou and Baroni, 2020). These typically follow a setup where success depends on cooperation between two or more participants/agents in a Lewis game. Here, players are prevented from communicating using conventional communicative means and instead must establish novel communication systems through repeated cooperation. Outcomes often show that players, human or machine, quickly establish novel signal-meaning mappings that enable them to communicate successfully. However, recent computational simulations using reinforcement learning agents often develop communicative systems different from those of humans (Galke et al., 2022)[1] unless specific key pressures are introduced to recover initially absent human patterns (Galke and Raviv, 2025).

It has been suggested that seemingly arbitrary aspects of linguistic structure may result from general learning and processing biases deriving from the structure of thought processes, perceptuo-motor factors, cognitive limitations, and pragmatics (Christiansen and Chater, 2008). A well-investigated cause for this phenomenon is the process of cumulative cultural evolution (Boyd et al., 1996; Tomasello, 1999), which is typically investigated using iterated learning experiments (Kirby et al., 2008). Here, information (e.g., a language) is repeatedly passed down from one generation to the next, where the information is modified and improved upon within each generation. The influential work by Kirby et al. (2008, 2015) demonstrated that when human individuals learned an artificial language previously learned by another individual, the language became easier to learn and displayed a higher degree of structure. Crucially, these results are mostly attributed to the fact that the language repeatedly goes through a learning bottleneck, in which individual cognitive constraints, such as memory constraints, gradually shape the language. Iterated learning has been used to demonstrate that structure emerges in various setups with, for example, continuous signals (Verhoef, 2012) or continuous meaning spaces (Carr et al., 2017), and it is argued that this process may have led to the statistical Zipfian structure of language (Arnon and Kirby, 2024). Yet, Raviv et al. (2019a) showed that structure can also emerge *without* generational transmission. In this case, a pressure for compressibility originating from communication with multiple interaction partners and expanding meaning spaces causes languages to become compositional. This effect is even more prominent if the

[1] But see Lian et al. (2023b, 2024); Zhang et al. (2024b) for recent work showing that the need to be understood (i.e. communicative success), noise, context sensitivity, and incremental sentence processing help induce human-like patterns of dependency length minimisation in reinforcement learning agents.

number of interaction partners is larger (Raviv et al., 2019b). The current chapter is inspired by the traditional methods described previously and extends them with our current most sophisticated models of natural language.

6.2.2 LLMs as models of language

LLMs are sophisticated models of natural languages, and growing evidence shows their ability to exhibit 'average' human behaviours. It is, for example, suggested that LLMs can model human moral judgements (Dillion et al., 2023) and transmission chain experiments revealed human-like content biases in GPT-3.5 (Acerbi and Stubbersfield, 2023). When LLMs are extended with records of experiences, Park et al. (2023) showed that groups of generative agents exhibit believable human-like individual and emergent social behaviours when they interact over extended periods. It is even suggested that human-LLM interactions in everyday life can potentially mediate human cultures through their influence on cultural evolutionary processes of variation, transmission and selection (Brinkmann et al., 2023; Yiu et al., 2024).

While previous work has investigated human-like behaviour at inference time, findings from cognitive science can also be used to improve model performance. Iterated learning can, for example, be incorporated into the training regime to extrapolate desirable behaviours. Zheng et al. (2024) have likewise shown that representations are easier to learn when vision-language contrastive learning is reframed as the Lewis signalling game between a vision agent and a language agent, ultimately improving compositional reasoning in vision-language models. However, this does not guarantee model improvements. Shumailov et al. (2024) have shown that LLMs, autoencoders and Gaussian mixture models drift when trained repeatedly on AI-generated data. In these cases, crucially, the generated content is slowly optimised to be understandable for models, *not* for humans, resulting in what they call model collapse. The authors therefore argue that genuine human interactions with systems will be increasingly important to prevent model collapse. While drift is often seen as an unwanted effect of unsupervised training, this is not surprising from a language evolution viewpoint since languages adapt to how they are learned and used (Smith, 2022). It was therefore suggested in Chapter 1 that languages should adapt to become more natural for humans *and* machines. This bears much resemblance to the idea that findings from cognitive science can prevent modal collapse (Smith et al., 2024) or inform modelling choices (Galke and Raviv, 2025). Here, we view LLMs from this evolutionary perspective.

Although inductive biases inherent to a language model's (pre-)training objectives (i.e. the cloze task and instruction tuning) and memory constraints are very different from those in humans, recent work has shown that GPT-2 models struggle to learn languages that contain unnatural word orders, lack hierarchical structure, or lack information locality (Kallini et al., 2024). This suggests that, even though the language processing mechanisms in Transformers are non-humanlike, LLMs exhibit a preference for structured languages similar to those of

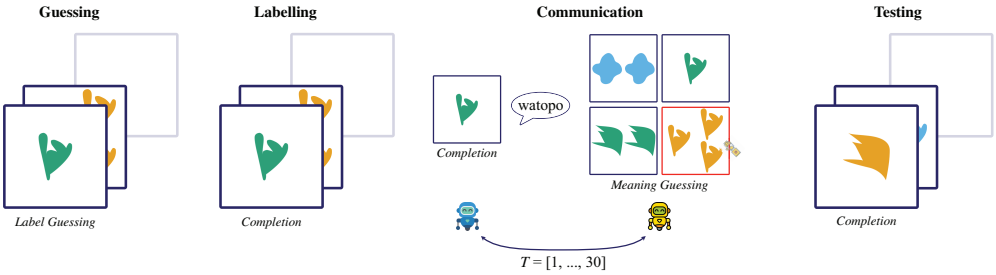


Figure 6.1: A graphical representation of the experimental blocks. The agents first go through a guessing block before labelling each of the 15 training stimuli in the labelling block. The communication block is done for 4 rounds, each consisting of 30 tasks T, where the agents alternate speaker-listener roles to be a speaker and listener for each stimulus once. Finally, the agents label 27 (15 original and 12 novel) stimuli in the testing block. Icons obtained from flaticon.com.

humans. Moreover, in an artificial language learning experiment similar to the work presented here, Galke et al. (2024) showed that compositional structure is advantageous for GPT-3 when learning an artificial language and that a higher degree of compositional structure also resulted in human-like generalisation for new unseen items. This chapter is different in that Galke et al. tested the ability of GPT-3 to learn languages that evolved during a *human* experiment (Raviv et al., 2019b, 2021), thus being optimised for human learners. We instead wish to investigate what kinds of languages evolve when they are optimised for *LLMs*.

6.3 Methodology

Our methodology is inspired by Kirby et al. (2015) and (Raviv et al., 2021). The complete simulation set-up consists of four blocks: guessing, labelling, communication and testing (Section 6.3.2 & Figure 6.1[2]). The agents perform the guessing, labelling, and testing block separately, but the communication block is interactive. The communication block is a classic referential game in which two agents communicate to discriminate a target stimulus from four distractor stimuli. They do so in four rounds, each consisting of 30 interactions T, alternating speaker-listener roles between interactions. In a single interaction round, the speaker observes a target stimulus (not the distractors) and utters a signal that describes the current stimulus. Using this signal, the listener must discriminate the correct target amongst a set of distractor stimuli. Cooperation is successful when the listener's guess is the target stimulus.[3]

[2]This is for illustration purposes only, we stress that our simulations are entirely run in the textual modality only to avoid the additional challenge of extracting relevant visual features and mapping these to artificial languages.

[3]All code, materials, and data are available on OSF: `https://osf.io/52yar/`.

6.3.1 Stimuli and initial languages

The meaning space consists of stimuli with three attributes. They have one of three shapes, one of three colours, and can appear in groups of one, two, or three shapes, creating 27 distinct stimuli. Initial signals for these stimuli were generated before each experiment according to the method used by Kirby et al. (2008). The signals are concatenations of 2, 3, or 4 randomly selected consonant-vowel (CV) syllables resulting in artificial non-existing signals (e.g., watopo, nafa, nomomeme). The CV syllables consist of one of eight consonants g, h, k, l, m, n, p, w and one of five vowels a, e, i, o, u. Out of 27 stimuli, only 15 stimuli are used during the guessing, labelling, and communication blocks. All 27 stimuli are used in the testing block such that we can assess whether the agents can generalise to novel stimuli. The training stimuli are selected randomly before each simulation, but we ensure that each attribute value is represented equally often across this set.

6.3.2 Simulation blocks

Each simulation consists of four blocks. In the first block, we assess whether agents can correctly guess a signal when presented with a stimulus. Second, in the labelling block, an agent repeatedly produces a signal for each stimulus given the initial training vocabulary. The signals generated in this block are taken as the learned vocabulary for that agent. In the third block, the agents communicate as described before, taking turns as speaker and listener until all rounds are completed and each stimulus appears twice per round (i.e., both agents produce a signal for each stimulus and make a guess for each stimulus). In this block, the interaction between the agents gradually alters each agent's individual vocabulary, much like how this is done in earlier simulations (De Boer, 2000; Steels and Loetzsch, 2012). Specifically, we update the current stimulus to be associated with the signal that is produced. After the communication block, the testing block tasks the agents to generate signals for the entire meaning space of 27 stimuli using the training vocabulary that was optimised in the labelling and communication block. Hence, they must generalise their strategies to unseen samples.

6.3.3 LLMs as agents

The LLMs used in our experiment were instruction-tuned instantiations of Llama-3 70B (Llama Team, 2024) with greedy sampling.[4] Since our method required LLM agents to follow instructions, we did not consider base models. In particular, we instructed them about the nature of their task and its collaborative goal. Though instruction-tuning using reinforcement learning from human feedback (RLHF) may influence the probabilities of some tokens fitting to instruction-following behaviour, the capacity to produce fluent language and knowledge

[4] Although we only report results on one model type, initial explorations with GPT-3.5 and Llama 2 7B showed similar behaviours to LLama-3 70B.

```
{'shape':3,'colour':'blue','amount':1,'word':'ninikonu'}
{'shape':1,'colour':'green','amount':3,'word':'hanosa'}
      ⋮
{'shape':2,'colour':'orange','amount':2,'word':'sanu'}
{'shape':1,'colour':'green','amount':3,'word':'[COMPLETE]
```

Prompt 6.3.1: A vocabulary snippet as used in a completion prompt. The complete prompts are visible in Section C.2.

is mostly acquired in the pre-training phase (Zhou et al., 2023; Lin et al., 2024). Moreover, since our method does not specifically tap into instruction-tuning behaviour, we do not expect much variance in the results should we use base models only. While human participants typically learn signal-meaning mappings through a learning block, we use LLMs' in-context learning ability (Brown et al., 2020) to teach them the languages. Specifically, we prepend our prompts with the items to be learned in a structured JSON-like format (Prompt 6.3.1). Given the observed behavioural similarities between humans and LLMs (Galke et al., 2024), we assume that a vocabulary of signal-meaning mappings in the context of a prompt provides enough (distributional) information for a LLM to learn an appropriate mapping between the attributes of the stimuli and signal syllables. Although the prompt structure 'invites' the LLM to infer a signal from the stimulus attributes, we are agnostic about how exactly and what kind of mapping the LLM deduces, but we are interested in the resulting behaviours.

Throughout a simulation, agents essentially perform one of two tasks: generation or guessing. The labelling block and speaking in the communication block involve generating signals. The guessing block and discrimination in the communication block involve guessing. The prompts for these tasks are extensions of those used by Galke et al. (2024), with slight adaptations to enable LLMs to discriminate between stimuli. Given that LLMs show a primacy and recency bias (Liu et al., 2024), the vocabulary is shuffled before each task such that ordering effects are minimal. System instructions depend on the task performed, but are largely similar and chosen to be as close as possible to instructions given to humans in experimental settings.

Generating signals. For signal generation in the labelling block, we use prompt completion (Prompt C.2.1). During labelling, the agents see the *entire* training set and generate a signal for each stimulus, effectively amounting to a look-up task since the stimulus is present in the prompt. On the other hand, the vocabulary presented to agents during communication and testing does *not* include the current stimulus, thus requiring the agents to extract an appropriate mapping and generalise to new stimuli (Prompt C.2.2). A human-like solution would be to map stimulus attributes (i.e. shape, colour, and amount) to syllables representing these attributes and create compositions that describe the stimulus. During communication, we incentivise the agents to communicate using a `communicativeSuccess` attribute which is set to 1 if the previous interaction for this stimulus was successful and zero otherwise. Adding this

attribute functions as a memory between interactions and provides a pressure for expressivity. It is hypothesised that the latter plays an important role in human language evolution since it prevents languages from becoming degenerate (Smith et al., 2013). Importantly, during testing, the vocabulary presented to the agents always includes the stimuli present in the train set (without the current stimulus), and stimuli from the test set are never present.

Guessing signals or meanings. For guessing and discrimination during communication, the agents need to respond with a choice corresponding to the speaker's signal. Unfortunately, LLMs are inconsistent and unreliable in answering multiple-choice questions (Khatun and Brown, 2024). In our initial exploration, this indeed proved to be unusable. Instead, for each distractor (signal or meaning), we run the prompt prefilled with that distractor through the model and select the distractor with the highest probability (Prompt C.2.3). Again, the agents observe the training vocabulary *with* the current stimulus in the guessing block. In the communication block, agents observe the training vocabulary *without* the current stimulus.

6.3.4 Metrics

We are firstly interested in investigating whether two agents settle on a language that enables them to communicate, measured by the percentage of successful interactions (*PercCom*) in a round. We use multiple metrics to measure structure in messages. The most common metric is topographic similarity (*TopSim*, Brighton and Kirby, 2006). Similar to Kirby et al. (2008), we report Z-scores of the Mantel test (Mantel, 1967) between signal similarities (normalised Levenshtein distance) and semantic similarities (the number of equal attributes between two meanings). A communication system with a high *TopSim* uses similar signals for similar meanings. We compute the *Ngram* diversity (Meister et al., 2023), being the average fraction of unique vs. total *Ngrams* for $N \in \{1, 2, 3, 4, 5\}$ in all produced signals. Low *Ngram* diversity across all signals implies the agents re-use parts of signals in different signals, hinting at compositional signals when it happens in combination with increased *TopSim*. We assess the degree of signal systematicity between the signals produced for unseen stimuli in the test block and the previous stimuli in the communication block using the generalisation score (*GenScore*, Raviv et al., 2021). Here, we first compute the pairwise semantic difference between each stimulus in the train and test scenes, followed by the pairwise normalised edit distance between the signals produced for these scenes. We then take the Pearson correlation between these differences across all stimuli. Intuitively, this measures whether similar scenes across both sets are similarly labelled, thereby suggesting generalisation.

6.4 Evaluation

We ran 15 simulations, each initialised with a random seed and unique artificial, unstructured, and holistic language.[5] Metrics were computed for each block, except for the generalisation score, which is only computed for the testing block. A human-like result would show increasingly successful interactions and increasing *TopSim* scores, while *Ngram* diversity should go down. If this is the case, we expect to observe higher generalisation scores since agents can compose new signals according to a learned structured strategy. We use linear mixed effects models to analyse the results of the communication block and to account for the random effects of each simulation's vocabulary. The slope ($\hat{\beta}$) determines the direction of the effect and the rate of change. Additionally, we use conditional R^2 (Nakagawa and Schielzeth, 2013), denoted by R^2_c, which considers fixed and random effects, to show how much variance can be explained by the model. Higher values of R^2_c indicate that the model captures more variance and that correlations are stronger. Finally, we report the marginal R^2_m, which is the variance explained by the fixed effects.

6.5 Results

6.5.1 Learning the artificial languages

We first assess whether LLMs were able to learn the initially unstructured languages. Given the nature of the guessing task, which is essentially a lookup task, unsurprisingly, LLMs were able to guess the correct signals for the stimuli almost perfectly ($M = .973, SD = .031$). However, labelling the same stimuli via completion proved much more difficult ($M = .453, SD = .152$) despite the presence of the correct signal in the prompt. This contrast is in line with work showing that LLM predictions are sensitive to task instructions and how predictions are extracted (Weber et al., 2023; Hu and Levy, 2023; Hu and Frank, 2024). Additionally, it corroborates the use of prefilled options in our guessing prompts during communication. Nevertheless, this performance is still better than that of humans[6] and is not unimpressive given the vast number of possible signals that can be produced. Finally, the expected struggle to correctly reproduce (i.e., learn) unstructured signals introduces some welcome variation to the agents' vocabulary, which is used at the start of the communication block.

[5] We are aware of the fragile nature of behavioural experiments with LLMs. Small perturbations to prompts can have large effects on the outcome (e.g. Weber et al., 2023; Hu and Levy, 2023; Hu and Frank, 2024; Giulianelli et al., 2024). This is also the case in our experiment. To ensure the reproducibility of the current findings, we use an open-source model, share all prompts, log probabilities, and data on OSF. Nonetheless, the probabilistic nature of LLMs will always warrant further investigation.

[6] In Chapter 7 we conduct an experiment involving humans and show that the guessing block is much easier than the labelling block.

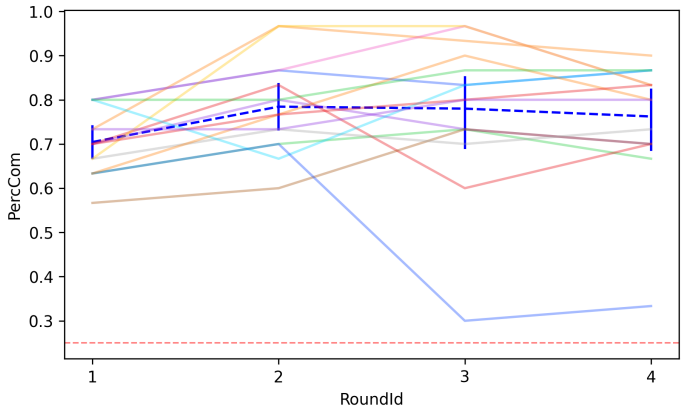


Figure 6.2: The communicative success (*PercCom*) over the communication rounds. Each coloured line indicates a simulation, and the dashed blue line displays the average with bars indicating the 95% confidence interval. The dashed red line delineates chance performance.

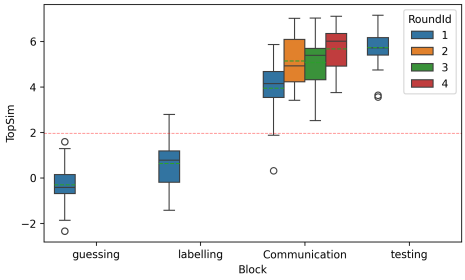
6.5.2 Agents communicate successfully

Once the agents have individually learned the vocabulary, they start communicating. Despite initially starting with different languages, approximately 70% of the interactions in the first round are successful (chance performance would amount to 25%). This increases somewhat in the following rounds to $\approx 75\%$, but not significantly (Figure 6.2). Interestingly, communicative success is not guaranteed; it fluctuates between rounds and can even decrease drastically in some simulations.

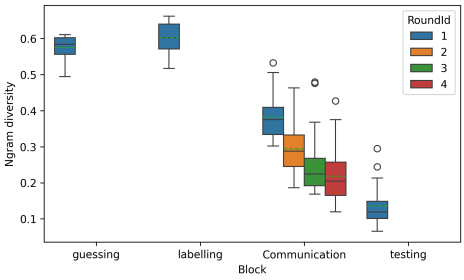
6.5.3 Communication results in structure

Although the initial languages are unstructured, some form of structure emerges due to repeated learning and use (Figure 6.3). This mostly happens during the communication block where *TopSim* increases significantly across rounds ($\hat{\beta} = .508 \pm .073, R_c^2 = .579, R_m^2 = .355, p < .001$) and *Ngram* decreases across rounds ($\hat{\beta} = -.054 \pm .004, R_c^2 = .812, R_m^2 = .558, p < .001$). This increase in structure benefits communicative success positively ($\hat{\beta} = .035 \pm .007, R_c^2 = .769, R_m^2 = .427, p < .001$). However, we also observe behaviour that is not human-like; the signals used to communicate become longer over the rounds ($\hat{\beta} = .557 \pm .044, R_c^2 = .919, R_m^2 = .505, p < .001$). This contradicts what is observed in human experiments, where we typically observe that messages become shorter and lie close to a theoretical frontier balancing expressivity and simplicity (Piantadosi et al., 2011; Kirby et al., 2015).

These results extend the findings of Galke et al. (2024) in that LLMs not only *learn* structured vocabularies better but also naturally *shape* languages to have some form of structure when they are optimised for their inherent preferences. In addition to the fact that LLMs struggle to learn



(a) TopSim scores over the agent's vocabulary in each block and round. The dashed red line indicates the $p < .05$ level.



(b) Ngram diversity scores over the agent's vocabulary in each block and round.

Figure 6.3: Communication clearly increases the structure of the vocabularies, as seen by the increasing *TopSim* scores and decreasing *Ngram* diversity.

impossible languages (Kallini et al., 2024), that reframing prompt instructions into a structured list improves the model response (Mishra et al., 2022), and given that we do not impose pressure to induce structure, the surprising outcome of our experiments may be the result of an apparent "structure bias" in LLMs.

6.5.4 Structure enables better generalisation

After the communication block, the agents engage in the final simulation block. Here, they generate signals for all 27 stimuli using the vocabulary that has evolved after learning and communication. We find that high *TopSim* languages allow for better generalisation ($r = 0.735, p < .001$, Figure 6.4). A qualitative inspection of the signals generated in the testing block of the simulation, which resulted in the highest *TopSim* after communication, reveals that this agent repeatedly re-uses parts of signals in different compositions (Table 6.1). For example: "su" refers to the amount one, "pepi" to two, "petite" to three. For shape 1, the signals "sunu" and "sutu" are used, "ginu" for shape 2, and shape 3 is referred to with "wipi" or "wipu". However, colours are less clearly demarcated by unique signal parts. This is also reflected in the ratio of unique signals produced during the test block ($M = 62.1\%, SD = 19.8\%$), indicating that some simulations sometimes result in repetitive use of the same signals for different meanings, resulting in a somewhat degenerate vocabulary. Such ambiguity may be the reason for non-perfect communicative success during communication. Nevertheless, it is clear that unseen stimuli are often labelled similarly to previously seen stimuli.

	Shape	Colour	Amount	Word
train set	3	orange	1	wipisu
	1	green	2	sutupepi
	2	green	1	ginisu
	3	green	1	wipisu
	1	blue	2	sunupepi
	1	green	3	sutupitite
	2	orange	1	ginusu
	3	blue	3	wipipitite
	3	green	3	wipupitite
	3	blue	1	wipisu
	1	blue	3	sunupitite
	2	orange	3	ginupitite
	2	blue	2	ginupepi
	1	orange	2	sunupepi
	2	orange	2	ginupepi
test set	1	orange	1	sutisu
	1	orange	3	sutupitite
	1	green	1	sutusu
	1	blue	1	sunusi
	2	green	2	ginupepi
	2	green	3	ginupitite
	2	blue	1	ginisu
	2	blue	3	ginupitite
	3	orange	2	wipupepi
	3	orange	3	wipipitite
	3	green	2	wipupepi
	3	blue	2	wipupepi

Table 6.1: An exemplary vocabulary that evolved in a simulation where the signals produced in the testing phase resulted in the highest *TopSim* score (7.13) after communication. The signals for the test stimuli share parts of signals and are composed similarly to train stimuli ($GenScore = .792$).

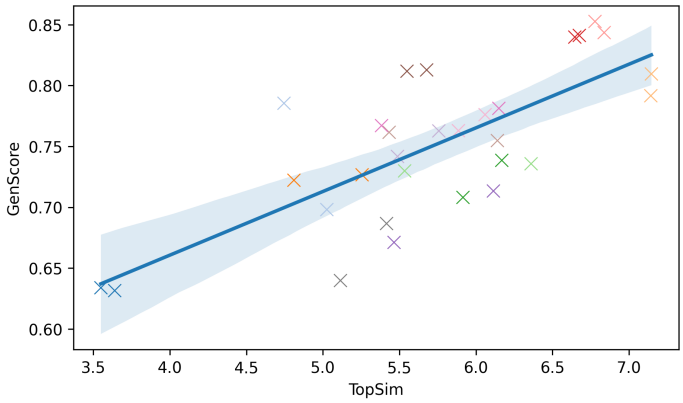


Figure 6.4: Languages that have evolved to be more structured allow for better generalisation to unseen test stimuli. Coloured crosses refer to individual simulations.

6.6 Iterated learning

The previous results showed that two LLMs can successfully communicate and slowly shape the language to become more structured. Provided that cumulative cultural evolution can extrapolate weak biases to have strong effects in socially learned systems like language (Smith, 2011), we extend our simulations by adding generations of learners. The first generation is initialised with a random unstructured language described in Section 6.3.1, but in the following generations, agents learn a portion of the signal-meaning mappings produced in the testing block by the agents of the previous generation. Only the vocabulary of the agent with the highest *TopSim* is transmitted to the next generation. We ran six transmission chains, each consisting of 8 generations. The seed generations for each chain were selected randomly from our initial 15 simulations.

6.6.1 Learnability increases

Iterated learning clearly increases the learnability of vocabularies (Figure 6.5). While LLMs in the first generation struggle to look up signals and reproduce them, a single generation of learning and using a language tremendously decreases the edit distance between ground truth signals and the produced signals. These results are remarkably similar to findings with human participants (Kirby et al., 2015), and show that the languages are optimised for LLMs' preferences.

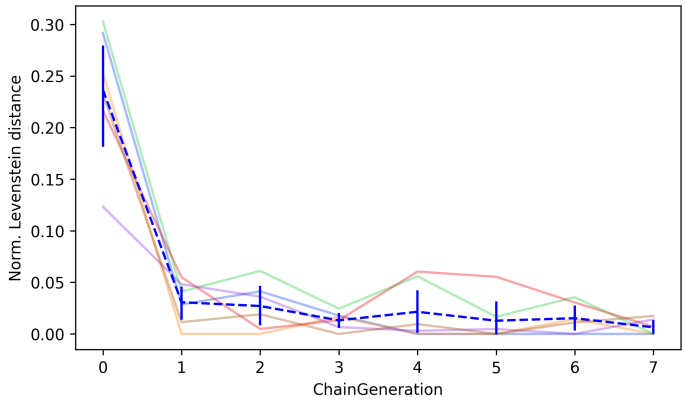


Figure 6.5: The normalised Levenshtein distance between the ground truth and the produced signal in the learning block. Solid lines indicate chains, and the dashed blue line indicates the average Levenshtein distance across simulations in a generation.

6.6.2 Communicative success and non-humanlike structures

Despite the increase in learnability, we do not observe an increase in communicative success due to iterated learning (Table 6.2, Figure C.1). This is possibly due to the already high scores of the first generation. Despite their increased learnability, the signals become significantly longer and more ambiguous. We take this non-humanlike solution to be an artefact of an absence of pressures for memorisation in LLMs. While human language is optimised to be compressible and expressive (Fedzechkina et al., 2012; Tamariz and Kirby, 2015; Kirby et al., 2015), the context windows of LLMs are considerably larger. In our case, Llama-3 70B has a context window of 8.2K tokens, which we do not exceed and therefore does not induce a pressure for compressibility.

Finally, the metrics to measure structure display a mixed picture. *TopSim*, does increase across generations but not significantly (Table 6.2). Yet, *Ngram* diversity decreases significantly across generations. For the evolution of these metrics across generations, see Section C.1. Qualitative inspections of several vocabularies show that some languages evolve into degenerate languages with repeating signals for different stimuli (i.e., underspecification). This is corroborated by a significantly lower number of uniquely produced signals in the last generation compared to the first simulation ($t(5) = 2.64, p = .046, M_{gen0} = .707, SD_{gen0} = .142, M_{gen7} = .519, SD_{gen7} = .119$). Together, this causes the *Ngram* diversity to be lower while clearly hurting communicative expressiveness. Even though degenerate languages are not uncommon in iterated learning experiments with humans (e.g., experiment 1 in Kirby et al., 2008), an additional pressure for expressivity typically prevents languages from becoming underspecified. Given the expressivity pressure that we imposed during the communication block, we expected to see less of such underspecification. The process of iterated learning, therefore, results in vocabularies

	$t(5)$	p	M_{gen0}	SD_{gen0}	M_{gen7}	SD_{gen7}
PercCom	.308	.770	.769	.077	.785	.123
TopSim	-1.42	.215	9.62	1.21	10.5	1.77
Ngram	2.83	.037	.158	.074	.071	.025

Table 6.2: The descriptives and statistics of the first ($gen0$) and last generation ($gen7$) in our chains. Paired t-tests show that *Ngram* diversity does significantly change resulting from generational transmission, while *TopSim* and *PercCom* do not.

that are optimised for the preferences of LLM agents but do so in a non-humanlike way.

6.7 Discussion

Our findings present a mixed picture; agents comprised of LLMs can learn and use artificial languages in a referential game. They do so by optimising the initially holistic vocabulary to fit better with the preferences of their language model, resulting in increased regularity and structure (Table 6.1). These human-like results are much in line with previous findings showing that structured languages can emerge from repeated interactions between interlocutors (i.a. Selten and Warglien, 2007; Verhoef et al., 2016b; Nölle et al., 2018; Raviv et al., 2019a). Yet, we also observe some degeneracy, i.e., many-to-one mappings of signals and attributes, and non-humanlike behaviours such as a tendency to produce long signals. Iterated learning further increases the learnability of the vocabulary but also extrapolates these non-humanlike behaviours further. Despite not being able to *directly* compare our results to human data, these findings are loosely comparable to earlier work involving human participants (Kirby et al., 2015; Raviv et al., 2019b) in which languages with similar properties emerge.

Table 6.1 moreover suggests that certain attributes, such as the colour attribute, in the inputs may be ignored, possibly due to the primacy and recency bias in LLMs (Liu et al., 2024). Optimising the instructive sentences by choosing sentences that maximise the fraction of valid model answers for each task, as suggested by Aher et al. (2023), may alleviate these ignorances and increase focus on relevant attributes. It is also possible that the LLMs do not 'experience' enough pressure to be understood by other agents, i.e., the *communicativeSuccess* attribute is not able to force a need to be expressive, which is deemed an essential pressure in computational simulations for human-like structures (Galke and Raviv, 2025). Despite these discrepancies, it is nevertheless interesting that some form of structure emerges.

Our results furthermore show variability between generations of learners. This is not uncommon in human experiments where processes of interaction and transmission sometimes generate fully systematic, compositional languages, but can also result in systems that lack structure entirely (Verhoef et al., 2022). In Chapter 2 we showed that differences in personal

biases may be a contributing factor to these differences. Since we do not initialise agents with different biases, these variations, originating in distributional information of the prepended vocabularies, are a natural human-like outcome of repeated exposure to and use of the language.

The evolution of degenerate vocabularies could be explained by the use of greedy decoding during signal generation, which does not necessarily produce the most human-like text (Holtzman et al., 2020; Meister et al., 2022, 2023) and may therefore also result in non-humanlike composition. Once an agent, perhaps mistakenly, duplicates a signal, its raw probabilities are increased when producing the next utterance, possibly resulting in a feedback loop that collapses onto a degenerate vocabulary. This effect may be further increased due to LLMs' inability to innovate (Bender et al., 2021; Yiu et al., 2024) and the choice of structured prompts that do not explicitly ask for innovation. Future work could attempt to increase the composition of novel signals by increasing the temperature parameter. Perhaps resulting in slightly more novel outputs as this forces exploration of the vocabulary embedding space (Peeperkorn et al., 2024), possibly alleviating the evolution of degenerate vocabularies and shifting the optimisation of the language to different solutions.

The rapid increase in learnability resulting from iterated learning proves that weak learning biases in language models, such as an observed simplicity bias (Chen et al., 2024), can be amplified by the process of generational transmission. Simulations with increased communicative difficulty, e.g., by increasing the number of distractors or the number of interaction partners, could reveal whether and how some form of memory constraint affects the learnability of languages, while also capturing the diversity and dynamic nature of language in the world more accurately. In general, systematic manipulations across model features (e.g., size, training data, or decoding strategies) may expose why we observe tendencies such as producing longer signals. Similar to what was proposed by Galke and Raviv (2025), we argue that careful manipulation of our setup can help reveal underlying mechanistic biases of language models and inform modelling choices when simulating language acquisition in LLMs. Taking into account the important role communication plays in shaping human language, LLM performance drastically increased when it was optimised for successful communication through RLHF.

Finally, we acknowledge that our results depend on several methodological considerations, including the model used, the prompt format, task instructions, and the tokenisation process. However, our primary goal was to investigate whether LLMs can be used in simulations of artificial language emergence. We aimed to stay as closely as possible to well-established experimental methods in the field of language emergence. We did not optimise for performance, human-like results, or compositional vocabularies. Instead, our goal was to reveal the natural behaviours of LLMs resulting from learning and using artificial languages. Future work could extend our findings by performing experiments in which humans collaborate with LLMs to investigate whether languages can evolve that are optimised for human *and* LLM preferences. Finally, as this chapter focused on experiments with a single LLM, future research should verify these findings across multiple LLM architectures to establish their generalisability.

6.8 Conclusion

Given the remarkable linguistic abilities of recent LLMs, we show how LLM-augmented agents behave in a classical referential game in which artificial languages, typically used in the field of language evolution, are learned and used. Primarily, our results suggest that LLMs can be used as artificial language learners to investigate the evolution of language. We showed that initially unstructured languages are optimised for improved learnability and allowed for successful communication. While we found some evidence of human-like compositional structures that enhance generalisation abilities, we also identified notable differences in the behavioural characteristics of LLMs compared to humans. Notably, iterated learning processes increased vocabulary learnability but also amplified such different characteristics further. As such, we extend existing research by revealing that structured languages are not merely easier for LLMs to learn. Critically, the inherent biases of LLMs also shape unstructured languages towards increased regularity. These findings contribute to a deeper understanding of how LLMs process and evolve language, potentially bridging the gap between computational models and natural language evolution. Finally, we hope to have shown that our setup is helpful in exposing the underlying mechanistic biases of LLMs and demystifying their uninterpretable nature.

7

Shaping Shared Languages

Languages are shaped by the inductive biases of their users. Using a classical referential game, we investigate how artificial languages evolve when optimised for inductive biases in humans and large language models (LLMs) via Human-Human, LLM-LLM and Human-LLM experiments. We show that referentially grounded vocabularies emerge that enable reliable communication in all conditions, even when humans and LLMs collaborate. Comparisons between conditions reveal that languages optimised for LLMs subtly differ from those optimised for humans. Interestingly, interactions between humans and LLMs alleviate these differences and result in vocabularies that are more human-like than LLM-like. These findings advance our understanding of how inductive biases in LLMs play a role in the dynamic nature of human language and contribute to maintaining alignment in human and machine communication. In particular, our work highlights the need to develop new methods that incorporate human interaction into the training processes of LLMs, and demonstrates that using communicative success as a reward signal can be a fruitful and novel direction.

Originally published as: Kouwenhoven, T., Peeperkorn, M., de Kleijn, R.E. and Verhoef, T. (2025). Shaping Shared Languages: Human and Large Language Models' Inductive Biases in Emergent Communication. In Kwok, J., editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI-25, International Joint Conferences on Artificial Intelligence Organization. Human-Centred AI

7.1 Introduction

Languages adapt to how they are learned and used. The primary reason is the continuous influence of individuals' (learning) biases and pressures that slowly shape languages to become more structured, easier to learn and communicatively efficient (Smith, 2022). Although a wealth of experiments in the field of language evolution have contributed to this knowledge (i.a. Kirby et al., 2014, 2015; Raviv et al., 2019a), only relatively recently have we started investigating whether these principles can be applied to large language models as well (Galke et al., 2024). For instance, more systematic and structured languages are typically easier for humans to learn when asked to learn novel artificial languages (Raviv et al., 2021). Recent work by Galke et al. (2024) revealed that the same is true for recurrent neural networks and transformer-based LLMs. Moreover, transmission of initially unstructured language systems over generations of human learners (i.e., iterated learning) increases structure and learnability in these languages (Kirby et al., 2015). To investigate whether this process leads to a similar outcome with LLMs, the work in Chapter 6 created a setting in which LLMs learned an initially holistic, unstructured artificial language and then repeatedly used it to communicate in a referential game. This showed that the linguistic structure of these languages increased, which enabled more successful communication between LLM agents, again mirroring observations from human experiments (Kirby et al., 2015).

With AI systems being increasingly incorporated into our daily lives (Brinkmann et al., 2023), it is argued that repeated interactions with machines become increasingly important to maintain alignment (Mikolov et al., 2018; Beuls and Van Eecke, 2024) and referential grounding (Chapter 1). In the case of humans, these repeated interactions cause languages to evolve in a way that accommodates the specific abilities and preferences of minority individuals at the group level (Josserand et al., 2024). Since the seemingly similar ways that languages adapt and optimise as a result of learning and use in both Human-Human and LLM-LLM interactions, the question that arises is whether these processes can also be used to experimentally evolve a language that is optimised for humans *and* LLMs. In other words, can humans and LLMs collaboratively shape a language that is easy to learn for both and allows for successful communication? If so, what do these languages look like?

This is investigated here by extending Chapter 6. Firstly, we provide experimental data of humans playing the same referential game used with LLMs in Chapter 6, allowing comparisons between languages evolved through LLM-LLM interactions with those resulting from Human-Human interactions. Secondly, we run experiments where humans collaborate with an LLM (Figure 7.1)[1]. While it is unclear how human and LLM abilities exactly differ, this allows us to test whether an artificial holistic language can be optimised for the inductive biases of two different types of language learners. If shared vocabularies of signals and meanings emerge that allow for successful communication, one could argue that there has been some form of

[1]This study was approved by the ethics department of Leiden University (2024-03-11-R.E. de Kleijn-V1-5354)

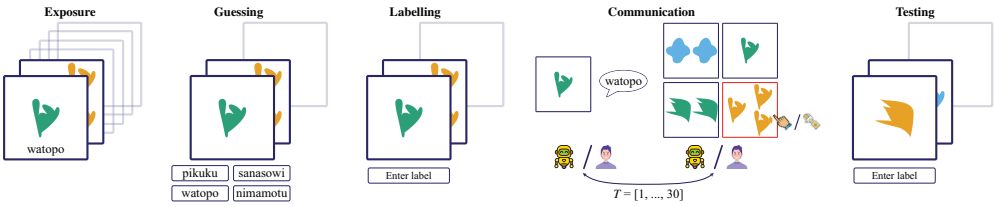


Figure 7.1: The experimental blocks in our experiment. Participants go through the exposure and guessing block twice before labelling each of the 15 training stimuli in the labelling block. The communication block is performed for 4 rounds each consisting of 30 tasks T, where participants alternate speaker-listener roles for each stimulus once. Participants label 27 (15 original and 12 novel) stimuli in the testing block. Image is adapted from Chapter 6. Icons obtained from flaticon.com

referential grounding, a prerequisite for successful communication (Clark and Brennan, 1991). Finally, Human-LLM collaboration allows investigating if and how the evolved language differs from languages that evolve within Human-Human and LLM-LLM interactions.

Our results show that structured and referentially grounded languages can emerge when humans and LLMs interact repeatedly in our experiment. The emergent languages from these interactions tend to be more human-like than LLM-like, suggesting that the LLMs are flexible towards the strong human preferences that shape the languages. Finally, languages optimised for LLMs result in less variation and are more degenerate than those optimised for humans.

7.2 Background

In this section, we discuss relevant work in the field of language evolution, the role of inductive biases in language evolution, and why this is relevant for LLM research.

7.2.1 Language evolution

Language allows us to communicate successfully because of the vocabulary we share, but also due to its open-ended nature, which enables the possibility of expressing novel meanings through compositional semantics. This defining feature of human language means that the meaning of any phrase is derived from the meanings of its individual components and the rules by which they are combined (Hockett, 1960). The evolution of compositionality has been investigated abundantly in the field of language evolution through human experiments (e.g. Kirby et al., 2008; Raviv et al., 2019a) and computational simulations (e.g. de Boer, 2006; Steels and Loetzsch, 2012; Lazaridou and Baroni, 2020). These experiments typically involve learning artificial languages or playing a signalling game. Here, *learning* artificial languages imposes a constraint for which it is believed to lead to more compressible and structured languages (Kirby et al., 2015). *Communication* in signalling games imposes a pressure for expressivity, requiring

participants to develop a vocabulary of signal-meaning mappings that allows them to communicate about novel stimuli. In this case, some form of referential grounding must be established through the process of repeated interactions. Participants—human or machine—generally establish novel signal-meaning mappings quickly, which enables successful communication.

7.2.2 Inductive biases

An important aspect of this chapter is the notion of biases. Here, we do not focus primarily on behavioural biases observed in humans (e.g., the confirmation bias), but rather are interested in implicit inductive biases that may result in biased language learning. This is relevant since seemingly arbitrary aspects of linguistic structure may actually result from general learning and processing biases deriving from the structure of thought processes, perceptuo-motor factors, cognitive limitations, and pragmatics (Christiansen and Chater, 2008). Especially so since fundamental predispositions influence how humans and artificial systems learn and process information (i.e., language). At a population level, these biases may *manifest* themselves as preferences for compressibility, simplicity, and efficiency—cognitive tendencies (Kirby et al., 2015; Tamariz and Kirby, 2015; Gibson et al., 2019) that naturally influence language evolution. For example, in the case of human systems (e.g., language) that are culturally transmitted, a memory constraint can enforce systems to be easy to learn and simple, because the hard-to-learn elements are less likely to be transmitted. Furthermore, the sound systems of human languages seem to be optimised for criteria such as acoustic distinctiveness or articulatory ease (Liljencrants and Lindblom, 1972; Lindblom and Maddieson, 1988) through a process of self-organisation (De Boer, 2000). Some even argue that humans' cognitive limitations may be *beneficial* for language acquisition (DeCaro et al., 2008; Poletiek et al., 2018).

Human constraints like these could well have evolved differently and are inherently different between humans and computational language learners, such as reinforcement learning agents and LLMs (although this is an ongoing debate (Kozachkov et al., 2023)). In the case of simulations of (reinforcement learning) agents, inductive biases typically do not match those present in humans. As such, they are often induced artificially by incorporating biases to guide learning dynamics as a means to recover human-like properties (for a review see Galke and Raviv, 2025). In the case of LLMs, which are fundamentally different from humans, we focus on increasingly apparent inductive biases of the Transformer architecture (Futrell and Mahowald, 2025) that may influence how languages evolve in the context of our experiment.

One example is a bias for simplicity. Rende et al. (2024) carefully cloned training data such that texts only contained between-token interactions up to a certain degree. Revealing that Transformers first learn low-degree between-token interactions, and only later learn high-degree interactions. Similarly, LLMs pick up grammar as the simplest explanation for data early during training. Only shortly thereafter, general linguistic capabilities arise (Chen et al., 2024). Moreover, Transformers seem to have an inductive bias favouring structure in (natural) language.

For example, GPT-2 models struggle more to learn impossible languages (e.g., languages lacking hierarchical structure or having unnatural or irreversible word orders) compared to English (Kallini et al., 2024), indicating that structure aids language learning. Additionally, the ability to generalise to novel stimuli increases when LLMs learn from more structured artificial languages (Galke et al., 2024). Recent work also revealed a primacy and recency bias in LLMs. They handle information better when it appears either at the beginning or towards the end of a prompt (Liu et al., 2024; Mina et al., 2025). Finally, LLMs have an inductive preference for verbose answers (Zheng et al., 2023; Saito et al., 2023), while humans prefer short, efficient answers (Gibson et al., 2019).

Although the underlying mechanisms of these biases differ between humans and machines, we find substantial overlap in terms of their behavioural effects. As such, we hypothesise that the aforementioned effects of continuous learning and use of language will also come into play when humans and machines collaborate, resulting in a language optimised for the preferences of both entities.

7.2.3 Why is this relevant for LLMs?

It is increasingly assumed that LLMs can be used as models of language (Millière, 2024) and that classical approaches from emergent communication can inform more human-like language learning in machines (Beuls and Van Eecke, 2024; Galke and Raviv, 2025). Moreover, language modelling and linguistics should complement each other (Futrell and Mahowald, 2025) as comparing LLMs to human language users, can help answer cognitive and typological questions (Warstadt and Bowman, 2022; van Dijk et al., 2023a). Vice-versa, methods from psychology can help to quantify inductive biases of LLMs (Griffiths et al., 2024; Galke and Raviv, 2025) or vision-and-language models (e.g. Chapter 4; Kouwenhoven et al., 2025) and compare them to known biases in humans.

For instance, the process of iterated learning, in which the transmitted information will ultimately come to mirror the minds of the learners (Griffiths and Kalish, 2007a), has been used to discover inherent LLM biases. Ren et al. (2024) showed that iterated learning causes subtle biases in LLM priors to be gradually amplified, Chapter 6 concluded that artificial languages can be optimised for LLM-augmented agents with iterated learning, and Shumailov et al. (2024) argue that generative models converge on uninterpretable junk when they are trained on AI-generated data. While the latter is typically seen as drift, crucially, we argue that what this shows is that the generated content is slowly shaped to be optimised for model preferences, *not* for human preferences. To prevent what Shumailov et al. (2024) refers to as model collapse, they argue that genuine human interactions with systems will become increasingly important. Similarly, Smith et al. (2024) responded that, like in human language transmission, the need to be expressive may prevent both the convergence on a few frequent uninformative sentences and the emergence of a long tail of uninterpretable junk.

These findings advance our understanding of internal LLM representations. This thereby contributes to maintaining alignment and mutual understandability between humans and machines in interaction. We address this by examining how adaptation processes unfold when humans and machines interact and develop a novel artificial language together.

7.3 Methodology

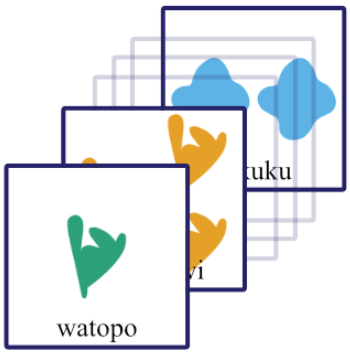
This experiment revolves around the classical referential Lewis game as implemented in Chapter 6, which is based on previous work in emergent communication (Raviv et al., 2021; Kirby et al., 2015, e.g.). We extend this setup to incorporate humans. In total, 45 participants participated in the experiment, 30 of whom formed 15 Human-Human pairs, and the remaining 15 interacted with an LLM in a Human-LLM setup. This allows us to directly compare languages adapted for human preferences to those adapted for the LLM-LLM simulations. But perhaps most interestingly, the Human-LLM condition provides an opportunity to investigate whether languages can be optimised for entities with different mechanisms and cognitive capacities (e.g., memory). If so, we can unravel what these look like.

During the experiment, participants first learn an artificial language and then use it to communicate with each other. The artificial language comprises a meaning space consisting of three attributes (shape, colour, and amount) that each can have three values, totalling to 27 unique stimuli. The corresponding labels are initialised following the design of Kirby et al. (2008), creating a holistic artificial language without structure (e.g., ”watopo”, ”sanasowi”, ”pikuku”) that contained a limited set of characters to prevent participants from writing English words. Participants first individually learn 15 random signal-meaning pairs through the exposure, guessing, and labelling blocks. Hereafter, participants are tasked to use the newly acquired language to communicate in a referential game. In this game, participants alternate between a speaker and listener role, where the speaker observes a target stimulus and labels it. Using this label as a signal, the listener is then tasked with identifying the correct target among three distractors. Cooperation is successful when the listeners' guess matches the target stimulus. After the communication block, there is a testing block in which participants individually label 27 meanings, including 12 unseen meanings, to assess how well they generalise to novel inputs. The duration of the entire experiment is roughly 70 minutes. An experiment overview is provided in Figure 7.1. [2]

7.3.1 LLMs as participants

Human participants learn the language by going through the exposure and guessing blocks twice. They iteratively go over the 15 training stimuli and may extract some apparent, but not present, patterns or consistencies. They are then tasked to label the stimuli, before moving on to

[2] All code, materials, and data are available on OSF: `https://osf.io/52yar/`.



{'shape':1,'colour':'green','amount':1,'word':'watopo'}
{'shape':3,'colour':'orange','amount':3,'word':'sanasowi'}
.
.
.
{'shape':2,'colour':'blue','amount':2,'word':'pikuku'}

Figure 7.2: Left: Humans learn the language by being exposed to stimuli and the corresponding signals in the exposure block. Right: LLMs learn the same vocabulary by virtue of in-context learning. A JSON-like structure containing the signal-meaning mappings is prepended to each prompt to serve as learning stimuli.

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|> You are a
 language learner who has to learn an artificial language with words
and their corresponding features. Your task is to complete the
vocabulary by generating a word that describes the last item. Only
respond with the word.<|eot_id|><|start_header_id|>user<|
end_header_id|>\n
{'shape':2,'colour':'orange','amount':1,'word':'giniwite'}
{'shape':1,'colour':'green','amount':3,'word':'hanosa'}
    :
{'shape':3,'colour':'blue','amount':2,'word':'tusetetu'}
{'shape':1,'colour':'green','amount':3,'word':'<|eot_id|><|
start_header_id|>assistant<|end_header_id|>[comp/prefill]
```

Prompt 7.3.1: A prompt snippet used for labelling and guessing. During communication, we add a `communicativeSuccess` attribute, update the system prompt to inform about the communicative task, and instruct that 'Communicative success is important'.

the communication block in which they interact with another human or a LLM. In either case, they were told that they interacted with a human. The LLM agents, however, are not updated and receive instructions to learn the languages by virtue of in-context learning. Specifically, the stimuli are presented in a structured, JSON-like format (Figure 7.2) that has proven to be effective in Galke et al. (2024) and Chapter 6. As such, we assume that these signal-meaning mappings in the context of a prompt provide enough (distributional) information for a LLM to learn a mapping between the attributes of the stimuli and signal syllables (Prompt 7.3.1). Although the prompt structure 'invites' the LLM to infer a signal from the stimulus attributes, we are agnostic about how exactly and what kind of mapping is deduced, but are interested in

the resulting behaviours. In the experiments, we use the instruction-tuned variant of Llama-3 70B (Llama Team, 2024) with greedy sampling. We opt for an instruction-tuned model since this allows us to specify the need for communicative success. This potentially affects how the model's inductive biases are expressed, but we leave this for future work.

One of essentially two tasks is performed throughout the experiment: labelling or guessing. The labelling block and speaking in the communication block involve labelling, and the guessing block and discrimination during communication involve guessing. Generating signals is achieved through prompt completion. Guessing is done by prefilling the prompt with distractor stimuli or labels and selecting the item with the highest probability. This alleviates LLMs' inconsistent behaviour in answering multiple-choice questions (Khatun and Brown, 2024), and follows recommendations from computational linguistics (Hendrycks et al., 2021; Wang et al., 2024). During communication, we add a `communicativeSuccess` attribute set to 1 if the previous interaction for this stimulus was successful and zero otherwise. This attribute functions as a memory between interactions and acts as a pressure for expressivity. In human language evolution, such pressure plays an important role since it prevents languages from becoming degenerate (Kirby et al., 2015). Importantly, the agents observe the training vocabulary in their context *with* the current stimulus in the guessing and labelling block, rendering them as simple look-up tasks. We do, however, *not* include the current stimulus during communication and testing, requiring the agents to extract an appropriate mapping and generalise to new stimuli. Akin to standard practice in older simulations (e.g. Steels and Loetzsch, 2012), the agent vocabularies are updated when labels are generated after the labelling block and during the communication block. This allows the vocabularies of signal-meaning mappings to evolve over the course of the simulation. As such, prompts are slightly different after each interaction. Moreover, given the primacy and recency bias in LLMs (Liu et al., 2024; Mina et al., 2025), we shuffle the vocabulary before creating prompts to account for unwanted ordering effects. Some example prompts used in the Human-LLM condition are displayed in Section D.1.

7.3.2 Metrics

Besides comparing the percentage of communicative success (*PercCom*), the primary goal of this work is to understand what a language looks like when optimised for different entities. Specifically, we investigate whether the languages display some degree of structure in the form of compositionality. In this experiment, this means that attribute values are denoted with label parts that are reused to describe other similar stimuli. Capturing this is not at all trivial, especially provided the freedom given to participants when they label stimuli. A common metric that gauges whether similar meanings map to similar signals is Topographic Similarity (*TopSim*; Brighton and Kirby, 2006). While providing a good indication of compositional language use, it does not account for variability in language, such as word-order freedom. It could therefore show an incomplete picture (i.e., a low *TopSim*) as languages can still be compositional

despite having multiple word orders, the existence of synonyms, or homonyms (Conklin and Smith, 2023). Hence, we report multiple metrics in addition to *TopSim* that together indicate the degree of compositionality. Specifically, we report on *synonymy* (one-to-many mappings), *homonymy* (many-to-one), and word order freedom (*Freedom*), for which Conklin and Smith (2023) proposed entropy-based metrics. A language where each attribute value is encoded by a single character in a position has low entropy, and thus a low *synonymy*. Languages with a uniform distribution over all characters to refer to an attribute value have high synonymy. *Homonymy* is similar; it looks at how many attribute-values a character in a position can refer to, i.e., when *homonymy* ≈ 1 characters can map to multiple attribute-values. Finally, we compute word-order freedom (*freedom*) to account for variability in the order by which labels are composed. It assesses whether each value of a specific attribute is encoded in a specific position of the label, i.e., there is little freedom, or whether attribute values can be encoded in any position of the label, i.e., displaying a high degree of word order freedom (*Freedom* ≈ 1).

Systematic generalisation to novel stimuli is assessed through the generalisation score *GenScore* from Raviv et al. (2021). It gauges whether the labels produced for unknown (i.e., testing) stimuli are labelled in consistent ways to labels produced for similar known (i.e., training) stimuli. In addition to character-based metrics, we assess whether participants reuse parts of labels in different labels by computing the *Ngram* diversity (Li et al., 2016) over all the produced labels in a block. *Ngram* diversity is the average ratio of unique vs. total *Ngrams* for $N \in \{1, 2, 3, 4\}$ in all labels. Low *Ngram* diversity implies that labels are composed of reused parts, and high diversity means that labels do not share many *Ngrams*, thus are very different. The percentage of unique labels captures the degree of degeneracy (*RatioUniLabels*). Finally, we measure whether a pressure for communicative success, known to drive efficiency in human experiments (Smith and Culbertson, 2020), results in shorter labels using *WordLength*.

7.4 Evaluation

We use linear mixed effect models to analyse our results. Specifically, we fit $PercCom \sim Metric + (1|RoundId)$ where $Metric$ can either be *TopSim*, *RatioUniLabels* and is the average value of two players in a round. To measure effects across conditions, we use $PercCom \sim Metric + Metric * Condition + (1|RoundId)$. The slope $\hat{\beta}$ determines the direction of the effect and the rate of change. Additionally, we use conditional R^2_c, and marginal R^2_m (Nakagawa and Schielzeth, 2013). The former considers fixed and random effects to show how much variance can be explained by the model. Higher values of R^2_c indicate that the model captures more variance and that correlations are stronger. R^2_m describes how much variance can be explained by the fixed effects. We report Pearson's R to describe the relationship between *TopSim* and *GenScore*, and use a paired T-test, or Welch's test when assumptions of normality and variance are not met, to assess whether the metrics differ significantly.

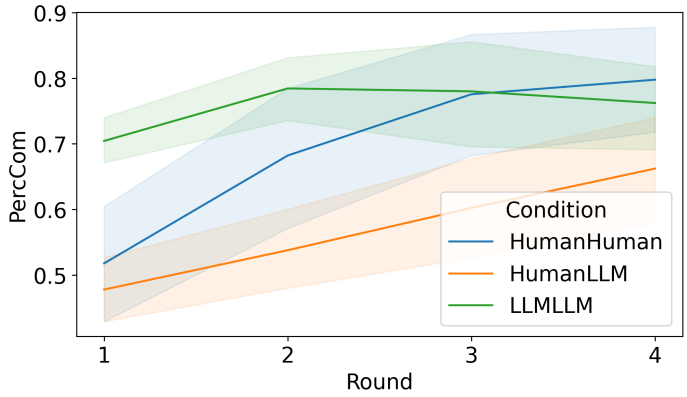


Figure 7.3: The average communicative performance (*PecrCom*) per round across the conditions. Communication steadily increases over rounds except for the LLM-LLM condition, in which coordination happens in the first round but does not increase afterwards.

7.5 Results

Human artificial language learning happened in a way that was expected based on earlier work (Kirby et al., 2015; Raviv et al., 2021). The results of these 15 Human-Human ($n = 30$) experiments act as a benchmark of human behaviour in our setup. We find that learning artificial languages is not a trivial task. After two rounds of exposure, labels were correctly guessed approximately half the time ($47.0\% \pm 49.9$). Freely labelling stimuli was done correctly only in $10.4\% \pm 30.6$ of the labels. Nevertheless, reliable communication protocols emerged during the communicating block; interactions were significantly more successful in the final round compared to the first round (Figure 7.3, $t(14) = -6.30, p < .001, d = 1.63, PercCom_{r1} = .518 \pm .176, PercCom_{r4} = .798 \pm .169$). *TopSim* positively influenced communication (*PercCom*) ($\hat{\beta} = .087 \pm .009, R^2_c = .731, R^2_m = .714, p < .001$) and generalisation to new stimuli was more consistent when the languages that evolved during communication displayed more *TopSim* ($r = .826, p < .001$).

A qualitative inspection revealed that, during communication, participants quickly replaced the labels they learned before, with only parts of labels 'surviving' this cut. Moreover, the number of shapes displayed (1, 2, or 3) was sometimes encoded by repeating the shape and colour labels several times, e.g., "pufepufe" was used to indicate two green shapes. Although expressive, this solution does not generalise to larger numbers and is therefore arguably not compositional.

Artificial language learning in LLMs was assessed in Chapter 6. Here, we briefly discuss the results of LLM-LLM ($n = 15$) simulations. LLMs guessed labels correctly for $97.3\% \pm 16.1$,

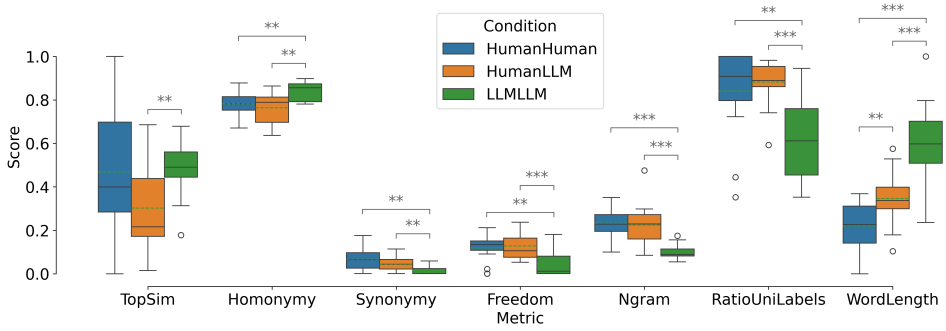


Figure 7.4: An overview of structure metrics used to measure compositional structure in the languages produced in the testing block. Generally, languages optimised for LLMs differ from those optimised for humans. Languages optimised for both mediate these differences. The asterisks indicate whether an independent Welch's t-test reveals a significant difference between the conditions where $* \; p < .05$, $** \; p < .01$, $*** \; p < .001$. *TopSim* and *WordLength* are normalised to values between 0 and 1 for visualisation purposes.

and the produced labels in the labelling block exactly matched the initial labels for $45.3\% \pm .498$. Indicating that learning the language is easier for LLMs than for humans. This is not surprising since the target stimulus is present in the prompt context in these blocks, and there is virtually no memory constraint. Communication happens reliably as well (Figure 7.3), however, communication can—but does not always—result in degenerate vocabularies with few uniquely used labels ($RatioUniLabels = .621 \pm .198$), significantly differing from human diversity in labels ($RatioUniLabels = .841 \pm .201, t(28) = -3.01, p = .005$, Figure 7.4). Interestingly, this happens even though the ratio of unique labels is modestly related to *PercCom* during communication ($\hat{\beta} = .300 \pm .134, R^2_c = .180, R^2_m = .092, p = .025$), suggesting that expressiveness is beneficial. A tentative explanation could be that aligning vocabularies happens much faster than in the other conditions. While humans would optimise languages whilst retaining expressiveness, LLMs start producing more duplicate and longer labels. We also ran additional simulations with other smaller models (i.e. Llama-3-8B (Llama Team, 2024), OLMo-2 7B, and OLMo-2 13B (Walsh et al., 2025)). The results are presented in Section D.3. In general, learning the artificial languages was comparable to Llama-3 70B for all models, but communication proved more difficult for the smaller models. Here, agents comprised of OLMo-2 7B models struggled the most and were unable to communicate robustly above chance levels.

What about Human-LLM communication? Our main contribution comes from the Human-LLM condition in which participants ($n = 15$) collaborated with LLMs. Successful communication necessitates that both entities adopt their behaviours and that a reliable referentially grounded vocabulary emerges. This is especially interesting since we observed that the ability to learn artificial languages differs between humans and LLMs, and that the optimised

languages differ in their use of homonyms, i.e., duplicate labels for different meanings. Despite these differences, communication *is* still possible ($PercCom = 0.662 \pm 0.161$, Figure 7.3). The final performance was lower than the other conditions, but the data suggest that prolonged interactions may result in higher communicative success. This exciting result shows that, even though the learning mechanisms of both entities may initially learn different signal-meaning relations, communication is possible. As such, the process of repeated learning and use of these artificial languages can overcome initial differences and indeed shape languages to be optimised for humans *and* LLMs. Out of 15 participants, 9 believed their partner was another human, despite communicating with an LLM. Performance did not change significantly as a result of this.

What do these languages look like? Having established that participants can communicate, we now examine if and how languages differ across different experimental conditions, focusing on the language metrics. Generally, we find that languages optimised for LLMs differ from those optimised for humans, and that languages optimised for *both* are more human-like than LLM-like (Figure 7.4). This is most notably visible in the ratio of uniquely produced labels, their respective lengths, and the *Ngram* diversity. Languages optimised for humans contain more unique labels, have higher *Ngram* diversity, and are shorter when compared to languages optimised for LLMs. The latter of which corroborates the well-known human preference for efficient communication (Smith and Culbertson, 2020). The length, number of unique labels, and diversity of label parts resulting from human-LLM collaboration seem to adhere more to human preferences than to LLM preferences. A similar pattern is visible for the compositionality metrics that allow for variation. There is more *homonymy* in LLM-optimised languages than in human-LLM languages, suggesting that the meanings of these words should be disambiguated by the context (i.e., the distractor stimuli) in which they appeared. This strategy is not straightforward and perhaps requires more cognitive capacity than available for humans, which could result in lower performance in the collaborative condition. It also seems that humans introduce synonymy into languages, i.e., they use more than one character to refer to specific attribute values. This introduces variability that can explain why the ratio of unique labels and *Ngram* diversity is higher in the collaborative condition. Finally, the word order of messages is somewhat flexible for humans, whereas LLMs tend to converge on a more fixed word order. The languages shaped by both seem to have human-like word order freedom. While these variations may introduce difficulties for LLMs to decode the meanings during communication, we do not find that *PercCom* is affected by the degree of *homonymy*, *synonymy*, or *freedom*.

The canonical *TopSim* metric suggests structure is lower when humans and LLMs collaborate compared to LLM simulations. This makes intuitive sense, given our observation that it was also more difficult to establish successful communication. Nevertheless, a linear mixed effects model fitted to predict *PercCom* with *TopSim*, the experimental condition, the interaction

between them, and a random effect for round revealed that *TopSim* strongly affects *PercCom* ($\hat{\beta} = .092 \pm .008, R^2_c = .580, R^2_m = .580, p < .001$). This means that irrespective of the experimental condition, a higher degree of structure in the produced labels was beneficial for communication. Moreover, generalisation to novel stimuli happened more consistently when the languages in the last round of communication displayed more structure. Again, confirmed by the mixed effects model predicting *GenScore* with *TopSim*, the condition, their interaction and round as random effect ($\hat{\beta} = .047 \pm .006, R^2_c = .784, R^2_m = .673, p < .001$, Figure D.1).

Generally, we find that while remarkably human-like, the languages shaped by intrinsic LLM constraints are in fact subtly different from those shaped by humans. Thereby providing a more nuanced view of what was observed in Chapter 6 as it only looked at *TopSim* and *NGram* diversity. Returning to the question of what languages optimised for entities with different inductive biases look like, they seem to be shaped in such a way as to conform more to human pressures than those present in LLMs.

7.6 Discussion

The primary goal of our work was to investigate if and how artificial languages differ when they are optimised for human *and* artificially intelligent language users. To do so, we extended Chapter 6, suggesting that LLMs can shape and use languages in referential communication. That setup was adapted to allow participants to interact with other human participants and with LLMs. This enabled controlled comparisons between the languages that evolve under different conditions. Our findings showed that human pairs, LLM pairs, and Human-LLM pairs can learn and successfully use languages in a referential game. This suggests that mechanisms that influence how language evolves for humans, specifically, learning and using a language repeatedly (e.g. Smith, 2022), also apply to computational and collaborative Human-LLM settings. In all conditions, successful communication was achieved by optimising an initially holistic, unstructured vocabulary to fit better with the inductive biases of the language users. We compared the languages across conditions and revealed that 1) while very human-like, LLM languages tend to be more strict (i.e., there is little variation), and that 2) languages adapted for human-LLM pairs tend to be more human-like than LLM-like (i.e., they are more diverse and have variation). Overall, our findings corroborate earlier claims that interactions between humans and machines are beneficial to establishing some form of referential grounding (Mikolov et al., 2018; Beuls and Van Eecke, 2024).

On the level of vocabulary, the ratio of uniquely produced labels by LLMs revealed that vocabularies can become degenerate. While this is also observed in human experiments when there is no pressure against it (Kirby et al., 2008), communicative success as a pressure is typically enough to prevent this (Kirby et al., 2015, e.g.). In contrast, even though the LLMs in our experiments were tested in a communicative setting, this did not prevent the languages from becoming underspecified. Possibly, this happened because the instruction to achieve

communicative success was not explicit enough and did not induce sufficient pressure for expressivity. Alternatively, it may be the case that the LLMs solved the problem in a non-humanlike manner and employed some kind of shortcut learning (Schwartz and Stanovsky, 2022; Mitchell and Krakauer, 2023). It could, for example, be that the distractors did not require the labels to be very specific, but instead allowed using underlying concurrences that were picked up by LLMs but not by humans. This would also explain the wide range of scores on this *RatioUniLabels*. On a character level, we observe related patterns in the form of high levels of *homonymy*, meaning that attribute values could be associated with multiple label characters, and that context was necessary to disentangle the correct meaning. While the duplicate labels can explain these scores for LLM simulations, this is not the case for humans. Here, the surprising behaviour of repeating label parts to indicate the *amount* attribute can explain the *homonymy* values.

The process of tokenisation plays an important role in these simulations. One could argue that it may help in learning some mapping between tokens and meanings. The meaning attributes and their values are common English words, while the initialised artificial languages consist of non-words that are tokenised into separate tokens. Meaning that the LLM is presented with a parsed set of attribute meanings and chunks of labels (i.e., the tokens). All that is left is to attend to a specific token given a particular meaning, which is precisely what a Transformer model is made for. Nevertheless, this does not undermine that these models *indeed* attribute attention correctly and that this produces human-like languages.

Similarly, our observation that a shared referential communicative system can be established is quite remarkable. Humans and LLMs may well use entirely different mechanisms and learn different relations between meanings and signals. Yet, their vocabularies become referentially grounded and are pragmatically understood by both humans and LLMs. This confirms that even though LLMs are not trained for this task, they can be used as relatively unbiased language learners (Wilcox et al., 2023), thereby providing a concrete example of how a pragmatic view of understanding, as argued for by van Dijk et al. (2023a), can be beneficial for collaborative tasks. This work also underscores the point made by Millière and Rathkopf (2024) that how LLMs or other AI models solve a cognitive task cannot be used as an argument against particular cognitive competences or language understanding, as long as the solution generalises.

Our results concretely corroborate the idea that insights from emergent communication literature can inform and improve language learning in language models (Smith et al., 2024; Beuls and Van Eecke, 2024; Galke and Raviv, 2025). We observed that just as languages accommodate for specific abilities and preferences in humans (Josserand et al., 2024), Human-LLM languages also adapt to the abilities and preferences of their users in that they are more human-like than LLM-like. Specifically, human preferences for simplicity and efficiency (Kirby et al., 2015; Gibson et al., 2019) likely drove vocabulary diversity while reducing lengths to human-like levels. This indicates that, in this experiment, LLMs are more flexible communicative partners than humans. These findings reinforce the idea that repeated interactions with humans are

crucial to maintaining referentially grounded human-like vocabularies instead of training only on recursively generated data (e.g. Shumailov et al., 2024) or using AI-augmented optimisation algorithms (e.g. Lee et al., 2024). In particular, this underscores the need for new methods that incorporate human interaction into the training processes of LLMs and shows that using communicative success as a reward signal can be a fruitful and novel approach.

Finally, we acknowledge that our results depend on methodological considerations, including the use of in-context learning, the model, the prompt format, and the sampling method. However, the primary goal was to extend previous work by investigating if languages optimised for human *and* LLM preferences can evolve. As such, we stayed close to well-established experimental methods in the field of language emergence and used prompts developed in Chapter 6. Importantly, we did not optimise for communicative success, human-like results, or compositional vocabularies.

7.7 Conclusion

Given the growing presence of contemporary LLMs in everyday life, there is an increasing need to understand their inductive biases to maintain alignment with humans. We tested whether general mechanisms of language learning and use have similar effects in an artificial language learning experiment conducted with Human-Human, LLM-LLM, and Human-LLM pairs. We show that referentially grounded vocabularies emerge in all conditions, indicating that initially unstructured artificial languages can be optimised for inductive biases of different language users. Comparisons across conditions revealed that, while similar to human vocabularies, LLM languages are subtly different. Interestingly, these differences are alleviated when humans and LLMs collaborate. This underscores that to achieve successful interactions between humans and machines, it is essential to optimise for communicative success. Overall, these findings advance our understanding of how LLMs may adapt to the dynamic nature of human language, contribute to its evolution, and maintain alignment with human understanding of language. While our setup only uses simple stimuli and basic languages, achieving this for human-level languages is a key research direction towards more natural language learning in LLMs.

8

Conclusions

This dissertation aimed to deepen our understanding of how inductive biases shape the emergence of languages across human, machine, and human-machine interactions by combining experimental and computational approaches. It proposed to take inspiration from the field of language evolution and investigate artificial language learning setups that differed in their linguistic complexity, interactivity, and language learners to unravel how human and artificial communication can complement each other. To achieve this, empirical studies were conducted on the evolution of language, comprising human, computational, and hybrid cognition. The chapters in this dissertation were structured such that they addressed increasingly complex linguistic phenomena but varied in their approaches, comprising modelling techniques, experiments, and behavioural probing of (vision-and-)language models. This concluding chapter synthesises our findings by addressing the Main Research Question (MRQ) in Section 8.1, by discussing each Research Question (RQ) outlined in Section 1.2.1, and by contextualising their broader implications. Finally, we reflect on this dissertation by addressing its limitations (Section 8.2.1) and provide an outlook on future work (Section 8.2.2)

8.1 Answers to Research Questions

RQ1 - (Chapter 2)

> *What role does diversity in biases for structure play in developing symbolic communicative systems?*

To answer this question, this chapter employed the Embodied Communication Game (ECG; Scott-Phillips et al., 2009) in which participant pairs ($N = 23$) were required to develop a symbolic communication system without having a conventional communication medium. We evaluated the process of establishing these systems and administered the Personal Need for Structure scale (PNS; Neuberg and Newsom, 1993) to measure participants' bias for structure since the social coordination of a shared language, which is initially unstructured, can be influenced by an individual's need for structure. We demonstrated that establishing effective communication is not trivial, with only 11 pairs successfully establishing a robust means of communication. This happened through establishing a common ground in the form of an initial convention, which enabled bootstrapping new signals from this common ground. Although not trivial, this resulted in symbolic vocabularies that were mutually understood and highly expressive. Interestingly, this process was more successful when participant pairs differed in their respective need for structure, specifically when they differed in their response to unpredictable environments.

From a broader perspective, our results suggest that diverse biases may be beneficial in creating communication systems, providing nuance to what is typically assumed: that alignment aids cooperation (e.g. Tylén et al., 2013; Scott-Phillips and Kirby, 2010). Given the prominent role of a human preference for simplicity and structure in language evolution research, these results suggest that it is interesting to investigate how differences in PNS affect a multitude of language evolution experiments. For example, in iterated learning experiments where small individual biases may have group-level effects (such as in Kirby et al., 2008; Theisen-White et al., 2011; Verhoef et al., 2011), or in experiments involving interactions in which humans trade-off expressivity and simplicity (e.g. Kirby et al., 2015; Raviv et al., 2019a). Moreover, these results resonate with the findings that communicating with multiple different interaction partners introduces pressures that result in more stable shared vocabularies for humans (Raviv et al., 2019a) and Reinforcement Learning (RL) agents (Rita et al., 2022a). Though highly speculative and extrapolated, in light of contemporary computational models, the results suggest that differences between human and machine understanding of language can be alleviated through interactions and provide an exciting area for future research.

RQ2 - (Chapter 3)

> *What insights about human sequential processing can be derived from modelling human behaviour in emergent communication?*

In this chapter, we used the behavioural data obtained in Chapter 2 and performed behaviour cloning to instil human sequential behaviour in the ECG in deep neural networks. As such, we used neural networks as observationally adequate approximations of human behaviour. We *approximated* latent human cognitive variables using computational tools to understand human behaviours that are important during the emergence of symbolic communication systems. We found that LSTMs can learn the behaviours associated with creating signal-meaning mappings, but did not find a correspondence between the latent cognitive variables and our cognitive measures of a bias for structure (*PNS*, *F1*, and *F2*). Nonetheless, we demonstrated that bidirectional LSTMs are better at capturing human behaviour than unidirectional LSTMs, suggesting that human sequential processing in the ECG takes into account both previous and future states when planning the next move. Moreover, we found a relation between participants' approximated learning rate and their exploration parameter. This relationship supports the view suggesting that humans combine random and uncertainty-directed exploration strategies to develop optimal behaviour (Jepma et al., 2016; Schulz and Gershman, 2019). Finally, our modelling results resonate with the belief that there exists a planning mechanism for sequential signal production in humans (Lashley et al., 1951; Cohen and Rosenbaum, 2004), thereby informing RL simulations of emergent communication.

The agents in RL simulations of emergent communication typically comprise unidirectional Recurrent Neural Networks (RNNs) that process sequential data in one direction (e.g. Chaabouni et al., 2022; Lian et al., 2024). Yet, our results revealed that bidirectional LSTMs are better at capturing human sequential data, suggesting RL simulations should use bidirectional LSTMs instead of unidirectional RNNs. Although we were unable to extract computational derivatives of human bias for structure in this particular setup, our methodology can be applied to other emergent communication setups. Such close comparisons to human data and computationally capturing human biases may reveal differences between human and RL behaviours and bring simulations closer to human experiments, as shown, for example, by Galke et al. (2024) and Lian et al. (2023a).

RQ3 - (Chapter 4)

> *To what extent do vision-and-language models exhibit human-like cross-modal associations such as the bouba-kiki effect?*

To answer this question, this chapter moved towards more contemporary models of language and explored probing vision-and-language models (VLMs) for a well-known cross-modal preference in humans. We adapted experiments from psycholinguistics conducted with humans

(Nielsen and Rendall, 2013) and conducted them with CLIP, ViLT, BLIP2, and GPT-4o, which differ in their architectures, training objective and data, and cross-modal attention mechanisms. While earlier work claimed strong associations in VLMs (Alper and Averbuch-Elor, 2023), our approach tested the existence of cross-modal associations more directly and revealed a more nuanced picture. Out of the four models tested, only CLIP and GPT-4o displayed *limited* evidence for associations between syllables and image features. This effect disappeared when we incorporated a more comprehensive dataset and after performing analyses of two-syllable pseudowords, suggesting that the results depend on the architecture, size, prompt, and training details of the model in question.

The work presented in this chapter contributes to a growing body of work that evaluates whether VLMs align with human perceptions (e.g. Muttenthaler et al., 2023; Jones et al., 2024; Shiono et al., 2025; Kouwenhoven et al., 2025). Overall, our findings inform discussions on the origin of the bouba-kiki effect in human cognition and, at the same time, contribute to the development of VLMs that align more closely with human cross-modal associations. While this highlights a limitation in current VLMs, it also provides a promising direction for future work: determining whether VLMs exhibit strong inductive preferences that exist but differ fundamentally from those of humans, rather than lacking such preferences entirely. For humans, non-arbitrary cross-modal associations may benefit language learning (Imai et al., 2008; Perry et al., 2015), artificial systems seem not to show sensitivity for these non-arbitrary mappings, which warrants further investigation. For instance, by investigating whether VLMs, like humans, can benefit from training them on data that has aligned cross-modal associations. These associations could be inspired by human associative patterns, but might also be suited to the preferences of VLMs (assuming such preferences exist). Doing so might perhaps enhance their general understanding of how words and their compositions relate to the world. Increasing such alignment between humans and machines may promote more natural interactions. In light of emergent communication research, these results underscore the importance of careful progression in using increasingly complex models (e.g. Bouchacourt and Baroni, 2018; Mahaut et al., 2025) in RL simulations, which recurs in the next research question.

RQ4 - (Chapter 5)

> *What role does representational alignment play in the emergence of compositional language in reinforcement learning?*

This chapter involved simulating the emergence of compositional language use with RL agents. Borrowing the broadly used Emergence of lanGuage in Games framework (EGG; Kharitonov et al., 2021), we trained deep RL agents equipped with contemporary vision models in a referential game. We tested their ability to communicate (compositionally) about MS COCO images (Lin et al., 2014), Gaussian Noise, and Winoground image pairs (Thrush et al., 2022). In an attempt to understand what these agents communicate about, we employed Representational

Similarity Analysis (Kriegeskorte et al., 2008) to assess the degree to which agent representations aligned with each other, and in particular, with their respective inputs. We confirmed earlier findings by Bouchacourt and Baroni (2018) through showing that emergent languages do not appear to encode human-like conceptual features. Instead, the agent representations seem to drift away from their inputs while the alignment between agent representations increases. This enabled them to communicate about noise input and demonstrated that RL agents rely on spurious rather than conceptual image features. Importantly, we showed that the degree of inter-agent alignment is strongly related to Topographic Similarity (*TopSim*; Brighton and Kirby, 2006), the most common metric of compositionality. Informed by this undesirable relation, we introduced an auxiliary loss function to mitigate it. Nevertheless, when tested on a strict compositionality benchmark (Winoground), we found no increased performance despite having aligned inter-agent *and* image-agent representations and higher *TopSim*.

These findings again underscore the trivial fact that humans and machines are different. Our agents exploited spurious correlations, resulting in shortcut learning, a form of understanding that is in many ways not human-like, but introduces a new 'alien' kind of problem-solving (Schwartz and Stanovsky, 2022; Mitchell and Krakauer, 2023). That is not to say that these models are not capable of developing human-like languages, but it means that scholars need to be aware of the importance and potential impact of representational alignment when claiming compositional or grounded language emergence in referential games. To this end, we suggest incorporating targeted out-of-distribution evaluations by repurposing datasets designed to assess targeted linguistic phenomena. Re-purposing such datasets can reveal more directly whether agents develop the attested communicative abilities that are trivial to humans. Doing so provides a more comprehensive analysis, rather than relying solely on metrics. In the case of compositionality, metrics (e.g., *TopSim*) could be accompanied with evaluations on visio-compositional or spatial reasoning tasks (e.g. Thrush et al., 2022; Diwan et al., 2022; Kamath et al., 2023). It is important to note, however, that our results only apply to the referential game. Another popular setup concerns the task of reconstructing the input that was given to a speaker based on its signal (Chaabouni et al., 2019a, 2020; Conklin and Smith, 2023; Lian et al., 2024, inter alia). A possible explanation for our findings was observed in recent work, which reveals that the training objective in these reconstruction games does seem to prevent agents from aligning their representations and promotes compositionality (Ben Zion et al., 2024). In particular, their work showed that the objective in the referential game, but not the reconstruction game, can promote semantically inconsistent communication protocols; semantically similar inputs do not necessarily produce the same message. Importantly, they show that the objective in the referential game (the objective employed in this chapter) can be solved with 'unintuitive systems', i.e., systems that do not rely on inputs to produce messages. This likely explains our findings.

RQ5 - (Chapter 6)

To what extent can Large Language Models learn and use artificial languages in emergent communication, mirroring human patterns of language evolution?

In this chapter, we aimed to unravel whether our current most sophisticated models of language can be used as subjects in emergent communication research. Inspired by the experimental design of Kirby et al. (2015) and Raviv et al. (2021), we developed an adapted version suitable for LLMs, and simulated the referential game with two LLM-augmented agents. This allowed testing whether general processes of language learning and use, which for humans shape linguistic systems towards communicative efficiency, also optimise languages for inductive LLM biases. The results revealed that this indeed is the case: LLMs can learn artificial languages and successfully use them to communicate. We found that initially holistic unstructured languages exhibited more structure after several communication rounds. If vocabularies evolved to display more pronounced structures, generalisation to novel stimuli also occurred more reliably. Nevertheless, the evolved vocabularies also displayed some degeneracy (i.e., underspecification). Finally, generational transmission contributed to the emergence of vocabularies that were easier to learn for LLMs, mirroring findings in human experiments, but this process also showed that languages did not necessarily adapt in a human-like way as the agents showed a tendency to produce unnecessarily long labels.

This chapter extended earlier work by Galke et al. (2024) through showing that more structured languages are not only easier to learn for LLMs, but their inductive biases also naturally shape languages to have some form of structure. It moreover provides an example of how methods from psycholinguistics, specifically iterated learning (which reveals biases that remain hidden when studying single learners), can be helpful in exposing the underlying mechanisms of LLMs and demystifying their uninterpretable nature. Hence, these findings contribute to a line of work that aims to reveal underlying inductive biases of LLMs (e.g. Zheng et al., 2023; Rende et al., 2024; Chen et al., 2024). While our investigation primarily focused on linguistic biases, its approach is similar to more behaviourally oriented studies that employ LLMs in socio-cultural scenarios to simulate believable human behaviour (e.g. Park et al., 2023; Perez et al., 2024). Likewise, we use LLM-augmented agents to observe and compare LLM behaviour with human behaviour. Overall, the presented results are remarkably human-like, suggesting that collaborative human and LLM language evolution is a fruitful idea, which recurs in the next research question.

RQ6 - (Chapter 7)

Can humans and Large Language Models develop shared vocabularies through collaborative communication?

To address this question, this chapter extended the LLM-only simulations of Chapter 6 by conducting an experiment that incorporated human participants. This involved Human-Human

and Human-LLM pairs to complement our earlier findings and facilitate comparison of the languages optimised for different entities. Humans collaborated with either other humans ($N = 30$) or with an instruction-tuned Llama3-70B model ($N = 15$). Our results demonstrated that across all conditions, referentially grounded languages emerged that enabled reliable communication. Surprisingly, a vocabulary of shared signal-meaning mappings emerged even when humans *and* LLMs collaborated. This indicates that initially unstructured artificial vocabularies can be optimised for the inductive biases of different language users who may well represent said vocabularies completely differently. Through analysing the (compositional) structure of the optimised languages with a series of metrics, we discovered that languages optimised for LLMs subtly differed from those optimised for humans. These differences were alleviated in our hybrid experiment where humans and LLMs collaborated. Specifically, the languages shaped for inductive biases of Human-LLM pairs displayed characteristics more closely resembling human-like patterns than LLM-like patterns.

In the context of this dissertation, these findings corroborate the claims that interactions between humans and machines are beneficial to establishing referential grounding (Mikolov et al., 2018; Bisk et al., 2020; Beuls and Van Eecke, 2024; Brandizzi, 2023). In line with the well-established idea that the meanings of signals originate from how they are used in language (Wittgenstein, 1953; Christiansen and Chater, 2022), these findings further advance our understanding of how LLMs play a role in the dynamic nature of language and contribute to maintaining alignment in human and machine communication. We take them as a concrete example of how insights from emergent communication literature can inform and improve language learning in language models (Smith et al., 2024; Galke and Raviv, 2025; Beuls and Van Eecke, 2024). Though it may be tempting to always resort to similarities in preferences, experiences, and mechanisms, our work showed that a pragmatic approach to referential meaning-making, which ignores how meanings are exactly represented (van Dijk et al., 2023a) but incorporates repeated interactions, can result in referentially grounded vocabularies as well. In particular, it underscores that to achieve successful interactions between humans and machines, it is essential to optimise for communicative success. We believe that developing additional training methods that incorporate these principles represents a promising direction for future research aimed at natural language understanding in LLMs.

MRQ

> *How can human and artificial cognition in emergent communication complement each other?*

Turning to our main research question, throughout this dissertation, we revisited established methods from the fields of language evolution and psycholinguistics to advance our understanding of both human and artificial cognition. Some involved human participants, as in more classical experimental setups, while others employed various artificially intelligent systems as

language learners, such as LLMs. This interdisciplinary perspective is demonstrated through six empirical studies, showing that collaboration and pollination across disciplines are fruitful in unravelling the intersections of human and artificial cognition.

Starting with the experiments that investigated more elementary concepts important for the evolution of language, our investigation of vision-and-language models in Chapter 4 demonstrated limited alignment between machine and human cross-modal associations such as the bouba-kiki effect. Out of four models, only CLIP and GPT-4o demonstrated limited evidence for a bouba-kiki effect. These findings highlight that the behaviour of multi-modal models is based on different underlying factors than those shaping human cognition. They underscore the need to better understand what determines multi-modal predictions if we wish to align human and machine cross-modal associations. However, our RL simulations demonstrated that emergent communication setups—specifically referential games—are not trivial candidates for incorporating such alignment goals. While the embeddings of vision-and-language models are useful for many downstream tasks, leveraging them as input features to evolve human-like languages with conceptual alignment proved challenging. The communication systems between artificial agents exhibited representational alignment patterns bearing no connection to their initial inputs, limiting their direct applicability to learning human-like systems (Chapter 5). Nevertheless, we believe that the interactive nature of these simulations provides fertile grounds to induce referentially grounded communicative systems into machines. Perhaps by incorporating more human-like bidirectional processing in artificial systems to simulate better the planning mechanisms essential for effective communication (Chapter 3), changing the train objective as discussed before, or integrating human-in-the-loop learning.

The final two chapters build on the knowledge obtained in earlier chapters. They investigated and used LLMs in a collaborative interactive setting. From a language evolution viewpoint, the chapters concerned with the most complex linguistic behaviours, and from a computational perspective, they employed the most competent computational methods. We found that LLMs can act as mature language learners in emergent communication experiments and that their inductive biases, like humans, shape languages towards more structure (Chapter 6). Methodologically, we demonstrated that iterated learning uncovers inductive biases present in LLMs, revealing that general processes of *learning and using language* have similar effects on how languages are shaped in LLMs and humans. When humans and LLMs collaborated in a communicative artificial language learning task, they established shared languages whose characteristics more closely resembled human patterns than those of LLMs (Chapter 7). This indicates that human cognitive biases can effectively guide artificial systems toward more natural language learning, fostering mutually understood and referentially grounded vocabularies. These findings corroborate our findings in Chapter 2 where we established that repeated *interactions* are crucial for grounding symbolic signals, as arbitrary movements acquired communicative meaning only through such exchanges. In the case of human-machine communication, interactions not only offered a way of establishing mutual agreement but also facilitated a

means to extrapolate shared behaviours despite inherent differences in cognitive biases. Put differently, differences in human and artificial cognitive preferences can complement each other instead of hindering them in communicative tasks, as long as there are interactions to facilitate this. This shows that the meanings of signals can be realised through their use, especially when the entities using them rely on fundamentally different mechanisms that may represent these meanings differently. Nevertheless, aligning human and machine understanding of language may benefit from human contributions in the form of grounded meaning, efficiency constraints, and cross-modal associations. Such insights from human cognition offer informative insights for modelling artificial cognition. Practically, this suggests that optimising for communicative success between humans and machines benefits from leveraging the strengths of both cognitive systems rather than attempting to make artificial cognition perfectly mimic human cognition.

In conclusion, human and artificial cognition complement each other through their inherent differences rather than despite them. This complementarity offers promising directions for developing communication systems that are adapted to the cognitive strengths of both humans and machines, potentially leading to more natural communicative interactions. At the outset of this dissertation, we posited that languages obtain their meaning when we put them into practice. We hope this work demonstrates why interactions are crucial in establishing referentially grounded communication between men and machines, and that the reader, like us, considers emergent communication to be a fruitful approach to establishing this.

8.2 Reflection

Since we conduct empirical research, some elements warrant further reflection. In addition to the discussion and limitations mentioned in the chapters, this section highlights aspects that influence the generalisability and conclusions derived in the previous sections. Addressing these limitations lays out opportunities for future work that could contribute to a more nuanced understanding of our findings.

8.2.1 Limitations

We first reflect on the sample sizes of our studies. We continue by discussing the practical implications of our proposition that communicative pressure should be incorporated into the training objective of contemporary models. We conclude the limitations of this dissertation by elaborating on our reliance on behavioural probes and the influence of prompting in VLMs and LLMs, which was important in disclosing to what degree human and artificial systems aligned.

Limited homogeneous sample sizes

The work presented in Chapters 2, 3, and 7 involved gathering human participants who collaborated in language evolution experiments. The sample sizes for these studies are not extremely large, and the participants come from European countries; most of them are pursuing or in possession of university degrees. As such, it is evident that our findings require larger, more heterogeneous sample sizes to improve their generalisability.

In language evolution studies, there is no such thing as a 'correct' answer, making evaluations non-trivial. Despite the existence of some metrics, collaborative experiments studying language evolution in the lab additionally require manual qualitative inspection since the solutions found by humans are idiosyncratic. To give an example: the generalisation metric employed in Chapter 6 is based on Raviv et al. (2021). It assesses the extent to which labels for known stimuli are similar to labels for unseen stimuli using two pairwise distance-based metrics. While insightful, it relies on the form and appropriateness of these metrics (Levenshtein distance and semantic distance) and, therefore, at best, only gauges generalisation. We also relied on manual inspection in Chapter 2, in which participants reported their grounding processes that needed to be verified through inspecting their behaviours. The nature of these experiments thus limits their scale and warrants further reflection.

In the case of Chapter 6 and Chapter 7, we also have a limited sample size as we relied mostly on Llama-3-70B-Instruct. The LLM-LLM simulations of Chapter 6 used only Llama-3-70B-Instruct, thus strictly limiting the generalisability of the claims involving communication. However, since our work builds on that of Galke et al. (2024), we also know that text-davinci-003 can, at the minimum, learn artificial languages. Thereby somewhat strengthening our conclusions. In the last chapter of this dissertation, we additionally conducted simulations with more—smaller and different—models, demonstrating that referential games can be used to reveal model-specific strengths and biases. For the simple practical reason that conducting human-based experiments takes time, humans only interacted with Llama-3-70B-Instruct. The evolution of shared referentially grounded vocabularies between humans and LLMs must therefore, for now, be seen as an exciting initial result that needs further empirical support.

Short-term alignment

The methods used in Chapter 6 and Chapter 7 rely on short-term alignment. They rely on the ability to learn from a few in-context examples (Brown et al., 2020) and follow instructions (Ouyang et al., 2022). While practical, such learning is only temporary and has no lasting impact on future behaviour, i.e., the models themselves have no history and their parameters are not updated. This stands in stark contrast to humans, who learn from and *in between* interactions. Participants relied heavily on previous interactions and engaged in mind-reading activities to accommodate partner behaviours. While the ability of LLMs to engage in such mind-reading activities is actively investigated (van Duijn et al., 2023; Kosinski, 2024; Shapira et al., 2024, inter

alia), the point we tried to make is that models should not only accommodate for successful interactions but also learn from this experience for future interactions, as it is only then that language can truly be grounded in experience. The practical solution for this, however, is not straightforward and has not been addressed in this dissertation. A fruitful direction could, for instance, be to carefully and incrementally update the reward model used for RLHF to incorporate the interactive, intentional, situated, and communicative nature of human language learning as was proposed by Beuls and Van Eecke (2024).

Reliance on behavioural probes and prompting

The conclusions drawn in Chapters 4, 6, and 7 rely on *behavioural* observations. While it is pragmatic to attribute meaning to behavioural observations (van Dijk et al., 2023a), it limits what can be concluded concerning the internal working of the employed models. In the case of cross-modal associations, it is unclear precisely what the underlying reasons are for not displaying a bouba-kiki effect. Similarly, while the evolved languages in Chapter 6 and Chapter 7 *show* some compositional structures, we did not touch upon *how* they 'interpret' these languages.

Prompting is also a fragile endeavour that often is not robust across different phrases encompassing the same meaning (e.g. Weber et al., 2023; Hu and Levy, 2023; Hu and Frank, 2024; Giulianelli et al., 2024). Our VLM and LLM chapters are, therefore, also subject to this. In the case of the bouba-kiki effect, we embedded labels into a simple sentence to provide more context. The textual representations of our stimuli in the final chapters were inspired by Galke et al. (2024), and we instructed LLMs to be communicative. In both cases, the results warrant further confirmation. For example, by using multiple *different* prompts designed to assess the same model's ability (as in: Allen et al., 2025; Kouwenhoven et al., 2025). Doing so still relies on behavioural observations, but removes the reliance on prompting, strengthening claims.

8.2.2 Future work

Here, we elaborate on directions that we deem fruitful for future work following the studies that constitute this dissertation.

1. **Hybrid experiments** – In Chapters 6 and 7 of this dissertation, we employed LLMs as approximates of mature language learners as if they were subjects with cognitive abilities (Binz and Schulz, 2023; Pellert et al., 2024; Binz and Schulz, 2024; Löhn et al., 2024). In doing so, we showed in Chapter 7 that humans and LLMs can effectively collaborate despite having different mechanisms, underscoring the importance of interactions. These findings enable comparisons between entity-specific and hybrid solutions, revealing the strengths and weaknesses of both, which can potentially result in symbiotic systems that leverage the capabilities of both to achieve more than could be achieved by a single entity. Importantly, the approach employed in this chapter opens up a range of possibilities for future research in which humans and machines collaborate actively on various tasks,

both within and outside the domain of language evolution. In particular, for domains such as education, therapy and healthcare, where it is vital that humans and machines adapt to the situation at hand to ensure successful and productive interaction (Ostrand and Berger, 2024).

2. **Beyond behavioural research** – There is an interesting dichotomy between open-source models that push the boundaries of obtaining small-scale models with strong linguistic capabilities through data-efficient training, and ever-growing closed-source language models challenging each other to be the 'best' model out there. While most scholars can only engage in behavioural studies with closed-source models, contrarily open-source variants such as OLMo2 (Walsh et al., 2025), BLOOMZ (Muennighoff et al., 2023), and Pythia (Biderman et al., 2023) enable investigating *what* is going on inside these models. Doing so provides a clearer picture of the relations that are learned and why these models perform well or not on specific tasks (e.g. Darcet et al., 2024; Skean et al., 2025). Our investigation of cross-modal associations in VLMs (Chapter 4) could, for example, be complemented by inspecting visual attention patterns to enhance our understanding of which visual features steer predictions. The signal-meaning mappings learned in Chapters 6 and 7 can be investigated by visualising attention patterns and token log probabilities. Furthermore, open-source models enable in-depth analysis of the impact of different fine-tuning steps, as demonstrated by Peeperkorn et al. (2025), who showed that fine-tuning has a negative effect on LLM output diversity.
3. **Language acquisition in LLMs** – By now, it is clear that LLMs have remarkable linguistic abilities. An increasingly growing body of research investigates *whether*, *when*, and *why* language models have these abilities (Misra and Mahowald, 2024; Chen et al., 2024; Kallini et al., 2024; Xu et al., 2025, inter alia). Such careful manipulation and inspection of training setups allow us to compare LLM language acquisition to how children acquire languages. For example, by training models on ecologically valid data (Warstadt et al., 2023), or by exploring to what extent child narratives aid language learning (van Dijk et al., 2023b). Together, these contribute to our understanding of how LLMs acquire languages. The degree to which these patterns are similar to humans, in turn, informs whether training objectives should be adjusted and may promote more human-like language learning in LLMs.
4. **Group communication** – An important argument in this dissertation is that interactions should have a more prominent role in language learning setups. However, the chapters in this dissertation only include at most two interlocutors, while humans are deeply embedded in culture and surrounded by others. Clearly, languages are a product of a diverse group of interacting minds, offering numerous opportunities for future research. In the realm of reinforcement learning, the NeLLCom-X framework (Lian et al., 2024) can be used to investigate the influence of learning and group dynamics on language universals. Furthermore, groups of interacting LLM-augmented agents can be used to

simulate cultural evolution involving more complex behaviours (e.g. Park et al., 2023; Perez et al., 2024) and unravel the dynamics of machine-generated cultural evolution (Brinkmann et al., 2023). From a language evolution perspective, the experiment in Chapter 7 can be extended to involve various group compositions of LLMs and humans to empirically test whether machines can participate in creating shared languages in group settings. This would be especially interesting since, at the group level, languages tend to adapt to preferences at the individual level (Josserand et al., 2024).

A

Kiki or Bouba?

A.1 Full set of images

This appendix presents the full set of images with visual shapes that were used in the experiments. Besides the original image pair from Köhler (1929, 1947) which was shown in Figure 4.1, we used four image pairs from Maurer et al. (2006), displayed in Figure A.1, four from Westbury (2005), displayed in Figure A.2, and 8 additional pairs we newly generated using a method inspired by the one described by Nielsen and Rendall (2013), displayed in Figure A.3. For each image pair, the Curved version is displayed on the left and the Jagged version on the right.

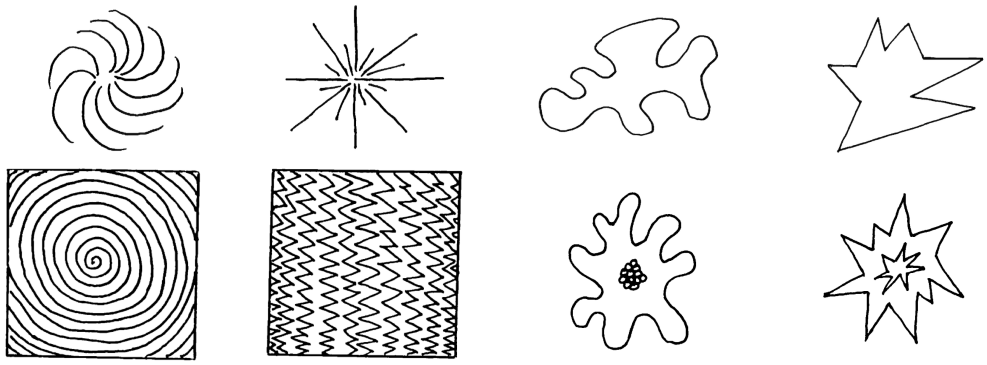


Figure A.1: Images from (Maurer et al., 2006)

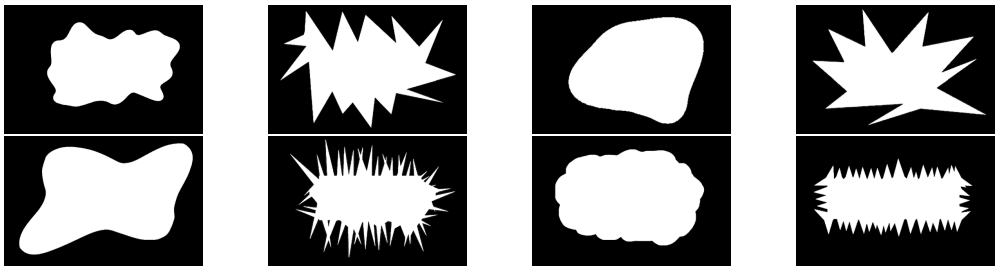


Figure A.2: Images from (Westbury, 2005)

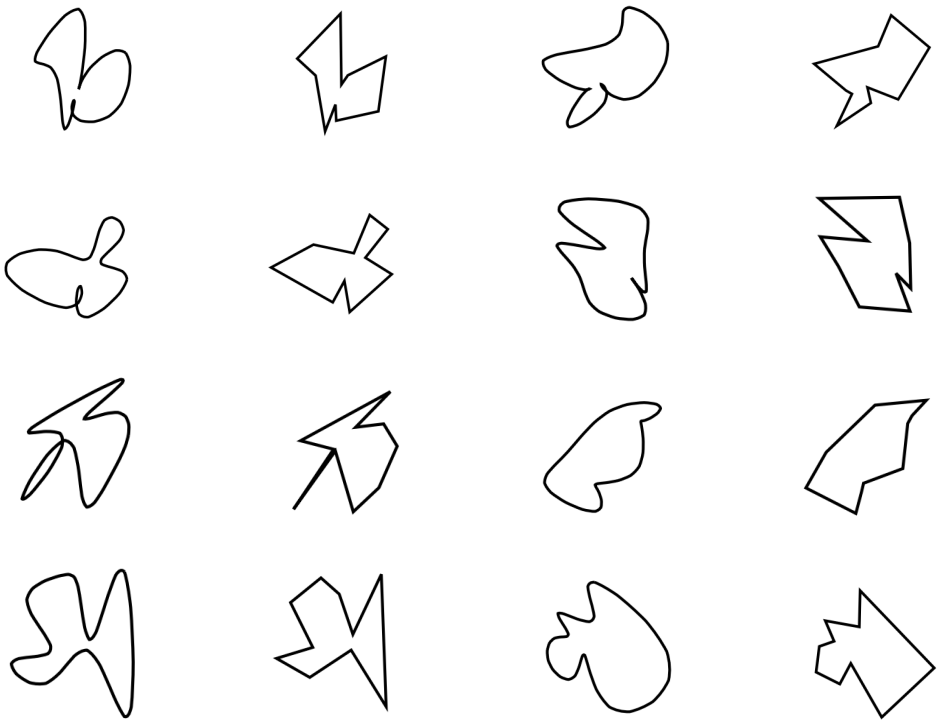


Figure A.3: Newly generated images

A.2 GPT-4o prompting

Image-label matching is not directly possible for GPT-4o since the probabilities of the input tokens cannot be accessed. We therefore prompt (Prompt A.2.1) this model, with the temperature being 0.0, to generate a syllable or pseudoword given an image and use the log probabilities of the generated tokens to calculate the probability for a label conditioned on an image. Just like in

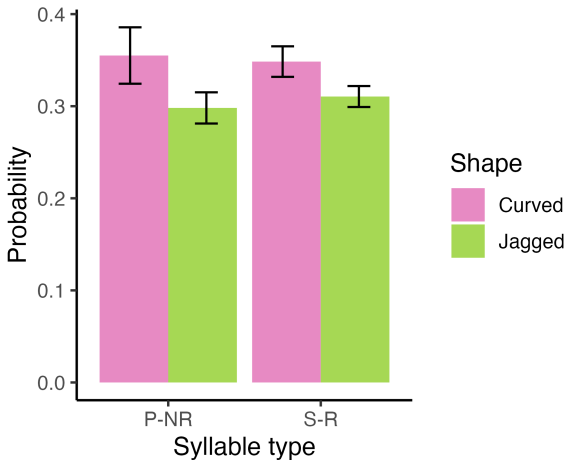


Figure A.4: Probability scores for GPT-4o when forced to generate a pseudoword for each image twice, once by combining two Jagged-associated syllables, and once with only the Curved-associated syllables as options.

the sentence setup used in the other models, our interest lies not primarily in the variability that may arise from using different prompts but rather focuses on the influence of the image on the predictions by using a simple and effective prompt that is identical for each image. Doing so allows us to use the resulting probabilities as a gauge for the models' preference of a label for a given image.

```
You are given an image for which you need to assign a label. Use {one
/two} of the following labels: {possible_labels}. Only respond with
the label.
```

Prompt A.2.1: The exact prompt used to obtain GPT-4o probabilities. $possible\_labels$ corresponds to the syllables of interest.

A.3 GPT-4o pseudoword probabilities

In Section 4.4.4 we describe the results of an experiment in which we asked GPT-4o to generate a pseudoword for each image twice, once when given only the set of Jagged-associated syllable options, and once with only the Curved-associated syllables as options. Figure A.4 shows the probabilities associated with these generated pseudowords. As concluded in the main text, no evidence for a preference to match P-NR syllables with Jagged shapes and S-R syllables with Curved shapes was found.

B

The Curious Case of Representational Alignment

B.1 Channel capacity

To test to what degree communicative success, *TopSim*, and representational alignment are confounded with the communication channel capacity, we ran simulations altering the vocabulary size ($V = \{3, 5, 10, 20, 40, 50, 100\}$) and message length ($L = \{2, 3, 5, 10, 50, 100\}$) resulting in 42 parameter settings per loss type. The parameters and seeds used to run the experiments in the main paper are displayed in Table B.1.

Overall, performance is relatively independent of the chosen configuration, but vocabulary size influences success more than message length (Figure B.1). The hyperparameters that resulted in the best validation accuracy (i.e., generalisation; Chaabouni et al., 2022) for the standard ce setup were $V = 40$ and $L = 2$. These parameters are used to produce the results

Parameter	**Value**
Batch size	32
Optimiser	Adam
Learning Rate (S & L)	0.01 & 0.001
Vocabulary size (V)	40
Message length (L)	2
Hidden size (S & L)	768 & 768
Embedding size	50
Listener cosine temperature	0.1
Seeds	16,22,41,56,67,77,14,78,99,23,82,40,51,37,62

Table B.1: Best-performing parameters resulting from the parameter sweep.

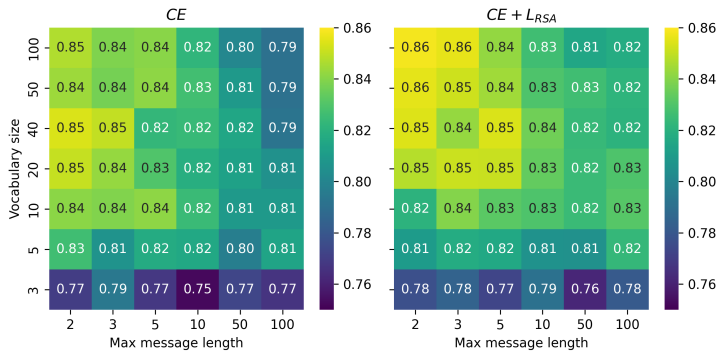


Figure B.1: The validation accuracy as a dependent factor of the vocabulary size and maximum message length. Values are averages across 15 seeds.

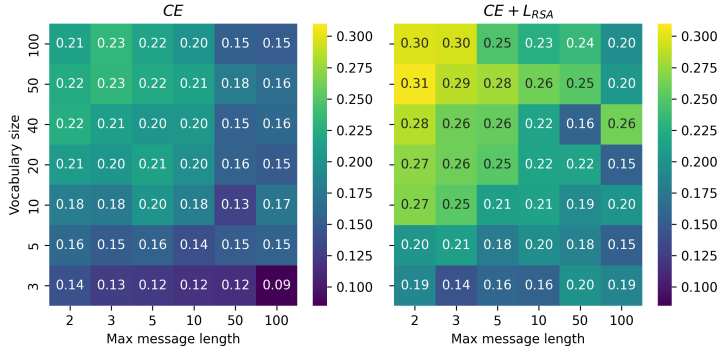


Figure B.2: *TopSim* as a dependent factor of the vocabulary size and maximum message length. Values are averages across 15 seeds.

in the main paper. Contra expectations, the vocabulary size also influenced *TopSim* more than message length. It, especially in the case of $ce + L_{RSA}$, is higher when messages are shorter but have access to a larger vocabulary (Figure B.2).

Figure B.3 shows that, regardless of the vocabulary capacity, inter-agent alignment (RSA_{sl}) increases while image-agent alignment (RSA_{si} and RSA_{li}) decreases with the ce loss. Interestingly, RSA_{sl} is agnostic to capacity but a larger vocabulary size, not message length, reduces the degree of drifting away from the input. We hypothesise this to result from lower pressure to compress rich continuous embeddings into smaller discrete vocabulary embeddings.

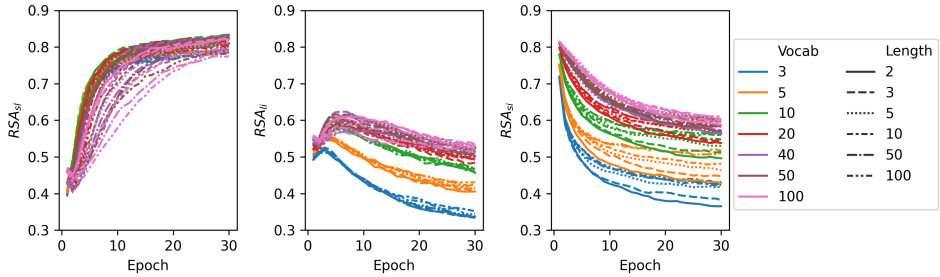


Figure B.3: Representational alignment metrics averaged over 15 simulations with the standard ce loss. Regardless of channel capacity, representational alignment always occurs while losing relation to the input.

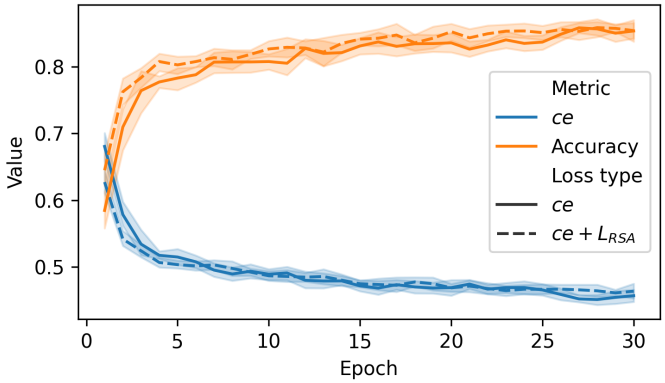


Figure B.4: Learning curves (accuracy) and cross-entropy loss (ce) for both loss settings. There is virtually no effect of the auxiliary term L_{RSA} on the cross entropy loss or communicative success.

B.2 Interaction between the alignment term and cross-entropy

To ensure that there is no impact of the alignment penalty on the pressure for communicative success, we ablated the L_{RSA} term of our proposed loss function and found that both, communicative success and ce are not affected by the alignment penalty (Figure B.4). Corroborating that only the ce term provides pressure for successful communication (Section 5.5.4).

B.3 Pre-trained vision modules

Although it is in principle possible to train the vision module of the agents from scratch (Dessi et al., 2021), in our work, agents' perception stems from a pre-trained vision-language model. Although there is reason to believe that DinoV2 embeddings capture high-level, conceptual image features useful for discriminating image pairs (Oquab et al., 2024), we assessed the degree to which the alignment problem occurs for different pre-trained models despite encoding the same objects. We ran additional simulations using image features obtained from ResNet (He et al., 2016) and CLIP (Radford et al., 2021) for 6 different parameter settings with the ce loss function. Here we used the parameters that resulted in the best, worst, mean, and quantile validation performance from the parameter sweep in Section B.1 (see Table B.2), and a sensible setup with $V = 10$ and $L = 5$.

Figure B.5 shows clearly that inter-agent alignment *increases* while agent-image alignment *decreases* for all models. In addition to the similar results reported by Bouchacourt and Baroni (2018) for VGG ConvNet embeddings, both 4096 and 1000 layers, our results confirm that the problem is agnostic to the input embeddings. Interestingly, agent representations drift most for CLIP embeddings. Nevertheless, the agents still develop a successful communication strategy, indicating that out-of-the-box CLIP embeddings are the least useful for agents in finding a (non-grounded) solution. No such differences are seen when the agents are trained with the additional alignment penalty term, inter-agent and image-agent alignment remain high for all models.

Message length (L)	Vocabulary Size (V)	Vision
2	40	DinoV2 CLIP ResNet
3	10	
5	5	
5	10	
10	3	
50	100	

Table B.2: The parameters for running additional simulations with CLIP and ResNet to assess the robustness of our results. Each combination was run for 15 different seeds. Note: results for the DinoV2 simulations are from the sweep.

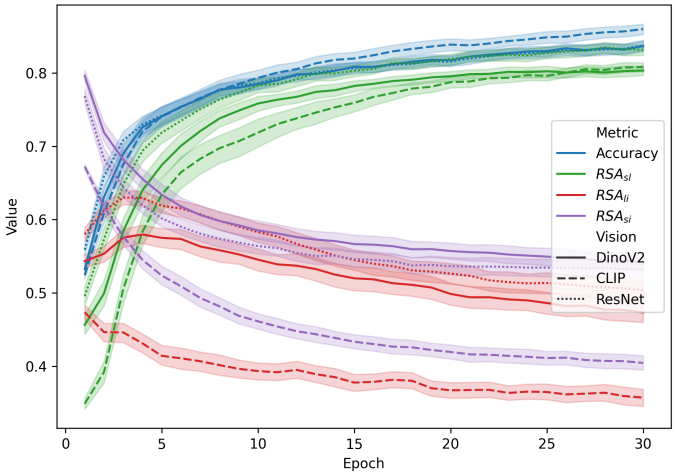


Figure B.5: Learning curves (accuracy) and *RSA* metrics for different vision models averaged over 6 parameter settings with 15 seeds each. The representational alignment problem always occurs. Line style corresponds to the vision module used to obtain image embeddings and colour indicates the metric. Areas indicate the 95% confidence intervals.

C

Searching for Structure

C.1 Additional results iterated learning

Figure C.1 shows how communicative success evolved across generations of language learners and users. There is no clear increase in communicative success. Figure C.2 shows that the average structure as measured by *TopSim* increases somewhat across generations, although not significantly. Interestingly, some generations display decreasing *TopSim*, indicating a loss of structure. This is reminiscent of findings in human iterated learning experiments, showing that processes of interaction and transmission sometimes generate fully systematic, compositional languages, but can also result in systems that lack structure entirely (Verhoef et al., 2022). In the case of *Ngram* diversity, we observe a decrease in the unique Ngrams produced, which indicates the languages re-use parts of signals more in later generations (Figure C.3).

C.2 Prompts

Our agents act based on prompts and system instructions. These are designed to be maximally close to the classical experimental setup and formatted similar to Galke et al. (2024). Prompt completion is used for labelling stimuli during the labelling and communication block. For the guessing task, we prefill the prompt with each possible word or distractor and pick the option with the highest probability. See the full prompts for labelling and guessing in Prompt C.2.1. Speaking during communication involved plain prompt completion (Prompt C.2.2). Discrimination during communication was done by prefilling the distractors attributes (Prompt C.2.3).

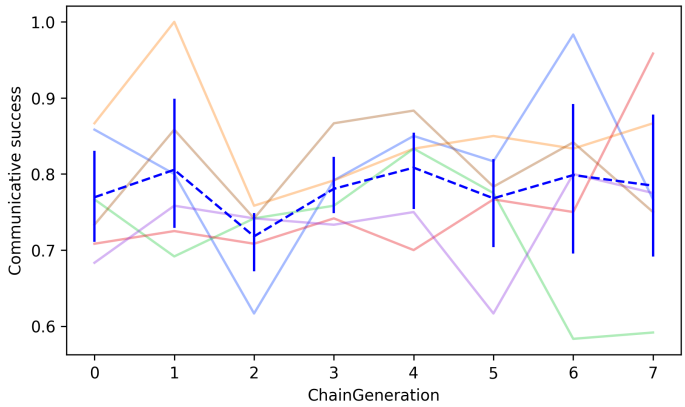


Figure C.1: The average communicative success across rounds for each generation. Each line represents a chain, and the dashed blue line indicates the average, with bars denoting the 95% confidence interval. See Table 6.2 for the descriptives of *PercCom*.

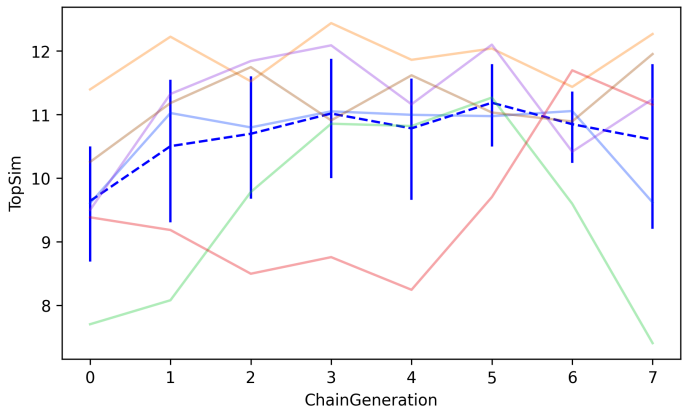


Figure C.2: The evolution of *TopSim* on the words produced in the testing block. Each line represents a chain, and the dashed blue line indicates the average, with bars denoting the 95% confidence interval. See Table 6.2 for the descriptives of *TopSim*.

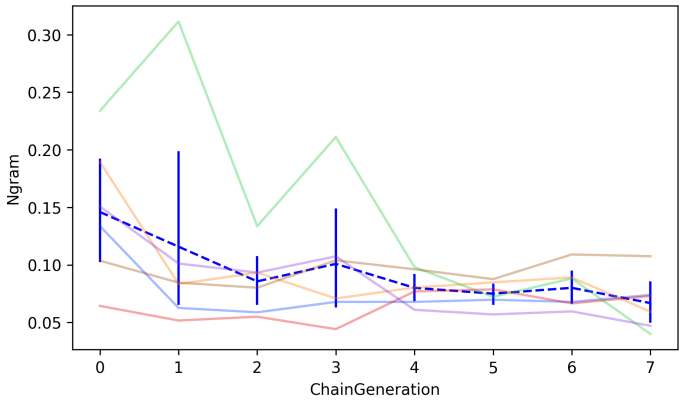


Figure C.3: The evolution of *Ngram* on the words produced in the testing block. Each line represents a chain, and the dashed blue line indicates the average, with bars denoting the 95% confidence interval. See Table 6.2 for the descriptives of *Ngram*.

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|> You are a
 language learner who has to learn an artificial language with words
and their corresponding features. Your task is to complete the
vocabulary by generating a word that describes the last item. Only
respond with the word.<|eot_id|><|start_header_id|>user<|
end_header_id|>

{'shape':2,'colour':'orange','amount':1,'word':'giniwite'}
{'shape':3,'colour':'green','amount':1,'word':'ginisu'}
{'shape':1,'colour':'orange','amount':2,'word':'pinisugi'}
{'shape':3,'colour':'green','amount':3,'word':'sutepi'}
{'shape':2,'colour':'orange','amount':2,'word':'winisu'}
{'shape':3,'colour':'orange','amount':1,'word':'niwi'}
{'shape':1,'colour':'blue','amount':2,'word':'sutuwite'}
{'shape':1,'colour':'blue','amount':3,'word':'tupitene'}
{'shape':3,'colour':'blue','amount':1,'word':'wipinepi'}
{'shape':2,'colour':'orange','amount':3,'word':'gigi'}
{'shape':1,'colour':'green','amount':2,'word':'nite'}
{'shape':3,'colour':'blue','amount':3,'word':'wite'}
{'shape':1,'colour':'green','amount':3,'word':'sune'}
{'shape':2,'colour':'blue','amount':2,'word':'ninene'}
{'shape':2,'colour':'green','amount':1,'word':'tusetetu'}
{'shape':1,'colour':'green','amount':3,'word':'<|eot_id|><|
start_header_id|>assistant<|end_header_id|>
[COMPLETION OR PREFFILED]
```

Prompt C.2.1: Completion Prompt used for labelling and guessing.

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|> You are a
 language learner who has to learn an artificial language with words
and their corresponding features. Your task is to generate a word
such that your communication partner can guess the correct meaning of
 the word. Communicative success is important. Only respond with the
word.<|eot_id|><|start_header_id|>user<|end_header_id|>

{'shape':1,'colour':'green','amount':3,'word':'sutupitite','
communicativeSuccess':1}
{'shape':2,'colour':'orange','amount':2,'word':'ginupepi','
communicativeSuccess':1}
{'shape':1,'colour':'orange','amount':2,'word':'sutupepi','
communicativeSuccess':1}
{'shape':1,'colour':'green','amount':2,'word':'sutupepi','
communicativeSuccess':0}
{'shape':2,'colour':'orange','amount':1,'word':'ginisu','
communicativeSuccess':1}
{'shape':2,'colour':'orange','amount':3,'word':'ginupitite',
'communicativeSuccess':1}
{'shape':3,'colour':'green','amount':1,'word':'wipisu','
communicativeSuccess':0}
{'shape':2,'colour':'green','amount':1,'word':'ginisu','
communicativeSuccess':1}
{'shape':1,'colour':'blue','amount':2,'word':'sunupepi','
communicativeSuccess':1}
{'shape':3,'colour':'green','amount':3,'word':'wipipitite','
communicativeSuccess':1}
{'shape':3,'colour':'orange','amount':1,'word':'wipisu','
communicativeSuccess':0}
{'shape':1,'colour':'blue','amount':3,'word':'sunupitite','
communicativeSuccess':1}
{'shape':3,'colour':'blue','amount':3,'word':'wipipitite','
communicativeSuccess':1}
{'shape':3,'colour':'blue','amount':1,'word':'wipisu','
communicativeSuccess':1}
{'shape':2,'colour':'blue','amount':2,'word':'<|eot_id|><|
start_header_id|>assistant<|end_header_id|>
[COMPLETION]
```

Prompt C.2.2: Speaking Prompt during communication.

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|> You are a
 language learner who has to learn an artificial language with words
and their corresponding features. Your task is to complete the
vocabulary by interpreting the intended meaning of the word generated
 by your communication partner. Communicative success is important.
Only respond with the complete last item.<|eot_id|><|start_header_id
|>user<|end_header_id|>

{'word':'wipipitite','shape':3,'colour':'blue','amount':3,'
communicativeSuccess':1}
{'word':'wipisu','shape':3,'colour':'orange','amount':1,'
communicativeSuccess':0}
{'word':'wipisu','shape':3,'colour':'green','amount':1,'
communicativeSuccess':0}
{'word':'sutupepi','shape':1,'colour':'orange','amount':2,'
communicativeSuccess':1}
{'word':'ginupepi','shape':2,'colour':'orange','amount':2,'
communicativeSuccess':1}
{'word':'sutupitite','shape':1,'colour':'green','amount':3,'
communicativeSuccess':1}
{'word':'wipipitite','shape':3,'colour':'green','amount':3,'
communicativeSuccess':1}
{'word':'wipisu','shape':3,'colour':'blue','amount':1,'
communicativeSuccess':1}
{'word':'ginisu','shape':2,'colour':'green','amount':1,'
communicativeSuccess':1}
{'word':'ginisu','shape':2,'colour':'orange','amount':1,'
communicativeSuccess':1}
{'word':'sunupepi','shape':1,'colour':'blue','amount':2,'
communicativeSuccess':1}
{'word':'sutupepi','shape':1,'colour':'green','amount':2,'
communicativeSuccess':0}
{'word':'sunupitite','shape':1,'colour':'blue','amount':3,'
communicativeSuccess':1}
{'word':'ginupitite','shape':2,'colour':'orange','amount':3,
'communicativeSuccess':1}
{'word':'ginupepi','shape':'<|eot_id|><|start_header_id|>assistant<|
end_header_id|>
[PREFILLED WITH DISTRACTOR ATTRIBUTES]
```

Prompt C.2.3: Guessing Prompt during communication.

D

Shaping Shared Languages

D.1 Prompts

The agents in our experiment act based on prompts and system instructions which are identical to those used in Chapter 6. These were designed to be maximally close to the classical experimental setup and formatted similar to Galke et al. (2024). During the labelling and guessing block, we use the completion Prompt D.1.1. In the labelling block, we simply ask the model to provide a completion. In the case of the guessing block, we prefill the word and pick the signal with the highest probability. See the full prompts for labelling and guessing (Prompt D.1.1), speaking (Prompt D.1.2), and discrimination (Prompt D.1.3) below.

As explained in the main body of our paper, we update the agent-specific vocabulary after each label prediction. This allows the vocabularies of signal-meaning mappings to evolve during the simulation. This entails that the prompts are also slightly different after each interaction or prediction. Moreover, given the observed bias for primacy and recency Liu et al. (2024) in LLMs, we shuffle the vocabulary before creating prompts to account for unwanted ordering effects.

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|> You are a
 language learner who has to learn an artificial language with words
and their corresponding features. Your task is to complete the
vocabulary by generating a word that describes the last item. Only
respond with the word.<|eot_id|><|start_header_id|>user<|
end_header_id|>

{'shape':3,'colour':'green','amount':1,'word':'tego'}
{'shape':1,'colour':'green','amount':3,'word':'wananima'}
{'shape':2,'colour':'blue','amount':3,'word':'wumawaka'}
{'shape':2,'colour':'green','amount':3,'word':'mafa'}
{'shape':3,'colour':'orange','amount':2,'word':'wawa'}
{'shape':1,'colour':'orange','amount':1,'word':'gofa'}
{'shape':1,'colour':'blue','amount':1,'word':'maka'}
{'shape':3,'colour':'blue','amount':1,'word':'kama'}
{'shape':3,'colour':'blue','amount':3,'word':'mawa'}
{'shape':2,'colour':'orange','amount':2,'word':'nawa'}
{'shape':2,'colour':'blue','amount':1,'word':'kaka'}
{'shape':3,'colour':'green','amount':2,'word':'matefama'}
{'shape':1,'colour':'orange','amount':3,'word':'kagonigo'}
{'shape':2,'colour':'green','amount':2,'word':'nimaniwu'}
{'shape':1,'colour':'orange','amount':2,'word':'wago'}
{'shape':1,'colour':'orange','amount':2,'word':'<|eot_id|><|
start_header_id|>assistant<|end_header_id|>
[COMPLETION OR PREFFILED]
```

Prompt D.1.1: An example completion prompt used for labelling and guessing.

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|> You are a
 language learner who has to learn an artificial language with words
and their corresponding features. Your task is to generate a word
such that your communication partner can guess the correct meaning of
 the word. Communicative success is important. Only respond with the
word.<|eot_id|><|start_header_id|>user<|end_header_id|>

{'shape':1,'colour':'green','amount':3,'word':'gofamama','
communicativeSuccess':1}
{'shape':2,'colour':'orange','amount':2,'word':'kakafa','
communicativeSuccess':1}
{'shape':1,'colour':'orange','amount':2,'word':'gofama','
communicativeSuccess':1}
{'shape':2,'colour':'blue','amount':1,'word':'kaka','
communicativeSuccess':1}
{'shape':2,'colour':'green','amount':2,'word':'kakafa','
communicativeSuccess':1}
{'shape':3,'colour':'green','amount':2,'word':'tegoma','
communicativeSuccess':1}
{'shape':1,'colour':'orange','amount':1,'word':'go','
communicativeSuccess':1}
{'shape':3,'colour':'green','amount':1,'word':'tega','
communicativeSuccess':1}
{'shape':2,'colour':'blue','amount':3,'word':'kakamama','
communicativeSuccess':1}
{'shape':3,'colour':'blue','amount':1,'word':'tego','
communicativeSuccess':1}
{'shape':1,'colour':'orange','amount':3,'word':'gofamama','
communicativeSuccess':1}
{'shape':2,'colour':'green','amount':3,'word':'kakamama','
communicativeSuccess':1}
{'shape':3,'colour':'orange','amount':2,'word':'tegoma','
communicativeSuccess':0}
{'shape':3,'colour':'blue','amount':3,'word':'tegomama','
communicativeSuccess':1}
{'shape':1,'colour':'blue','amount':1,'word':'<|eot_id|><|
start_header_id|>assistant<|end_header_id|>
[COMPLETION]
```

Prompt D.1.2: An example speaking prompt during communication. In this particular case, the speaker produced the label *'goa'* which was correctly interpreted by the human listener.

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|> You are a
 language learner who has to learn an artificial language with words
and their corresponding features. Your task is to complete the
vocabulary by interpreting the intended meaning of the word generated
 by your communication partner. Communicative success is important.
Only respond with the complete last item.<|eot_id|><|start_header_id
|>user<|end_header_id|>

{'word':'kakafa','shape':2,'colour':'orange','amount':2,'
communicativeSuccess':1}
{'word':'go','shape':1,'colour':'orange','amount':1,'
communicativeSuccess':1}
{'word':'kakafa','shape':2,'colour':'green','amount':2,'
communicativeSuccess':1}
{'word':'goa','shape':1,'colour':'blue','amount':1,'
communicativeSuccess':1}
{'word':'kakamama','shape':2,'colour':'green','amount':3,'
communicativeSuccess':1}
{'word':'tego','shape':3,'colour':'blue','amount':1,'
communicativeSuccess':1}
{'word':'kaka','shape':2,'colour':'blue','amount':1,'
communicativeSuccess':1}
{'word':'tegoma','shape':3,'colour':'orange','amount':2,'
communicativeSuccess':0}
{'word':'gofamama','shape':1,'colour':'green','amount':3,'
communicativeSuccess':1}
{'word':'kakamama','shape':2,'colour':'blue','amount':3,'
communicativeSuccess':1}
{'word':'gofama','shape':1,'colour':'orange','amount':2,'
communicativeSuccess':1}
{'word':'tega','shape':3,'colour':'green','amount':1,'
communicativeSuccess':1}
{'word':'tegomama','shape':3,'colour':'blue','amount':3,'
communicativeSuccess':1}
{'word':'gofamama','shape':1,'colour':'orange','amount':3,'
communicativeSuccess':1}
{'word':'tegama','shape':'<|eot_id|><|start_header_id|>assistant<|
end_header_id|>
[PREFILLED WITH DISTRACTOR ATTRIBUTES]
```

Prompt D.1.3: An example guessing prompt during communication. Here the human speaker has produced the label *'tegama'* which was correctly interpreted by the listener.

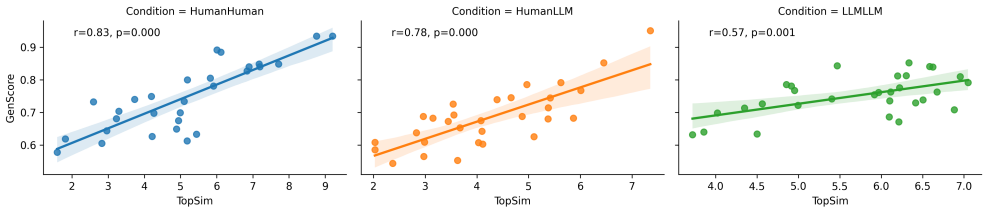


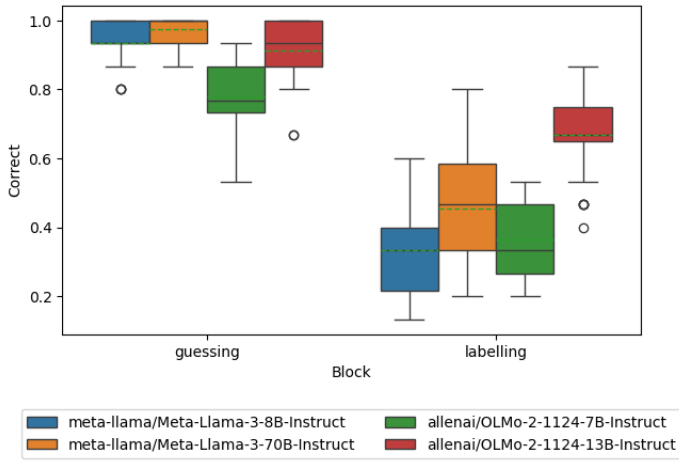
Figure D.1: Across conditions, generalisation to novel stimuli was more consistent with known samples when the labels produced during the last round of communication showed a higher degree of *TopSim*.

D.2 Generalisation to novel stimuli

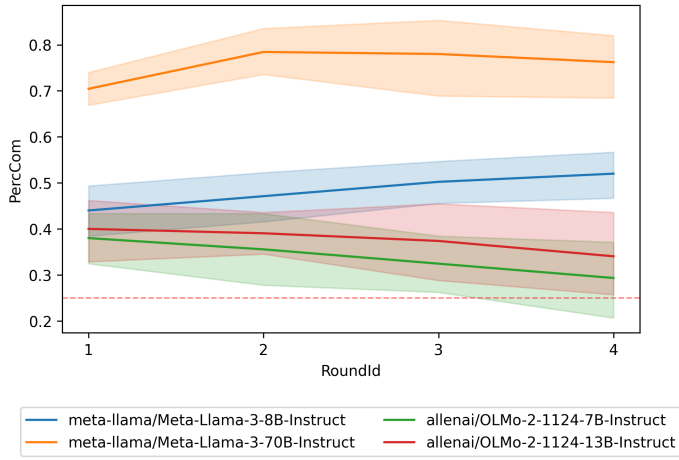
In the testing block of our experiments, we examined whether the participants and LLMs could generalise to novel stimuli. Figure D.1 shows the relationship between *TopSim* and *GenScore* that was modelled by a mixed effects model in Chapter 7. It reveals clearly that generalisation to novel items happened more consistently when the vocabulary evolved to be more structured.

D.3 Testing additional LLMs

To extend the findings of Chapter 6, we ran additional simulations with three different LLMs, Llama-3-8B (Llama Team, 2024), OLMo-2 7B, and OLMo-2 13B (Walsh et al., 2025) using the same 15 seeds. Figure D.2a shows that learning the artificial languages is also possible for smaller and different LLMs. Interestingly, out of all models, agents comprised of a OLMo-2 13B model perform best during the labelling task. While agents with OLMo-2 13B can also communicate reliably above change performance ($t(13) = 1.96, p = .036$), they struggle much more (*PercCom* $\approx$ 35%, Figure D.2b, chance performance amounts to 25%). Moreover, we observe that compared to their larger versions, smaller models struggle more. Llama-3 8B achieves $\approx$ 50% of successful communication and OLMo-2 8B only 30%. The latter is not significantly above chance ($t(14) = 1.04, p = .158$). We take these results as additional evidence that our setup can be used to discover LLM-specific constraints and indicating that larger LLMs benefit more from instruction following Lou et al. (2024). We leave the precise dynamics of both to future work.



(a) Performance on the guessing and labelling task. The newly tested LLMs can learn the languages equally well or better (e.g. OLMO-2 13B outperforms Llama-3 70B).



(b) Communicative performance across rounds for different models. Note that the dashed red line indicates chance performance.

Figure D.2: The results of learning and using languages for two different LLMs with two different sizes. Figure D.2a shows the degree to which LLMs can learn the languages and Figure D.2b shows how well these models can use the language during communication.

Bibliography

Abramova, E. and Fernández, R. (2016). Questioning arbitrariness in language: a data-driven study of conventional iconicity. In Knight, K., Nenkova, A., and Rambow, O., editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 343–352, San Diego, California. Association for Computational Linguistics.

Acerbi, A. and Stubbersfield, J. M. (2023). Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44):e2313790120.

Agrawal, A., Batra, D., Parikh, D., and Kembhavi, A. (2018). Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980.

Aher, G. V., Arriaga, R. I., and Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR.

Allen, K., Dasgupta, I., Kosoy, E., and Lampinen, A. K. (2025). The in-context inductive biases of vision-language models differ across modalities.

Alper, M. and Averbuch-Elor, H. (2023). Kiki or bouba? sound symbolism in vision-and-language models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 78347–78359. Curran Associates, Inc.

Alper, M., Fiman, M., and Averbuch-Elor, H. (2023). Is BERT blind? exploring the effect of vision-and-language pretraining on visual language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6778–6788.

Arnon, I. and Kirby, S. (2024). Cultural evolution creates the statistical structure of language. *Scientific Reports*, 14(1):5255.

Auersperger, M. and Pecina, P. (2022). Defending compositionality in emergent languages. In Ippolito, D., Li, L. H., Pacheco, M. L., Chen, D., and Xue, N., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 285–291, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Baker, N., Lu, H., Erlikhman, G., and Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLOS Computational Biology*, 14(12):1–43.

Ben Zion, R., Carmeli, B., Paradise, O., and Belinkov, Y. (2024). Semantics and spatiality of emergent communication. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C., editors, *Advances in Neural Information Processing Systems*, volume 37, pages 110156–110196. Curran Associates, Inc.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Bergen, B. K. (2004). The psychological reality of phonaesthemes. *Language*, 80(2):290–311.

Beuls, K. and Van Eecke, P. (2024). Humans learn language from situated communicative interactions. what about machines? *Computational Linguistics*, 50(3):1277–1311.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., and Van Der Wal, O. (2023). Pythia: A suite for analyzing large language models across training and scaling. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., Éltető, N., Griffiths, T. L., Haridi, S., Jagadish, A. K., Ji-An, L., Kipnis, A., Kumar, S., Ludwig, T., Mathony, M., Mattar, M., Modirshanechi, A., Nath, S. S., Peterson, J. C., Rmus, M., Russek, E. M., Saanum, T., Scharfenberg, N., Schubert, J. A., Buschoff, L. M. S., Singhi, N., Sui, X., Thalmann, M., Theis, F., Truong, V., Udandarao, V., Voudouris, K., Wilson, R., Witte, K., Wu, S., Wulff, D., Xiong, H., and Schulz, E. (2025). A foundation model to predict and capture human cognition. *Nature*.

Binz, M. and Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Binz, M. and Schulz, E. (2024). Turning large language models into cognitive models. In *The Twelfth International Conference on Learning Representations*.

Bisazza, A., Üstün, A., and Sportel, S. (2021). On the difficulty of translating free-order case-marking languages. *Transactions of the Association for Computational Linguistics*, 9:1233–1248.

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N., and Turian, J. (2020). Experience grounds language. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., and Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, 113(39):10818–10823.

Blondel, M., Teboul, O., Berthet, Q., and Djolonga, J. (2020). Fast differentiable sorting and ranking. In *International Conference on Machine Learning*, pages 950–959. PMLR.

Bottini, R., Barilari, M., and Collignon, O. (2019). Sound symbolism in sighted and blind. the role of vision and orthography in sound-shape correspondences. *Cognition*, 185:62–70.

Bouchacourt, D. and Baroni, M. (2018). How agents see things: On visual representations in an emergent language game. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 981–985, Brussels, Belgium. Association for Computational Linguistics.

Boyd, R., Richerson, P. J., et al. (1996). Why culture is common, but cultural evolution is rare. In *Proceedings-british academy*, volume 88, pages 77–94. Oxford University Press Inc.

Brandizzi, N. (2023). Toward more human-like ai communication: A review of emergent communication research. *IEEE Access*, 11:142317–142340.

Brandizzi, N. and Iocchi, L. (2022). Emergent communication in human-machine games. In *Emergent Communication Workshop at ICLR 2022*.

Brighton, H. and Kirby, S. (2006). Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artificial Life*, 12(2):229–242.

Brinkmann, L., Baumann, F., Bonnefon, J.-F., Derex, M., Müller, T. F., Nussberger, A.-M., Czaplicka, A., Acerbi, A., Griffiths, T. L., Henrich, J., et al. (2023). Machine culture. *Nature Human Behaviour*, 7(11):1855–1868.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, 100(5):1–54.

Carr, J. W., Smith, K., Cornish, H., and Kirby, S. (2017). The cultural evolution of structured languages in an open-ended, continuous world. *Cognitive Science*, 41(4):892–923.

Chaabouni, R., Kharitonov, E., Bouchacourt, D., Dupoux, E., and Baroni, M. (2020). Compositionality and generalization in emergent languages. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4427–4442, Online. Association for Computational Linguistics.

Chaabouni, R., Kharitonov, E., Dupoux, E., and Baroni, M. (2019a). Anti-efficient encoding in emergent communication. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Chaabouni, R., Kharitonov, E., Lazaric, A., Dupoux, E., and Baroni, M. (2019b). Word-order biases in deep-agent emergent communication. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5166–5175, Florence, Italy. Association for Computational Linguistics.

Chaabouni, R., Strub, F., Altché, F., Tarassov, E., Tallec, C., Davoodi, E., Mathewson, K. W., Tieleman, O., Lazaridou, A., and Piot, B. (2022). Emergent communication at scale. In *International Conference on Learning Representations*.

Chang, T. A. and Bergen, B. K. (2022). Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, 10:1–16.

Chen, A., Shwartz-Ziv, R., Cho, K., Leavitt, M. L., and Saphra, N. (2024). Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. In *The Twelfth International Conference on Learning Representations*.

Cheng, E., Rita, M., and Poibeau, T. (2023). On the correspondence between compositionality and imitation in emergent neural communication. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12432–12447, Toronto, Canada. Association for Computational Linguistics.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Moschitti, A., Pang, B., and Daelemans, W., editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press.

Chomsky, N. (2023). The false promise of chatgpt. *The New York Times*.

Christiansen, M. and Chater, N. (2022). *The Language Game: How Improvisation Created Language and Changed the World*. Basic Books.

Christiansen, M. H. and Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, 31(5):489–509.

Christiansen, M. H. and Kirby, S. (2003). Language evolution: The hardest problem in science? In *Language Evolution*. Oxford University Press.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204.

Clark, H. and Brennan, S. (1991). *Grounding in Communication*, volume 13, pages 127–149. American Psychological Association.

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

Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). What does BERT look at? an analysis of BERT's attention. In Linzen, T., Chrupała, G., Belinkov, Y., and Hupkes, D., editors, *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Cohen, R. G. and Rosenbaum, D. A. (2004). Where grasps are made reveals how grasps are planned: generation and recall of motor plans. *Experimental Brain Research*, 157(4):486–495.

Conklin, H. and Smith, K. (2023). Compositionality with variation reliably emerges in neural networks. In *The Eleventh International Conference on Learning Representations*.

Contreras Kallens, P. and Christiansen, M. H. (2024). Distributional semantics: Meaning through culture and interaction. *Topics in Cognitive Science*, n/a(n/a).

Contreras Kallens, P., Kristensen-McLachlan, R. D., and Christiansen, M. H. (2023). Large language models demonstrate the potential of statistical learning in language. *Cognitive Science*, 47(3):e13256.

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., and Konkle, T. (2023). What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *bioRxiv*.

Culbertson, J. and Smolensky, P. (2012). A bayesian model of biases in artificial language learning: The case of a word-order universal. *Cognitive Science*, 36(8):1468–1498.

Cuskley, C., Dingemanse, M., Kirby, S., and Van Leeuwen, T. M. (2019). Cross-modal associations and synesthesia: Categorical perception and structure in vowel–color mappings in a large online sample. *Behavior research methods*, 51:1651–1675.

Cuskley, C. and Kirby, S. (2013). Synesthesia, Cross-Modality, and Language Evolution. In *Oxford Handbook of Synesthesia*. Oxford University Press.

Cuskley, C., Simner, J., and Kirby, S. (2017). Phonological and orthographic influences in the bouba–kiki effect. *Psychological research*, 81:119–130.

Ćwiek, A., Fuchs, S., Draxler, C., Asu, E. L., Dediu, D., Hiovain, K., Kawahara, S., Koutalidis, S., Krifka, M., Lippus, P., Lupyan, G., Oh, G. E., Paul, J., Petrone, C., Ridouane, R., Reiter, S., Schümchen, N., Szalontai, Á., Ünal-Logacev, Ö., Zeller, J., Perlman, M., and Winter, B. (2022). The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1841):20200390.

Dagan, G., Hupkes, D., and Bruni, E. (2021). Co-evolution of language and agents in referential games. In Merlo, P., Tiedemann, J., and Tsarfaty, R., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2993–3004, Online. Association for Computational Linguistics.

Darcet, T., Oquab, M., Mairal, J., and Bojanowski, P. (2024). Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*.

Davis, C. P., Morrow, H. M., and Lupyan, G. (2019). What does a horgous look like? nonsense words elicit meaningful drawings. *Cognitive Science*, 43(10):e12791.

De Boer, B. (2000). Self-organization in vowel systems. *Journal of phonetics*, 28(4):441–465.

de Boer, B. (2006). *Computer modelling as a tool for understanding language evolution*, pages 381–406. Springer Netherlands, Dordrecht.

De Kleijn, R., Kachergis, G., and Hommel, B. (2014). Everyday robotic action: lessons from human action control. *Frontiers in Neurorobotics*, 8:13.

de Kleijn, R., Kachergis, G., and Hommel, B. (2018). Predictive movements and human reinforcement learning of sequential action. *Cognitive Science*, 42(S3):783–808.

de Kleijn, R., Sen, D., and Kachergis, G. (2022). A critical period for robust curriculum-based deep reinforcement learning of sequential action in a robot arm. *Topics in Cognitive Science*, 14(2):311–326.

Deacon, T. W. (1997). *The Symbolic Species: The Co-evolution of Language and the Brain*. ISSR Library. W.W. Norton.

DeCaro, M. S., Thomas, R. D., and Beilock, S. L. (2008). Individual differences in category learning: Sometimes less working memory capacity is better than more. *Cognition*, 107(1):284–294.

Dessi, R., Kharitonov, E., and Marco, B. (2021). Interpretable agent communication from scratch (with a generic visual processor emerging on the side). In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26937–26949. Curran Associates, Inc.

Dillion, D., Tandon, N., Gu, Y., and Gray, K. (2023). Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.

Dingemanse, M. (2012). Advances in the cross-linguistic study of ideophones. *Language and Linguistics Compass*, 6(10):654–672.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., and Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*, 19(10):603–615.

Diwan, A., Berry, L., Choi, E., Harwath, D., and Mahowald, K. (2022). Why is winoground hard? investigating failures in visuolinguistic compositionality. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2250, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dubova, M. and Moskvichev, A. (2020). Effects of supervision, population size, and self-play on multi-agent reinforcement learning to communicate. In *The 2020 Conference on Artificial Life*, volume ALIFE 2020: The 2020 Conference on Artificial Life of *Artificial Life Conference Proceedings*, pages 678–686.

Eva, S., Silvia, H., and Dáša, M. (2014). Personal need for structure in relation to language variables. *Procedia - Social and Behavioral Sciences*, 159:665–670. 5th World Conference on Psychology, Counseling and Guidance, WCPCG-2014, 1-3 May 2014, Dubrovnik, Croatia.

Evans, K. K. and Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision*, 10(1):6–6.

Evans, N. and Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5):429–448.

Fay, N., Garrod, S., and 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.

Fedzechkina, M., Jaeger, T. F., and Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, 109(44):17897–17902.

Fort, M. and Schwartz, J.-L. (2022). Resolving the bouba-kiki effect enigma by rooting iconic sound symbolism in physical properties of round and spiky objects. *Scientific reports*, 12(1):19172.

Futrell, R. and Mahowald, K. (2025). How linguistics learned to stop worrying and love the language models.

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

Galantucci, B. and Garrod, S. (2010). Experimental semiotics: A new approach for studying the emergence and the evolution of human communication. *Interaction Studies*, 11(1):1–13.

Galke, L., Ram, Y., and Raviv, L. (2022). Emergent communication for understanding human language evolution: What's missing? In *Emergent Communication Workshop at ICLR 2022*.

Galke, L., Ram, Y., and Raviv, L. (2024). Deep neural networks and humans both benefit from compositional language structure. *Nature Communications*, 15(1):10816.

Galke, L. and Raviv, L. (2024). Emergent communication and learning pressures in language models: a language evolution perspective. *arXiv preprint arXiv:2403.14427*.

Galke, L. and Raviv, L. (2025). Learning and communication pressures in neural networks: Lessons from emergent communication. *Language Development Research*, 5(1):116–143.

Garrod, S., Fay, N., Lee, J., Oberlander, J., and MacLeod, T. (2007). Foundations of representation: Where might graphical symbol systems come from? *Cognitive Science*, 31(6):961–987.

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., and Brendel, W. (2021). Partial success in closing the gap between human and machine vision. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 23885–23899. Curran Associates, Inc.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2019). Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*.

Gemini Team, G. (2024). Gemini: A family of highly capable multimodal models.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., and Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5):389–407.

Giulianelli, M., Malagutti, L., Gastaldi, J. L., DuSell, B., Vieira, T., and Cotterell, R. (2024). On the proper treatment of tokenization in psycholinguistics. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18556–18572, Miami, Florida, USA. Association for Computational Linguistics.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. (2017). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Griffiths, T. L. and Kalish, M. L. (2007a). Language evolution by iterated learning with bayesian agents. *Cognitive Science*, 31(3):441–480.

Griffiths, T. L. and Kalish, M. L. (2007b). Language evolution by iterated learning with bayesian agents. *Cognitive Science*, 31(3):441–480.

Griffiths, T. L., Zhu, J.-Q., Grant, E., and McCoy, R. T. (2024). Bayes in the age of intelligent machines. *Current Directions in Psychological Science*, 33(5):283–291.

Guo, S., Ren, Y., Havrylov, S., Frank, S., Titov, I., and Smith, K. (2019). The emergence of compositional languages for numeric concepts through iterated learning in neural agents.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Havrylov, S. and Titov, I. (2017). Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, page 2146–2156. Curran Associates, Inc.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2021). Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Hermann, K., Chen, T., and Kornblith, S. (2020). The origins and prevalence of texture bias in convolutional neural networks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19000–19015. Curran Associates, Inc.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Hockett, C. F. (1959). Animal "languages" and human language. *Human Biology*, 31(1):32–39.

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

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2020). The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Hsieh, C.-Y., Zhang, J., Ma, Z., Kembhavi, A., and Krishna, R. (2023). Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 31096–31116. Curran Associates, Inc.

Hu, J. and Frank, M. (2024). Auxiliary task demands mask the capabilities of smaller language models. In *First Conference on Language Modeling*.

Hu, J. and Levy, R. (2023). Prompting is not a substitute for probability measurements in large language models. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Hubbard, T. L. (1996). Synesthesia-like mappings of lightness, pitch, and melodic interval. *The American Journal of Psychology*, 109(2):219–238.

Hurford, J. (2007). *The Origins of Meaning: Language in the Light of Evolution*. James R. Hurford. Oxford University Press.

Iida, H. and Funakura, H. (2024). Investigating iconicity in vision-and-language models: A case study of the bouba/kiki effect in japanese models. In *Proceedings of the 46th Annual Conference of the Cognitive Science Society*, volume 46.

Imai, M. and Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical transactions of the Royal Society B: Biological sciences*, 369(1651):20130298.

Imai, M., Kita, S., Nagumo, M., and Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition*, 109(1):54–65.

Jabri, A., Joulin, A., and van der Maaten, L. (2016). Revisiting visual question answering baselines. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 727–739, Cham. Springer International Publishing.

Jepma, M., Murphy, P. R., Nassar, M. R., Rangel-Gomez, M., Meeter, M., and Nieuwenhuis, S. (2016). Catecholaminergic regulation of learning rate in a dynamic environment. *PLOS Computational Biology*, 12(10):1–24.

Jones, C. R., Bergen, B., and Trott, S. (2024). Do multimodal large language models and humans ground language similarly? *Computational Linguistics*, 50(3):1415–1440.

Josserand, M., Pellegrino, F., Grosseck, O., Dediu, D., and Raviv, L. (2024). Adapting to individual differences: An experimental study of language evolution in heterogeneous populations. *Cognitive Science*, 48(11):e70011.

Juzek, T. S. and Ward, Z. B. (2025). Why does ChatGPT "delve" so much? exploring the sources of lexical overrepresentation in large language models. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S., editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6397–6411, Abu Dhabi, UAE. Association for Computational Linguistics.

Kallini, J., Papadimitriou, I., Futrell, R., Mahowald, K., and Potts, C. (2024). Mission: Impossible language models. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14691–14714, Bangkok, Thailand. Association for Computational Linguistics.

Kamath, A., Hessel, J., and Chang, K.-W. (2023). What's "up" with vision-language models? investigating their struggle with spatial reasoning. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175, Singapore. Association for Computational Linguistics.

Karamcheti, S., Nair, S., Balakrishna, A., Liang, P., Kollar, T., and Sadigh, D. (2024). Prismatic VLMs: Investigating the design space of visually-conditioned language models. In *Forty-first International Conference on Machine Learning*.

Kemp, C. and Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, 336(6084):1049–1054.

Kharitonov, E., Chaabouni, R., Bouchacourt, D., and Baroni, M. (2020). Entropy minimization in emergent languages. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5220–5230. PMLR.

Kharitonov, E., Dessì, R., Chaabouni, R., Bouchacourt, D., and Baroni, M. (2021). EGG: a toolkit for research on Emergence of lanGuage in Games. `https://github.com/facebookresearch/EGG`.

Khatun, A. and Brown, D. G. (2024). A study on large language models' limitations in multiple-choice question answering.

Kim, W., Son, B., and Kim, I. (2021). Vilt: Vision-and-language transformer without convolution or region supervision. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kirby, S. (2017). Culture and biology in the origins of linguistic structure. *Psychonomic bulletin & review*, 24(1):118–137.

Kirby, S. and Christiansen, M. H. (2003). From language learning to language evolution. In *Language Evolution*. Oxford University Press.

Kirby, S., Cornish, H., and 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.

Kirby, S., Griffiths, T., and Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, 28:108–114.

Kirby, S. and Tamariz, M. (2022). Cumulative cultural evolution, population structure and the origin of combinatoriality in human language. *Philosophical Transactions of the Royal Society B*, 377(1843):20200319.

Kirby, S., Tamariz, M., Cornish, H., and Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102.

Köhler, W. (1929). *Gestalt Psychology*. New York: Horace Liveright.

Köhler, W. (1947). *Gestalt Psychology*. (2nd ed.) New York: Horace Liveright.

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

Kottur, S., Moura, J., Lee, S., and Batra, D. (2017). Natural language does not emerge 'naturally' in multi-agent dialog. In Palmer, M., Hwa, R., and Riedel, S., editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2962–2967, Copenhagen, Denmark. Association for Computational Linguistics.

Kouwenhoven, T., Shahrasbi, K., and Verhoef, T. (2025). Cross-modal associations in vision and language models: Revisiting the bouba-kiki effect.

Kozachkov, L., Kastanenka, K. V., and Krotov, D. (2023). Building transformers from neurons and astrocytes. *Proceedings of the National Academy of Sciences*, 120(34):e2219150120.

Kriegeskorte, N., Mur, M., and Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4.

Lashley, K. S. et al. (1951). *The problem of serial order in behavior*, volume 21. Bobbs-Merrill Oxford.

Lazaridou, A. and Baroni, M. (2020). Emergent multi-agent communication in the deep learning era.

Lazaridou, A., Hermann, K. M., Tuyls, K., and Clark, S. (2018). Emergence of linguistic communication from referential games with symbolic and pixel input. In *International Conference on Learning Representations*.

Lazaridou, A., Peysakhovich, A., and Baroni, M. (2017). Multi-agent cooperation and the emergence of (natural) language. In *International Conference on Learning Representations*.

Lazaridou, A., Pham, N. T., and Baroni, M. (2016). Towards multi-agent communication-based language learning. *arXiv preprint arXiv:1605.07133*.

Lazaridou, A., Potapenko, A., and Tieleman, O. (2020). Multi-agent communication meets natural language: Synergies between functional and structural language learning. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7663–7674, Online. Association for Computational Linguistics.

Lee, H., Phatale, S., Mansoor, H., Mesnard, T., Ferret, J., Lu, K., Bishop, C., Hall, E., Carbune, V., Rastogi, A., and Prakash, S. (2024). Rlaif vs. rlhf: scaling reinforcement learning from human feedback with ai feedback. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Lewis, D. (1969). Convention: A philosophical study. *Cambridge, MA*.

Li, F. and Bowling, M. (2019). Ease-of-teaching and language structure from emergent communication. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Li, H., Nourkhiz Mahjoub, H., Chalaki, B., Tadiparthi, V., Lee, K., Moradi Pari, E., Lewis, C., and Sycara, K. (2024). Language grounded multi-agent reinforcement learning with human-interpretable communication. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C., editors, *Advances in Neural Information Processing Systems*, volume 37, pages 87908–87933. Curran Associates, Inc.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. (2016). A diversity-promoting objective function for neural conversation models. In Knight, K., Nenkova, A., and Rambow, O., editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Li, J., Li, D., Savarese, S., and Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Lian, Y., Bisazza, A., and Verhoef, T. (2021). The effect of efficient messaging and input variability on neural-agent iterated language learning. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10121–10129, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lian, Y., Bisazza, A., and Verhoef, T. (2023a). Communication Drives the Emergence of Language Universals in Neural Agents: Evidence from the Word-order/Case-marking Trade-off. *Transactions of the Association for Computational Linguistics*, 11:1033–1047.

Lian, Y., Bisazza, A., and Verhoef, T. (2023b). Communication drives the emergence of language universals in neural agents: Evidence from the word-order/case-marking trade-off. *Transactions of the Association for Computational Linguistics*, 11:1033–1047.

Lian, Y., Verhoef, T., and Bisazza, A. (2024). NeLLCom-X: A comprehensive neural-agent framework to simulate language learning and group communication. In Barak, L. and Alikhani, M., editors, *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 243–258, Miami, FL, USA. Association for Computational Linguistics.

Liljencrants, J. and Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, 48(4):839–862.

Lin, B. Y., Ravichander, A., Lu, X., Dziri, N., Sclar, M., Chandu, K., Bhagavatula, C., and Choi, Y. (2024). The unlocking spell on base LLMs: Rethinking alignment via in-context learning. In *The Twelfth International Conference on Learning Representations*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Lindblom, B. and Maddieson, I. (1988). Phonetic universals in consonant systems. In Li, C. and Hyman, L. M., editors, *Language, Speech and Mind*, pages 62–78. Routledge, London.

Little, H., Eryılmaz, K., and de Boer, B. (2017). Signal dimensionality and the emergence of combinatorial structure. *Cognition*, 168:1–15.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Llama Team, M. (2024). The llama 3 herd of models.

Locke, J. (1847). *An essay concerning human understanding*, volume 114. Kay & Troutman.

Lockwood, G. and Dingemanse, M. (2015). Iconicity in the lab: a review of behavioral, developmental, and neuroimaging research into sound-symbolism. *Frontiers in Psychology*, 6.

Löhn, L., Kiehne, N., Ljapunov, A., and Balke, W.-T. (2024). Is machine psychology here? on requirements for using human psychological tests on large language models. In Mahamood, S., Minh, N. L., and Ippolito, D., editors, *Proceedings of the 17th International Natural Language Generation Conference*, pages 230–242, Tokyo, Japan. Association for Computational Linguistics.

Lou, R., Zhang, K., and Yin, W. (2024). Large language model instruction following: A survey of progresses and challenges. *Computational Linguistics*, 50(3):1053–1095.

Lowe, R., Gupta, A., Foerster, J., Kiela, D., and Pineau, J. (2020). On the interaction between supervision and self-play in emergent communication. In *International Conference on Learning Representations*.

Lupyan, G. and Dale, R. (2016). Why are there different languages? the role of adaptation in linguistic diversity. *Trends in Cognitive Sciences*, 20(9):649–660.

Mahaut, M., Dessi, R., Franzon, F., and Baroni, M. (2025). Referential communication in heterogeneous communities of pre-trained visual deep networks. *Transactions on Machine Learning Research*.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., and Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2_Part_1):209–220.

Marks, L. E. (1974). On associations of light and sound: The mediation of brightness, pitch, and loudness. *The American Journal of Psychology*, 87(1/2):173–188.

Maurer, D., Pathman, T., and Mondloch, C. J. (2006). The shape of boubas: Sound–shape correspondences in toddlers and adults. *Developmental science*, 9(3):316–322.

McCrae, R. R., Costa, Jr, P. T., and Martin, T. A. (2005). The NEO–PI–3: A more readable revised NEO personality inventory. *Journal of Personality Assessment*, 84(3):261–270.

Meister, C., Pimentel, T., Wiher, G., and Cotterell, R. (2023). Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121.

Meister, C., Wiher, G., Pimentel, T., and Cotterell, R. (2022). On the probability–quality paradox in language generation. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 36–45, Dublin, Ireland. Association for Computational Linguistics.

Melara, R. D. (1989). Dimensional interaction between color and pitch. *Journal of Experimental Psychology: Human Perception and Performance*, 15(1):69.

Meteyard, L., Stoppard, E., Snudden, D., Cappa, S. F., and Vigliocco, G. (2015). When semantics aids phonology: A processing advantage for iconic word forms in aphasia. *Neuropsychologia*, 76:264–275. Special Issue: Semantic Cognition.

Michel, P., Rita, M., Mathewson, K. W., Tieleman, O., and Lazaridou, A. (2023). Revisiting populations in multi-agent communication. In *The Eleventh International Conference on Learning Representations*.

Mikolov, T., Joulin, A., and Baroni, M. (2018). A roadmap towards machine intelligence. In *Computational Linguistics and Intelligent Text Processing*, pages 29–61, Cham.

Millière, R. and Rathkopf, C. (2024). Anthropocentric bias and the possibility of artificial cognition. In *ICML 2024 Workshop on LLMs and Cognition*.

Millière, R. (2024). Language models as models of language.

Mina, M., Ruiz-Fernández, V., Falcão, J., Vasquez-Reina, L., and Gonzalez-Agirre, A. (2025). Cognitive biases, task complexity, and result intepretability in large language models. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S., editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1767–1784, Abu Dhabi, UAE. Association for Computational Linguistics.

Mishra, S., Khashabi, D., Baral, C., Choi, Y., and Hajishirzi, H. (2022). Reframing instructional prompts to GPTk's language. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 589–612, Dublin, Ireland. Association for Computational Linguistics.

Misra, K. and Mahowald, K. (2024). Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 913–929, Miami, Florida, USA. Association for Computational Linguistics.

Mitchell, M. and Krakauer, D. C. (2023). The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1928–1937, New York, New York, USA. PMLR.

Mollo, D. C. and Millière, R. (2023). The vector grounding problem.

Montero, M. L., Ludwig, C. J., Costa, R. P., Malhotra, G., and Bowers, J. (2021). The role of disentanglement in generalisation. In *International Conference on Learning Representations*.

Mordatch, I. and Abbeel, P. (2018). Emergence of grounded compositional language in multi-agent populations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Morey, R., Rouder, J., Jamil, T., Urbanek, S., Forner, K., and Ly, A. (2018). Bayesfactor: Computation of bayes factors for common designs (r package version 0.9. 12-4.2)[computer software]. *Retrieved form https://CRAN. R-project. org/package= BayesFactor*.

Mu, J. and Goodman, N. (2021). Emergent communication of generalizations. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 17994–18007. Curran Associates, Inc.

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Le Scao, T., Bari, M. S., Shen, S., Yong, Z. X., Schoelkopf, H., Tang, X., Radev, D., Aji, A. F., Almubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raff, E., and Raffel, C. (2023). Crosslingual generalization through multitask finetuning. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Muttenthaler, L., Dippel, J., Linhardt, L., Vandermeulen, R. A., and Kornblith, S. (2023). Human alignment of neural network representations. In *The Eleventh International Conference on Learning Representations*.

Nakagawa, S. and Schielzeth, H. (2013). A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142.

Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage*, 222:117254.

Neuberg, S. L. and Newsom, J. T. (1993). Personal need for structure: Individual differences in the desire for simpler structure. *Journal of Personality and Social Psychology*, 65(1):113.

Nielsen, A. and Rendall, D. (2011). The sound of round: evaluating the sound-symbolic role of consonants in the classic takete-maluma phenomenon. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 65(2):115.

Nielsen, A. and Rendall, D. (2012). The source and magnitude of sound-symbolic biases in processing artificial word material and their implications for language learning and transmission. *Language and Cognition*, 4(2):115–125.

Nielsen, A. and Rendall, D. (2013). Parsing the role of consonants versus vowels in the classic takete-maluma phenomenon. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 67(2):153.

Nikolaus, M. and Fourtassi, A. (2021). Evaluating the acquisition of semantic knowledge from cross-situational learning in artificial neural networks. In Chersoni, E., Hollenstein, N., Jacobs, C., Oseki, Y., Prévot, L., and Santus, E., editors, *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 200–210, Online. Association for Computational Linguistics.

Nussenbaum, K. and Hartley, C. A. (2019). Reinforcement learning across development: What insights can we draw from a decade of research? *Developmental Cognitive Neuroscience*, 40:100733.

Nölle, J., Staib, M., Fusaroli, R., and Tylén, K. (2018). The emergence of systematicity: How environmental and communicative factors shape a novel communication system. *Cognition*, 181:93–104.

Oliphant, M. (2002). *Learned systems of arbitrary reference: The foundation of human linguistic uniqueness*. Cambridge University Press.

Onnis, L., Lim, A., Cheung, S., and Huettig, F. (2022). Is the mind inherently predicting? exploring forward and backward looking in language processing. *Cognitive Science*, 46(10):e13201.

OpenAI (2024). Gpt-4 technical report.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. (2024). DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*.

Ostrand, R. and Berger, S. E. (2024). Humans linguistically align to their conversational partners, and language models should too. In *ICML 2024 Workshop on LLMs and Cognition*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Parise, C. V. and Spence, C. (2009). 'when birds of a feather flock together': Synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLOS ONE*, 4(5):1–7.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Peeperkorn, M., Kouwenhoven, T., Brown, D., and Jordanous, A. (2024). Is temperature the creativity parameter of large language models? In Grace, K., Llano, M. T., Martins, P., and Hedblom, M. M., editors, *Proceedings of the 15th International Conference on Computational Creativity*, pages 226–235. Association for Computational Creativity.

Peeperkorn, M., Kouwenhoven, T., Brown, D., and Jordanous, A. (2025). Mind the gap: Conformative decoding to improve output diversity of instruction-tuned large language models.

Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., and Strohmaier, M. (2024). Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5):808–826. PMID: 38165766.

Perez, J., Léger, C., Ovando-Tellez, M., Foulon, C., Dussauld, J., Oudeyer, P.-Y., and Moulin-Frier, C. (2024). Cultural evolution in populations of large language models.

Perlman, M., Dale, R., and Lupyan, G. (2015). Iconicity can ground the creation of vocal symbols. *Royal Society Open Science*, 2(8):150152.

Perniss, P., Thompson, R., and Vigliocco, G. (2010). Iconicity as a general property of language: Evidence from spoken and signed languages. *Frontiers in Psychology*, 1.

Perry, L. K., Perlman, M., and Lupyan, G. (2015). Iconicity in english and spanish and its relation to lexical category and age of acquisition. *PLOS ONE*, 10(9):1–17.

Piantadosi, S. T. (2024). Modern language models refute chomsky's approach to language. In *From fieldwork to linguistic theory: A tribute to Dan Everett*, pages 353–414. Language Science Press.

Piantadosi, S. T., Tily, H., and Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

Pimentel, T., McCarthy, A. D., Blasi, D., Roark, B., and Cotterell, R. (2019). Meaning to form: Measuring systematicity as information. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1751–1764, Florence, Italy. Association for Computational Linguistics.

Poletiek, F. H., Conway, C. M., Ellefson, M. R., Lai, J., Bocanegra, B. R., and Christiansen, M. H. (2018). Under what conditions can recursion be learned? effects of starting small in artificial grammar learning of center-embedded structure. *Cognitive Science*, 42(8):2855–2889.

Quinn, M. (2001). Evolving communication without dedicated communication channels. In Kelemen, J. and Sosík, P., editors, *Advances in Artificial Life*, pages 357–366, Berlin, Heidelberg. Springer Berlin Heidelberg.

Quinn, M., Smith, L., Mayley, G., and Husbands, P. (2003). Evolving controllers for a homogeneous system of physical robots: structured cooperation with minimal sensors. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 361(1811):2321–2343.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Ramachandran, V. S. and Hubbard, E. M. (2001). Synaesthesia–a window into perception, thought and language. *Journal of Consciousness Studies*, 8(12):3–34.

Raviv, L., de Heer Kloots, M., and Meyer, A. (2021). What makes a language easy to learn? a preregistered study on how systematic structure and community size affect language learnability. *Cognition*, 210:104620.

Raviv, L., Meyer, A., and Lev-Ari, S. (2019a). Compositional structure can emerge without generational transmission. *Cognition*, 182:151–164.

Raviv, L., Meyer, A., and Lev-Ari, S. (2019b). Larger communities create more systematic languages. *Proceedings of the Royal Society B: Biological Sciences*, 286(1907):20191262.

Ray, A., Radenovic, F., Dubey, A., Plummer, B., Krishna, R., and Saenko, K. (2023). Cola: A benchmark for compositional text-to-image retrieval. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 46433–46445. Curran Associates, Inc.

Ren, Y., Guo, S., Labeau, M., Cohen, S. B., and Kirby, S. (2020). Compositional languages emerge in a neural iterated learning model. In *International Conference on Learning Representations*.

Ren, Y., Guo, S., Qiu, L., Wang, B., and Sutherland, D. J. (2024). Bias amplification in language model evolution: An iterated learning perspective. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Ren, Y., Jin, R., Zhang, T., and Xiong, D. (2025). Do large language models mirror cognitive language processing? In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S., editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2988–3001, Abu Dhabi, UAE. Association for Computational Linguistics.

Rende, R., Gerace, F., Laio, A., and Goldt, S. (2024). A distributional simplicity bias in the learning dynamics of transformers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Rita, M., Chaabouni, R., and Dupoux, E. (2020). "LazImpa": Lazy and impatient neural agents learn to communicate efficiently. In Fernández, R. and Linzen, T., editors, *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 335–343, Online. Association for Computational Linguistics.

Rita, M., Michel, P., Chaabouni, R., Pietquin, O., Dupoux, E., and Strub, F. (2024). Language evolution with deep learning. *arXiv preprint arXiv:2403.11958*.

Rita, M., Strub, F., Grill, J.-B., Pietquin, O., and Dupoux, E. (2022a). On the role of population heterogeneity in emergent communication. In *International Conference on Learning Representations*.

Rita, M., Tallec, C., Michel, P., Grill, J.-B., Pietquin, O., Dupoux, E., and Strub, F. (2022b). Emergent communication: Generalization and overfitting in lewis games. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 1389–1404. Curran Associates, Inc.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Saito, K., Wachi, A., Wataoka, K., and Akimoto, Y. (2023). Verbosity bias in preference labeling by large language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Schaeffer, R., Miranda, B., and Koyejo, S. (2023). Are emergent abilities of large language models a mirage? In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 55565–55581. Curran Associates, Inc.

Schmidtke, D., Conrad, M., and Jacobs, A. M. (2014). Phonological iconicity. *Frontiers in Psychology*, 5.

Schulz, E. and Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology*, 55:7–14. Machine Learning, Big Data, and Neuroscience.

Schuster, M. and Paliwal, K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Schwartz, R. and Stanovsky, G. (2022). On the limitations of dataset balancing: The lost battle against spurious correlations. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2182–2194, Seattle, United States. Association for Computational Linguistics.

Scott-Phillips, T. C. and Kirby, S. (2010). Language evolution in the laboratory. *Trends in Cognitive Sciences*, 14(9):411–417.

Scott-Phillips, T. C., Kirby, S., and Ritchie, G. R. (2009). Signalling signalhood and the emergence of communication. *Cognition*, 113(2):226–233.

Selten, R. and Warglien, M. (2007). The emergence of simple languages in an experimental coordination game. *Proceedings of the National Academy of Sciences*, 104(18):7361–7366.

Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., Sap, M., and Shwartz, V. (2024). Clever hans or neural theory of mind? stress testing social reasoning in large language models. In Graham, Y. and Purver, M., editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273, St. Julian's, Malta. Association for Computational Linguistics.

Shiono, D., Brassard, A., Ishizuki, Y., and Suzuki, J. (2025). Evaluating model alignment with human perception: A study on shitsukan in LLMs and LVLMs. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S., editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11428–11444, Abu Dhabi, UAE. Association for Computational Linguistics.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., and Gal, Y. (2024). Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Simner, J., Cuskley, C., and Kirby, S. (2010). What sound does that taste? cross-modal mappings across gustation and audition. *Perception*, 39(4):553–569. PMID: 20515002.

Skean, O., Arefin, M. R., Zhao, D., Patel, N., Naghiyev, J., LeCun, Y., and Shwartz-Ziv, R. (2025). Layer by layer: Uncovering hidden representations in language models.

Smith, K. (2011). Learning Bias, Cultural Evolution of Language, and theBiological Evolution of the Language Faculty. *Human Biology*, 83(2):261 – 278.

Smith, K. (2022). How language learning and language use create linguistic structure. *Current Directions in Psychological Science*, 31(2):177–186.

Smith, K. and Culbertson, J. (2020). Communicative pressures shape language during communication (not learning): Evidence from casemarking in artificial languages.

Smith, K., Kirby, S., Guo, S., and Griffiths, T. L. (2024). Ai model collapse might be prevented by studying human language transmission. *Nature*, 633(8030):525.

Smith, K., Tamariz, M., and Kirby, S. (2013). Linguistic structure is an evolutionary trade-off between simplicity and expressivity. In *Proceedings of the annual meeting of the cognitive science society*, volume 35.

Steels, L. (1999). *The Talking Heads Experiment.* Laboratorium, Antwerpen, Antwerpen.

Steels, L. (2006). Experiments on the emergence of human communication. *Trends in Cognitive Sciences*, 10(8):347–349.

Steels, L. (2012a). *Grounding Language through Evolutionary Language Games*, pages 1–22. Springer US, Boston, MA.

Steels, L. (2012b). Self-organization and selection in cultural language evolution. In *Experiments in Cultural Language Evolution*, pages 1–37. John Benjamins, Amsterdam.

Steels, L. and Loetzsch, M. (2012). The grounded naming game. In *Experiments in Cultural Language Evolution*, volume 3, pages 41–59. John Benjamins.

Steinert-Threlkeld, S., Zhou, X., Liu, Z., and Downey, C. M. (2022). Emergent communication fine-tuning (EC-FT) for pretrained language models. In *Emergent Communication Workshop at ICLR 2022*.

Steinmetz, J.-P., Loare, E., and Houssemand, C. (2011). Rigidity of attitudes and behaviors: A study on the validity of the concept. *Individual Differences Research*, 9(2):84 – 106.

Sucholutsky, I. and Griffiths, T. (2023). Alignment with human representations supports robust few-shot learning. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 73464–73479. Curran Associates, Inc.

Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., Love, B. C., Grant, E., Groen, I., Achterberg, J., Tenenbaum, J. B., Collins, K. M., Hermann, K. L., Oktar, K., Greff, K., Hebart, M. N., Jacoby, N., Zhang, Q., Marjieh, R., Geirhos, R., Chen, S., Kornblith, S., Rane, S., Konkle, T., O'Connell, T. P., Unterthiner, T., Lampinen, A. K., Müller, K.-R., Toneva, M., and Griffiths, T. L. (2023). Getting aligned on representational alignment.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA.

Svecova, V. and Pavlovicova, G. (2016). Screening the personal need for the structure and solving word problems with fractions. *SpringerPlus*, 5(1):1–9.

Tamariz, M. and Kirby, S. (2015). Culture: Copying, compression, and conventionality. *Cognitive Science*, 39(1):171–183.

Tamariz, M., Roberts, S. G., Martínez, J. I., and Santiago, J. (2018). The interactive origin of iconicity. *Cognitive Science*, 42(1):334–349.

ter Hoeve, M., Kharitonov, E., Hupkes, D., and Dupoux, E. (2022). Towards interactive language modeling.

Theisen-White, C., Kirby, S., and Oberlander, J. (2011). Integrating the horizontal and vertical cultural transmission of novel communication systems. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, pages 956–961, Boston, MA. Cognitive Science Society.

Thompson, M. M., Naccarato, M. E., and Parker, K. E. (1989). Assessing cognitive need: The development of the personal need for structure and personal fear of invalidity scales. In *annual meeting of the Canadian Psychological Association, Halifax, Nova Scotia, Canada*.

Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., and Ross, C. (2022). Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5238–5248.

Tieleman, O., Lazaridou, A., Mourad, S., Blundell, C., and Precup, D. (2019). Shaping representations through communication: community size effect in artificial learning systems. *arXiv preprint arXiv:1912.06208*.

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Harvard University Press.

Tomasello, M. (2008). *Origins of human communication*. MIT Press, Cambridge, Mass.; London.

Turoman, N. and Styles, S. J. (2017). Glyph guessing for 'oo'and 'ee': Spatial frequency information in sound symbolic matching for ancient and unfamiliar scripts. *Royal Society open science*, 4(9):170882.

Tylén, K., Fusaroli, R., Bundgaard, P. F., and Østergaard, S. (2013). Making sense together: A dynamical account of linguistic meaning-making. *Semiotica*, 2013(194):39–62.

van Dijk, B., Kouwenhoven, T., Spruit, M., and van Duijn, M. J. (2023a). Large language models: The need for nuance in current debates and a pragmatic perspective on understanding. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12641–12654, Singapore. Association for Computational Linguistics.

van Dijk, B., van Duijn, M., Verberne, S., and Spruit, M. (2023b). ChiSCor: A corpus of freely-told fantasy stories by Dutch children for computational linguistics and cognitive science. In Jiang, J., Reitter, D., and Deng, S., editors, *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 352–363, Singapore. Association for Computational Linguistics.

van Dijk, B., van Duijn, M. J., Kloostra, L., Spruit, M., and Beekhuizen, B. (2024). Using a language model to unravel semantic development in children's use of a dutch perception verb. In Zock, M., Chersoni, E., Hsu, Y.-Y., and de Deyne, S., editors, *Proceedings of the Workshop on Cognitive Aspects of the Lexicon @ LREC-COLING 2024*, pages 98–106, Torino, Italia. ELRA and ICCL.

van Duijn, M., van Dijk, B., Kouwenhoven, T., de Valk, W., Spruit, M., and van der Putten, P. (2023). Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests. In Jiang, J., Reitter, D., and Deng, S., editors, *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 389–402, Singapore. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Verhoef, T. (2012). The origins of duality of patterning in artificial whistled languages. *Language and cognition*, 4(4):357–380.

Verhoef, T., de Boer, B., et al. (2011). Language acquisition age effects and their role in the preservation and change of communication systems. *Linguistics in Amsterdam*, 4(1).

Verhoef, T., Kirby, S., and De Boer, B. (2014). Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals. *Journal of Phonetics*, 43:57–68.

Verhoef, T., Kirby, S., and De Boer, B. (2016a). Iconicity and the emergence of combinatorial structure in language. *Cognitive science*, 40(8):1969–1994.

Verhoef, T., Roberts, S. G., and Dingemanse, M. (2015). Emergence of systematic iconicity: Transmission, interaction and analogy. In *37th Annual Meeting of the Cognitive Science Society (CogSci 2015)*, pages 2481–2486. Cognitive Science Society.

Verhoef, T., Walker, E., and Marghetis, T. (2016b). Cognitive biases and social coordination in the emergence of temporal language. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, pages 2615–2620, Austin, TX. Cognitive Science Society.

Verhoef, T., Walker, E., and Marghetis, T. (2022). Interaction dynamics affect the emergence of compositional structure in cultural transmission of space-time mappings. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44, pages 2133–2139.

Vital, F., Sardinha, A., and Melo, F. S. (2025). Implicit repair with reinforcement learning in emergent communication. In *24th International Conference on Autonomous Agents and Mutliagent Systems (AAMAS 2025)*.

Walsh, P., Soldaini, L., Groeneveld, D., Lo, K., Arora, S., Bhagia, A., Gu, Y., Huang, S., Jordan, M., Lambert, N., Schwenk, D., Tafjord, O., Anderson, T., Atkinson, D., Brahman, F., Clark, C., Dasigi, P., Dziri, N., Guerquin, M., Ivison, H., Koh, P. W., Liu, J., Malik, S., Merrill, W., Miranda, L. J. V., Morrison, J., Murray, T., Nam, C., Pyatkin, V., Rangapur, A., Schmitz, M., Skjonsberg, S., Wadden, D., Wilhelm, C., Wilson, M., Zettlemoyer, L., Farhadi, A., Smith, N. A., and Hajishirzi, H. (2025). 2 olmo 2 furious.

Wang, X., Ma, B., Hu, C., Weber-Genzel, L., Röttger, P., Kreuter, F., Hovy, D., and Plank, B. (2024). "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416, Bangkok, Thailand. Association for Computational Linguistics.

Ward, J., Huckstep, B., and Tsakanikos, E. (2006). Sound-colour synaesthesia: to what extent does it use cross-modal mechanisms common to us all? *Cortex*, 42(2):264–280.

Warstadt, A. (2022). *Artificial Neural Networks as Models of Human Language Acquisition*. New York University.

Warstadt, A. and Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. *Algebraic Structures in Natural Language*, pages 17–60.

Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., and Cotterell, R. (2023). Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., and Cotterell, R., editors, *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Washburn, M. F. (1916). *Movement and mental imagery: Outlines of a motor theory of the complexer mental processes*. Houghton Mifflin.

Weber, L., Bruni, E., and Hupkes, D. (2023). Mind the instructions: a holistic evaluation of consistency and interactions in prompt-based learning. In Jiang, J., Reitter, D., and Deng, S., editors, *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 294–313, Singapore. Association for Computational Linguistics.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Westbury, C. (2005). Implicit sound symbolism in lexical access: Evidence from an interference task. *Brain and language*, 93(1):10–19.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wilcox, E. G., Futrell, R., and Levy, R. (2023). Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, pages 1–44.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256.

Winter, B., Perlman, M., Perry, L. K., and Lupyan, G. (2017). Which words are most iconic?: Iconicity in english sensory words. *Interaction Studies*, 18:443–464.

Wittgenstein, L. (1953). *Philosophical Investigations*. Basil Blackwell, Oxford.

Xu, T., Kuribayashi, T., Oseki, Y., Cotterell, R., and Warstadt, A. (2025). Can language models learn typologically implausible languages?

Xu, Z., Niethammer, M., and Raffel, C. A. (2022). Compositional generalization in unsupervised compositional representation learning: A study on disentanglement and emergent language. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 25074–25087. Curran Associates, Inc.

Yiu, E., Kosoy, E., and Gopnik, A. (2024). Transmission versus truth, imitation versus innovation: What children can do that large language and language-and-vision models cannot (yet). *Perspectives on Psychological Science*, 19(5):874–883. PMID: 37883796.

Zhang, J., Huang, J., Jin, S., and Lu, S. (2024a). Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20.

Zhang, Y., Verhoef, T., van Noord, G., and Bisazza, A. (2024b). Endowing neural language learners with human-like biases: A case study on dependency length minimization. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5819–5832, Torino, Italia. ELRA and ICCL.

Zhang, Y., Zhang, C., Tang, Y., and He, Z. (2024c). Cross-modal concept learning and inference for vision-language models. *Neurocomputing*, 583:127530.

Zheng, C., Zhang, J., Kembhavi, A., and Krishna, R. (2024). Iterated learning improves compositionality in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13785–13795.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., Zhang, S., Ghosh, G., Lewis, M., Zettlemoyer, L., and Levy, O. (2023). Lima: less is more for alignment. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Summary

Language is a uniquely human trait that enables us to collaborate and exchange ideas on a daily basis. Although it is now taken for granted that we understand each other's language, the way this shared understanding came to be remains a mystery. A prominent theory in language evolution proposes that repeated interactions provide anchor points where signals and meanings become linked. Cultural processes can then spread these meaningful signals across populations. Together, these processes help shape a language such that it aligns with users' cognitive preferences, such as a preference for simple and structured signals. Sustained pressure from such preferences results in a shared language that is structured, expressive, and learnable.

A new type of language user is playing an increasingly prominent role in our society. Artificially intelligent systems, such as Large Language Models, can now be considered mature language users. However, the way they make decisions fundamentally differs from how humans do. Therefore, it is essential to investigate these largely unknown forms of cognition. This dissertation does so from the perspective of language evolution. It uses methods from that field, which are not only applicable to humans but are also suitable for unravelling cognitive patterns and preferences in non-human systems. The resulting findings contribute to our understanding of language evolution and inform us about how non-human cognition processes linguistic and visual information. In doing so, this dissertation addresses the question of how human and non-human cognition can complement each other in research on the evolution of language.

First, it examines how rudimentary signals emerge and the role that neural networks can play in this process. Experiments with humans confirm that shared rudimentary signals arise from repeated interactions and that it can help when conversation partners differ in their need for structure. This latter point nuances existing theory, which suggests that shared preferences are advantageous. By simulating human behaviour in this task using computational models, we see that bidirectional mechanisms best explain this behaviour–meaning that focusing on the surrounding context (both before and after) of a communicative signal is essential, just as it is in language processing.

Next, the dissertation investigates whether multimodal models display human-like cross-modal associations and studies, using self-learning models, the evolution of structured language

built from rudimentary elements. Where humans consistently name a rounded object "bouba" instead of "kiki," four multimodal models barely exhibit this pattern. This confirms that modern multimodal models struggle to link visual and textual information. Even when these models play a communication game where they develop their own language, they use different visual features than humans do. As a result, they are unable to distinguish between two images containing the same types of objects arranged differently–even when explicitly trained to learn the correct features.

Finally, the dissertation explores whether modern language models can serve as participants in language evolution experiments and whether humans and machines can collaboratively learn a language. When language models repeatedly interact in a communication game, a language emerges with compositional properties in which syllables are recombined to describe different objects. This bears striking similarities to human language evolution, even though the underlying mechanisms differ fundamentally. Simulations involving cultural transmission–where successive generations learn and use the language–also show that the language adapts to the user preferences of the language models in a way similar to how it does with humans. Still, there are differences: language models more frequently produce a single signal with multiple meanings (homonyms), and longer signals, whereas humans tend to prefer short, expressive utterances. The final contribution of this dissertation is an experiment in which participants collaborate with a Llama3-70B language model to develop a language. Despite the fundamentally different mechanisms for language acquisition and use, expressive and meaningful languages emerge. These languages contain compositional properties and show more human-like traits than languages developed without human involvement. These results support the idea that language adapts to its users and suggest that interactions, where communicative success matters, can play a role in natural language acquisition by machines.

Through this interdisciplinary approach to language evolution, this dissertation contributes to a line of research that studies both human and non-human cognition from a human-centred perspective. It shows that both forms of cognition can complement each other precisely because they differ. This finding offers promising possibilities for the development of communication systems adapted to the cognitive capacities of both humans and machines, potentially enabling more natural interactions between them.

Samenvatting

Taal is een unieke menselijke eigenschap waarmee we dagelijks samenwerken en ideeën uitwisselen. Hoewel het nu een gegeven is dat we elkaars taal begrijpen, is de manier waarop dit gedeeld begrip is ontstaan tot op heden nog een raadsel. Een prominent gedachtegoed in taalevolutie stelt dat herhaaldelijke interacties een ijkmoment bieden waarin signalen en betekenissen aan elkaar worden gelinkt. Culturele processen kunnen deze betekenisvolle signalen vervolgens verspreiden op populatieniveau. Samen zorgen ze er voor dat een taal zich aanpast aan de cognitieve voorkeuren van gebruikers, zoals een voorkeur voor simpele en gestructureerde signalen. Langdurige druk van zulke voorkeuren resulteert in een gedeelde taal die gestructureerd, expressief en te leren is.

Een nieuw soort taalgebruiker krijgt een steeds prominentere rol in onze samenleving. Kunstmatig intelligente systemen, zoals Large Language Models, kunnen inmiddels worden beschouwd als volwaardige taalgebruikers. Echter verschilt de manier waarop ze beslissingen nemen fundamenteel van hoe mensen dat doen. Daarom is het essentieel om deze grotendeels onbekende vormen van cognitie te onderzoeken. Dit proefschrift doet dat vanuit het perspectief van taalevolutie. Het gebruikt methoden uit dat veld, die niet alleen toepasbaar zijn op mensen, maar ook geschikt zijn om cognitieve patronen en voorkeuren bij niet-menselijke systemen te ontrafelen. De bevindingen die hieruit voorvloeien dragen bij aan het begrip van taalevolutie én informeren hoe niet-menselijke cognitie talige en visuele informatie verwerkt. Hiermee beantwoordt dit proefschrift de vraag hoe menselijke en niet-menselijke cognitie elkaar kunnen complementeren in onderzoek naar de evolutie van taal.

Allereerst wordt onderzocht hoe rudimentaire signalen onstaan, en welke rol neurale netwerken daarbij kunnen spelen. Experimenten met mensen bevestigen dat gedeelde rudimentaire signalen ontstaan uit herhaaldelijke interacties en dat het kan helpen als gesprekspartners verschillen in hun behoefte aan structuur. Dit laatste nuanceert bestaand gedachtegoed dat gedeelde voorkeuren voordelig zijn. Door het gedrag van mensen in deze taak na te bootsen met computermodellen zien we dat bidirectionele mechanismen dit gedrag het beste kunnen verklaren, wat wil zeggen dat een focus op de omliggende context (zowel voor als na) van een communicatief signaal essentieel is, net als bij het verwerken van taal.

Vervolgens wordt onderzocht of multimodale modellen menselijke cross-modale associaties

vertonen en bestudeert dit proefschrift, met behulp van zelflerende modellen, de evolutie van gestrucureerde taal waarin signalen opgebouwd zijn uit rudimentaire elementen. Waar mensen consistent een rondvormig object 'bouba' zouden noemen in plaats van 'kiki', vertonen vier multimodale modellen dit patroon nauwelijks. Dit bevestigt dat het voor moderne multimodale modellen lastig is om visuele en tekstuele informatie aan elkaar te linken. Zelfs wanneer dit soort multimodale modellen een communicatiespel spelen waarbij een eigen taal ontwikkelt wordt blijkt dat er andere visuele kenmerken worden gebruikt dan de kenmerken die mensen gebruiken. Hierdoor zijn ze niet in staat om met de geleerde taal twee plaatjes met dezelfde soort objecten in een andere opstelling van elkaar te onderscheiden—zelfs niet wanneer de modellen expliciet gedwongen worden om de juiste kenmerken te leren.

Afsluitend onderzoekt dit proefschrift of moderne taalmodellen kunnen fungeren als proefpersonen in taalevolutie-experimenten en bevraagt het in een collaboratief experiment of mens en machine samen een taal kunnen leren. Wanneer taalmodellen herhaaldelijk met elkaar interacteren in een communicatiespel onstaat er een taal met compositionele eigenschappen waarin lettergrepen in nieuwe combinaties worden gebruikt om verschillende objecten te beschrijven. Dit heeft opvallende overeenkomsten met menselijke taalevolutie, ondanks dat de onderliggende mechanismen fundamenteel van elkaar verschillen. Ook uit simulaties met culturele overdracht—waarin opeenvolgende generaties taal leren en gebruiken—blijkt dat de taal zich aanpast naar gebruikersvoorkeuren van de taalmodellen op een manier die vergelijkbaar is met mensen. Toch zijn er ook verschillen: taalmodellen produceren vaker één signaal met meerdere betekenissen (homoniemen), en lange signalen, terwijl mensen de voorkeur geven aan korte, expressieve uitingen. De laatste bijdrage van dit proefschrift betreft een experiment waarin partipanten samenwerken met een Llama3-70B taalmodel en een taal ontwikkelen. Ondanks de fundamenteel verschillende mechanismen voor taalverwerving en gebruik, ontstaan er expressieve en betekenisvolle talen. Deze talen bevatten compositionele eigenschappen en vertonen meer menselijke eigenschappen dan talen ontwikkeld zonder menselijke inbreng. Hiermee ondersteunen de resultaten het idee dat taal zich aanpast aan haar gebruikers en suggereren ze dat interacties, waarin communicatief succes belangrijk is, een rol kunnen spelen in natuurlijke taalverwerving door machines.

Met deze interdisciplinaire benadering van taalevolutie draagt dit proefschrift bij aan een onderzoekslijn die vanuit een menselijk oogpunt zowel menselijke als niet-menselijke cognitie onderzoekt. Het laat zien dat beide vormen van cognitie elkaar kunnen complementeren omdat ze van elkaar verschillen. Deze bevinding biedt veelbelovende mogelijkheden voor de ontwikkeling van talen die zijn aangepast aan de cognitieve capaciteiten van zowel mensen als machines waardoor we mogelijk op een natuurlijkere manier kunnen interacteren.

Acknowledgements

Life as a PhD candidate is somewhat strange. Many bright minds surround you, but your quest is very personal: only you can do it. Interacting with people is thus paramount. Without them, I do not think there is a thing called a 'successful' dissertation. I want to thank some people with whom I frequently interacted and who helped making my life as a PhD candidate a success.

I was very fortunate to have Tessa Verhoef as my co-promoter. We shared many unforgettable moments: from Sushi, Macaques, and Ramen bowls in Japan, to nailing statistical analyses above the Atlantic Ocean, to hiking in Wisconsin, Bangkok, and Singapore. Tessa, thank you for your support and dedication, which allowed me to research what I am passionate about. I hope to share many more moments in the future.

Bram van Dijk and Max Peeperkorn, I have long thought about my acknowledgements to you, and to this moment, I still do not know how to do you justice. You have been my sparring partners, inspiration sources, and complaining buddies, I could go on for hours. But most importantly, you are *very* good friends. Bram, what started with coffee and a stroll around Leiden became true companionship. Besides your ever-listening ear and critical view, I dearly value our mutually grown hobby of introducing new things to each other, like Hans van Werven cigars, eating ramen with chopsticks, and slurping oysters. Max, our story also starts with a multitude of strolls, this time through Linz. Our weekly meetings, in which we discussed programming issues, paper writing, and simple chitchat, were something I always looked forward to. Beyond the Fridays, you were always available–no matter what time of day. Even to the point that Anniko once asked me whether I also helped you! I hope I did. Besides work, you are a true friend with whom I hope to spend more time in the future!

Initially, it seemed challenging to position my work between linguistics and computer science, but I believe that this has become one of its strengths. Thanks to my promotor Stephan Raaijmakers, I had the opportunity and freedom to settle between these fields. Roy de Kleijn, as my co-promotor, you helped set up experiments and were always there to listen, thank you. I know where to find you on October 2nd.

Max van Duijn, even though you had no formal role during my PhD, you were always interested, giving me the feeling that my work mattered. Besides, you taught me a great deal about the world of roasted beans. Thank you, Max. To learning from each other when we

collaborate in the future!

I am grateful to Kiana Shahrasbi, Koen Poortvliet, and Neval Kara for their help in conducting experiments during my PhD. Of course, I also want to thank all the participants; without you, there would be nothing to analyse. I hope you enjoyed taking part.

Thanks to all the people at the office, including (but not limited to) Bernard Hilpert, Dan Xu, Danica Mast, Evani Lachmansing, Koen Ponse, Marcello Gómez-Maureira, Matthias Müller-Brockhausen, Marianne Bossema, Peter van der Putten, Ramira van der Meulen, Rob Saunders, Ross Towns, and Sabijn Perdijk. You, in one way or another, contributed to a great social environment that cannot be underestimated when the rest of your work deals with anti-social computers that never laugh at your jokes. I want to thank some people in particular. Max van Haastrecht, thank you for co-organising the PhD seminar and for all the things you did for PhD candidates in general. Zane Kripe, we simply vibe. Thank you for your listening ears, your everlasting smile and our conversations. Michiel van der Meer, times are long and sweet. Singapore made me hungry for more, and I hope to travel to many more conferences together. Giulio Barbero, our squabble game keeps us, and perhaps everyone else, entertained, thanks! Lennard ('The Northerner') Froma, not a single day goes by without laughing to tears. To lots of jokes in the future. Felix Kleuker, I enjoyed our strolls, but I will never forgive you for throwing me under the bus twice during Werewolf. Lennard and Felix, if, after all, the life of an academic does not suit you, know that you can always pursue a career in tree chopping. Alice Mulder, it turns out that you are not only an incredible illustrator, but collaborative human-AI cover-making is also a piece of cake for you, thank you.

Beyond academia, I want to thank the friends who kept me from losing myself in the world of academia. Thank you Annelotte Mulder, Bjorn de Jong, Bob van Brussel, Boris van Hattum, De Collectie, Freek Boelders, Fábio Matos da Costa, Janneke van Oirschot, Jurre van Rijswijck, Lara Kluts, Luuk Hopman, Mark Snoek, Nicole de Groot, Paulien Elfring, Roomburg Heren 1, Ties Eigenhuis, Tania Sluckin, and Willeke Verduijn.

Thanks to Anita Israel, Kristian Vinke, Noor Janssen, and Xander Janssen for always keeping the door open in Zeeland. Perhaps we will experience some self-driving in seafaring; only time will tell.

Pursuing a research career was only possible because my lovely family always supported me. Seeing you feels like a small getaway where all that matters is simply being together. Thank you dearly Susan Kouwenhoven, Mylène Rutten, Huub Kouwenhoven, Caroline Schippers, and Grégory Dianzenza. I care a great deal about you.

Finally, I want to express my deepest gratitude to Anniko Janssen. You bring me joy, laughter, craziness, love, and many more things that are much more valuable and go far beyond the positives science can offer. No matter the activity—be it cooking, travelling, or dancing—being together always makes it fun. Life with you is like a gift that keeps on giving.

List of publications

Asterisks denote equal contributions.

1 **Kouwenhoven, T.**, Towns, R., Verhoef, T. (2022). Modelling the emergence of vocal grooming. In Ravignani, A., Asano, R., Valente, D., Ferretti, F., Hartmann, S., Hayashi, M., Jadoul, Y., Martins, M., Oseki, Y., Rodrigues, E. D., Vasileva, O., & Wacewicz, S., editors, *Proceedings of the Joint Conference on Language Evolution (JCoLE)* pages 434-436. Max Planck Institute for Psycholinguistics.

2 **Kouwenhoven, T.**, Verhoef, T., De Kleijn, R., and Raaijmakers, S. (2022). Emerging grounded shared vocabularies between human and machine, inspired by human language evolution. *Frontiers in Artificial Intelligence*, 5:886349

3 **Kouwenhoven, T.**, de Kleijn, R.E., Raaijmakers, S.A., Verhoef, T.(2023). Need for Structure and the Emergence of Communication. In J. Culbertson, A. Perfors., H. Rabagliati. & V. Ramenzoni., editors, *Proceedings of the Annual Meeting of the Cognitive Science Society*, Volume 44, pages 549-555. Cognitive Science Society.

4 **Kouwenhoven, T.**, Verhoef, T., Raaijmakers, S.A., de Kleijn, R.E. (2023). Modelling Human Sequential Behaviour with Deep Learning Neural Networks in Emergent Communication. In M. Goldwater., F. K. Anggoro., B. K. Hayes., & D. C. Ong., editors, *Proceedings of the Annual Meeting of the Cognitive Science Society*, Volume 44, pages 549-555. Cognitive Science Society.

5 Van Duijn, M.J.*, Van Dijk, B.M.A.*, **Kouwenhoven, T.***, De Valk, W.M., Spruit, M.R., and Van Der Putten, P.W.H. (2023). Theory of Mind in Large Language Models: Examining Performance of 11 State-of-the-Art Models vs. Children Aged 7-10 on Advanced Tests. In Jiang, J., Reitter, D., and Deng, S., editors, *Proceedings of the 27th Conference on Computational Natural Language Learning*, pages 389-402, Singapore. Association for Computational Linguistics.

6 van Dijk, B., **Kouwenhoven, T.**, Spruit, M., and van Duijn, M. J. (2023a). Large language models: The need for nuance in current debates and a pragmatic perspective on understanding. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Con-*

ference on Empirical Methods in Natural Language Processing, pages 12641–12654, Singapore. Association for Computational Linguistics

7 Peeperkorn, M., **Kouwenhoven, T.**, Brown, D., and Jordanous, A. (2024). Is temperature the creativity parameter of large language models? In Grace, K., Llano, M. T., Martins, P., and Hedblom, M. M., editors, *Proceedings of the 15th International Conference on Computational Creativity*, pages 226–235. Association for Computational Creativity

8 **Kouwenhoven, T.**, Peeperkorn, M., Van Dijk, B., and Verhoef, T. (2024). The curious case of representational alignment: Unravelling visio-linguistic tasks in emergent communication. In Kuribayashi, T., Rambelli, G., Takmaz, E., Wicke, P., and Oseki, Y., editors, *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 57–71, Bangkok, Thailand. Association for Computational Linguistics

9 Verhoef, T.*, Shahrasbi, K., and **Kouwenhoven, T***. (2024). What does kiki look like? cross-modal associations between speech sounds and visual shapes in vision-and-language models. In Kuribayashi, T., Rambelli, G., Takmaz, E., Wicke, P., and Oseki, Y., editors, *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 199–213, Bangkok, Thailand. Association for Computational Linguistics

10 **Kouwenhoven, T.**, Peeperkorn, M., and Verhoef, T. (2025b). Searching for structure: Investigating emergent communication with large language models. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S., editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9977–9991, Abu Dhabi, UAE. Association for Computational Linguistics

11 **Kouwenhoven, T.**, Peeperkorn, M., de Kleijn, R., and Verhoef, T. (2025a). Shaping shared languages: Human and large language models' inductive biases in emergent communication. In Kwok, J., editor, *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI-25, International Joint Conferences on Artificial Intelligence Organization. Human-Centred AI

12 Peeperkorn, M., **Kouwenhoven, T.**, Brown, D., and Jordanous, A. (2025). Mind the gap: Conformative decoding to improve output diversity of instruction-tuned large language models. *Preprint*

12 Pitta, E., **Kouwenhoven, T.**, Verhoef, T. (2025). Probing Vision-Language Understanding through the Visual Entailment Task: promises and pitfalls. In Henrique, L.C., Rui, S., Koponen, M., Pareja-Lora, A., editors, *Proceedings of the Second LUHME Workshop*, Bologna, Spain.

13 **Kouwenhoven, T***., Shahrasbi, K., Verhoef, T.* (2025). Cross-modal Associations in Vision and Language Models: Revisiting the bouba-kiki effect. *Preprint*

included in this dissertation.

Curriculum Vitae

Tom Kouwenhoven was born on January 26, 1995, in Nijmegen, the Netherlands. He graduated in 2013 with a VWO-plus diploma from the Kandinsky College in Nijmegen, Gelderland. From 2014 to 2017, Tom pursued a bachelor's degree in Lifestyle Informatics (now known as Artificial Intelligence) at the Vrije Universiteit in Amsterdam. During this time, he completed a minor in Computation Arts at Concordia University in Montréal, Canada. He obtained his master's degree in 2020 from the University of Leiden, where he completed the Media Technology program.

In 2021, Tom started as a PhD candidate at the Leiden Institute for Advanced Computer Science, which he successfully completed in 2025. During his doctoral studies, he collaborated with researchers from various (inter)national institutions – most notably with Max Peeperkorn. One of these collaborations, exploring the effect of large language models' temperature parameter on their creative writing abilities, received the Best Student Paper Award at the 2024 International Conference on Computational Creativity in Jönköping, Sweden. Alongside his research, Tom completed training in transferable skills, including communication for science, presenting skills for PhDs, and scientific conduct. Since May 2025, Tom has been working as a postdoctoral researcher at Leiden University, affiliated with the Hybrid Intelligence consortium.

SIKS dissertation series

2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
04 Laurens Rietveld (VUA), Publishing and Consuming Linked Data
05 Evgeny Sherkhonov (UvA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
07 Jeroen de Man (VUA), Measuring and modeling negative emotions for virtual training
08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
09 Archana Nottamkandath (VUA), Trusting Crowdsourced Information on Cultural Artefacts
10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
11 Anne Schuth (UvA), Search Engines that Learn from Their Users
12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
13 Nana Baah Gyan (VUA), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
16 Guangliang Li (UvA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17 Berend Weel (VUA), Towards Embodied Evolution of Robot Organisms
18 Albert Meroño Peñuela (VUA), Refining Statistical Data on the Web
19 Julia Efremova (TU/e), Mining Social Structures from Genealogical Data
20 Daan Odijk (UvA), Context & Semantics in News & Web Search
21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
22 Grace Lewis (VUA), Software Architecture Strategies for Cyber-Foraging Systems
23 Fei Cai (UvA), Query Auto Completion in Information Retrieval
24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26 Dilhan Thilakarathne (VUA), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30 Ruud Mattheij (TiU), The Eyes Have It
31 Mohammad Khelghati (UT), Deep web content monitoring
32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33 Peter Bloem (UvA), Single Sample Statistics, exercises in learning from just one example
34 Dennis Schunselaar (TU/e), Configurable Process Trees: Elicitation, Analysis, and Enactment
35 Zhaochun Ren (UvA), Monitoring Social Media: Summarization, Classification and Recommendation
36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40 Christian Detweiler (TUD), Accounting for Values in Design
41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
42 Spyros Martzoukos (UvA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44 Thibault Sellam (UvA), Automatic Assistants for Database Exploration
45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46 Jorge Gallego Perez (UT), Robots to Make you Happy
47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48 Tanja Buttler (TUD), Collecting Lessons Learned
49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis

50 Yan Wang (TiU), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime
02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
05 Mahdieh Shadi (UvA), Collaboration Behavior
06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
08 Rob Konijn (VUA), Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
10 Robby van Delden (UT), (Steering) Interactive Play Behavior
11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
12 Sander Leemans (TU/e), Robust Process Mining with Guarantees
13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
14 Shoshannah Tekofsky (TiU), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
15 Peter Berck (RUN), Memory-Based Text Correction
16 Aleksandr Chuklin (UvA), Understanding and Modeling Users of Modern Search Engines
17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
18 Ridho Reinanda (UvA), Entity Associations for Search
19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
22 Sara Magliacane (VUA), Logics for causal inference under uncertainty
23 David Graus (UvA), Entities of Interest — Discovery in Digital Traces
24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
25 Veruska Zamborlini (VUA), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
28 John Klein (VUA), Architecture Practices for Complex Contexts
29 Adel Alhuraibi (TiU), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT
30 Wilma Latuny (TiU), The Power of Facial Expressions
31 Ben Ruijl (UL), Advances in computational methods for QFT calculations
32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
35 Martine de Vos (VUA), Interpreting natural science spreadsheets
36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
37 Alejandro Montes Garcia (TU/e), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
38 Alex Kayal (TUD), Normative Social Applications
39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval
44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement
46 Jan Schneider (OU), Sensor-based Learning Support
47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration
48 Angel Suarez (OU), Collaborative inquiry-based learning

2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations
02 Felix Mannhardt (TU/e), Multi-perspective Process Mining
03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
05 Hugo Huurdeman (UvA), Supporting the Complex Dynamics of the Information Seeking Process
06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems
08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems
09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology
11 Mahdi Sargolzaei (UvA), Enabling Framework for Service-oriented Collaborative Networks
12 Xixi Lu (TU/e), Using behavioral context in process mining
13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
14 Bart Joosten (TiU), Detecting Social Signals with Spatiotemporal Gabor Filters
15 Naser Davarzani (UM), Biomarker discovery in heart failure
16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
17 Jianpeng Zhang (TU/e), On Graph Sample Clustering
18 Henriette Nakad (UL), De Notaris en Private Rechtspraak
19 Minh Duc Pham (VUA), Emergent relational schemas for RDF
20 Manxia Liu (RUN), Time and Bayesian Networks

	21	Aad Slootmaker (OU), EMERGO: a generic platform for authoring and playing scenario-based serious games
	22	Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
	23	Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
	24	Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
	25	Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
	26	Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
	27	Maikel Leemans (TU/e), Hierarchical Process Mining for Scalable Software Analysis
	28	Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
	29	Yu Gu (TiU), Emotion Recognition from Mandarin Speech
	30	Wouter Beek (VUA), The "K" in "semantic web" stands for "knowledge": scaling semantics to the web
2019	01	Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification
	02	Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
	03	Eduardo Gonzalez Lopez de Murillas (TU/e), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
	04	Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
	05	Sebastiaan van Zelst (TU/e), Process Mining with Streaming Data
	06	Chris Dijkshoorn (VUA), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
	07	Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
	08	Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes
	09	Fahimeh Alizadeh Moghaddam (UvA), Self-adaptation for energy efficiency in software systems
	10	Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
	11	Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
	12	Jacqueline Heinerman (VUA), Better Together
	13	Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
	14	Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
	15	Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
	16	Guangming Li (TU/e), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
	17	Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts
	18	Gerard Wagenaar (UU), Artefacts in Agile Team Communication
	19	Vincent Koeman (TUD), Tools for Developing Cognitive Agents
	20	Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
	21	Cong Liu (TU/e), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
	22	Martin van den Berg (VUA),Improving IT Decisions with Enterprise Architecture
	23	Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
	24	Anca Dumitrache (VUA), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
	25	Emiel van Miltenburg (VUA), Pragmatic factors in (automatic) image description
	26	Prince Singh (UT), An Integration Platform for Synchromodal Transport
	27	Alessandra Antonaci (OU), The Gamification Design Process applied to (Massive) Open Online Courses
	28	Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
	29	Daniel Formolo (VUA), Using virtual agents for simulation and training of social skills in safety-critical circumstances
	30	Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
	31	Milan Jelisavcic (VUA), Alive and Kicking: Baby Steps in Robotics
	32	Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
	33	Anil Yaman (TU/e), Evolution of Biologically Inspired Learning in Artificial Neural Networks
	34	Negar Ahmadi (TU/e), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES
	35	Lisa Facey-Shaw (OU), Gamification with digital badges in learning programming
	36	Kevin Ackermans (OU), Designing Video-Enhanced Rubrics to Master Complex Skills
	37	Jian Fang (TUD), Database Acceleration on FPGAs
	38	Akos Kadar (OU), Learning visually grounded and multilingual representations
2020	01	Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
	02	Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
	03	Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
	04	Maarten van Gompel (RUN), Context as Linguistic Bridges
	05	Yulong Pei (TU/e), On local and global structure mining
	06	Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
	07	Wim van der Vegt (OU), Towards a software architecture for reusable game components
	08	Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search
	09	Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
	10	Alifah Syamsiyah (TU/e), In-database Preprocessing for Process Mining
	11	Sepideh Mesbah (TUD), Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models
	12	Ward van Breda (VUA), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
	13	Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
	14	Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
	15	Konstantinos Georgiadis (OU), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
	16	Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
	17	Daniele Di Mitri (OU), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
	18	Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
	19	Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
	20	Albert Hankel (VUA), Embedding Green ICT Maturity in Organisations
	21	Karine da Silva Miras de Araujo (VUA), Where is the robot?: Life as it could be
	22	Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar

23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
24 Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
25 Xin Du (TU/e), The Uncertainty in Exceptional Model Mining
26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization
27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context
28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
30 Bob Zadok Blok (UL), Creatief, Creatiever, Creatiefst
31 Gongjin Lan (VUA), Learning better – From Baby to Better
32 Jason Rhuggenaath (TU/e), Revenue management in online markets: pricing and online advertising
33 Rick Gilsing (TU/e), Supporting service-dominant business model evaluation in the context of business model innovation
34 Anna Bon (UM), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development
35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production

2021 01 Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space
02 Rijk Mercuur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models
03 Seyyed Hadi Hashemi (UvA), Modeling Users Interacting with Smart Devices
04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
05 Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems
06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
07 Armel Lefebvre (UU), Research data management for open science
08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play
10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning
14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support
15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm
16 Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues
17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks
18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks
19 Roberto Verdecchia (VUA), Architectural Technical Debt: Identification and Management
20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems
21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes
22 Sihang Qiu (TUD), Conversational Crowdsourcing
23 Hugo Manuel Proença (UL), Robust rules for prediction and description
24 Kaijie Zhu (TU/e), On Efficient Temporal Subgraph Query Processing
25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications
26 Benno Kruit (CWI/VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables
27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You
28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs

2022 01 Judith van Stegeren (UT), Flavor text generation for role-playing video games
02 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey
03 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare
04 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework
05 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization
06 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding
07 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics
08 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design
09 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach
10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines
11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring
12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases
13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge
14 Michiel Overeem (UU), Evolution of Low-Code Platforms
15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining
16 Pieter Gijsbers (TU/e), Systems for AutoML Research
17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification
18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation
19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation
20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media
21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments
22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations
23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents
24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values
25 Anna L.D. Latour (UL), Optimal decision-making under constraints and uncertainty

26 Anne Dirkson (UL), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences
27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences
28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems
29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality
30 Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays
31 Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management
32 Cezara Pastrav (UU), Social simulation for socio-ecological systems
33 Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management
34 Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change
35 Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction

---

2023 01 Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions
02 Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts
03 Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations
04 Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval
05 Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications
06 António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment
07 Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning
08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning
09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques
10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing
11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications
12 Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries
13 Injy Sarhan (UU), Open Information Extraction for Knowledge Representation
14 Selma Čaušević (TUD), Energy resilience through self-organization
15 Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models
16 Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters
17 Meike Nauta (UT), Explainable AI and Interpretable Computer Vision – From Oversight to Insight
18 Gustavo Penha (TUD), Designing and Diagnosing Models for Conversational Search and Recommendation
19 George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals
20 Arkadiy Dushatskiy (TUD), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning
21 Gerrit Jan de Bruin (UL), Network Analysis Methods for Smart Inspection in the Transport Domain
22 Alireza Shojaifar (UU), Volitional Cybersecurity
23 Theo Theunissen (UU), Documentation in Continuous Software Development
24 Agathe Balayn (TUD), Practices Towards Hazardous Failure Diagnosis in Machine Learning
25 Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs
26 Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour
27 Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions
28 Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts
29 Tim Draws (TUD), Understanding Viewpoint Biases in Web Search Results

---

2024 01 Daphne Miedema (TU/e), On Learning SQL: Disentangling concepts in data systems education
02 Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems
03 Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis
04 Mike Huisman (UL), Understanding Deep Meta-Learning
05 Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair
06 Azqa Nadeem (TUD), Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence
07 Parisa Shayan (TiU), Modeling User Behavior in Learning Management Systems
08 Xin Zhou (UvA), From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation
09 Giso Dal (UT), Probabilistic Inference Using Partitioned Bayesian Networks
10 Cristina-Iulia Bucur (VUA), Linkflows: Towards Genuine Semantic Publishing in Science
11 withdrawn
12 Peide Zhu (TUD), Towards Robust Automatic Question Generation For Learning
13 Enrico Liscio (TUD), Context-Specific Value Inference via Hybrid Intelligence
14 Larissa Capobianco Shimomura (TU/e), On Graph Generating Dependencies and their Applications in Data Profiling
15 Ting Liu (VUA), A Gut Feeling: Biomedical Knowledge Graphs for Interrelating the Gut Microbiome and Mental Health
16 Arthur Barbosa Câmara (TUD), Designing Search-as-Learning Systems
17 Razieh Alidoosti (VUA), Ethics-aware Software Architecture Design
18 Laurens Stoop (UU), Data Driven Understanding of Energy-Meteorological Variability and its Impact on Energy System Operations
19 Azadeh Mozafari Mehr (TU/e), Multi-perspective Conformance Checking: Identifying and Understanding Patterns of Anomalous Behavior
20 Ritsart Anne Plantenga (UL), Omgang met Regels
21 Federica Vinella (UU), Crowdsourcing User-Centered Teams
22 Zeynep Ozturk Yurt (TU/e), Beyond Routine: Extending BPM for Knowledge-Intensive Processes with Controllable Dynamic Contexts
23 Jie Luo (VUA), Lamarck's Revenge: Inheritance of Learned Traits Improves Robot Evolution
24 Nirmal Roy (TUD), Exploring the effects of interactive interfaces on user search behaviour
25 Alisa Rieger (TUD), Striving for Responsible Opinion Formation in Web Search on Debated Topics
26 Tim Gubner (CWI), Adaptively Generating Heterogeneous Execution Strategies using the VOILA Framework
27 Lincen Yang (UL), Information-theoretic Partition-based Models for Interpretable Machine Learning
28 Leon Helwerda (UL), Grip on Software: Understanding development progress of Scrum sprints and backlogs

29 David Wilson Romero Guzman (VUA), The Good, the Efficient and the Inductive Biases: Exploring Efficiency in Deep Learning Through the Use of Inductive Biases
30 Vijanti Ramautar (UU), Model-Driven Sustainability Accounting
31 Ziyu Li (TUD), On the Utility of Metadata to Optimize Machine Learning Workflows
32 Vinicius Stein Dani (UU), The Alpha and Omega of Process Mining
33 Siddharth Mehrotra (TUD), Designing for Appropriate Trust in Human-AI interaction
34 Robert Deckers (VUA), From Smallest Software Particle to System Specification - MuDForM: Multi-Domain Formalization Method
35 Sicui Zhang (TU/e), Methods of Detecting Clinical Deviations with Process Mining: a fuzzy set approach
36 Thomas Mulder (TU/e), Optimization of Recursive Queries on Graphs
37 James Graham Nevin (UvA), The Ramifications of Data Handling for Computational Models
38 Christos Koutras (TUD), Tabular Schema Matching for Modern Settings
39 Paola Lara Machado (TU/e), The Nexus between Business Models and Operating Models: From Conceptual Understanding to Actionable Guidance
40 Montijn van de Ven (TU/e), Guiding the Definition of Key Performance Indicators for Business Models
41 Georgios Siachamis (TUD), Adaptivity for Streaming Dataflow Engines
42 Emmeke Veltmeijer (VUA), Small Groups, Big Insights: Understanding the Crowd through Expressive Subgroup Analysis
43 Cedric Waterschoot (KNAW Meertens Instituut), The Constructive Conundrum: Computational Approaches to Facilitate Constructive Commenting on Online News Platforms
44 Marcel Schmitz (OU), Towards learning analytics-supported learning design
45 Sara Salimzadeh (TUD), Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making
46 Georgios Stathis (Leiden University), Preventing Disputes: Preventive Logic, Law & Technology
47 Daniel Daza (VUA), Exploiting Subgraphs and Attributes for Representation Learning on Knowledge Graphs
48 Ioannis Petros Samiotis (TUD), Crowd-Assisted Annotation of Classical Music Compositions

---

2025 01 Max van Haastrecht (UL), Transdisciplinary Perspectives on Validity: Bridging the Gap Between Design and Implementation for Technology-Enhanced Learning Systems
02 Jurgen van den Hoogen (JADS), Time Series Analysis Using Convolutional Neural Networks
03 Andra-Denis Ionescu (TUD), Feature Discovery for Data-Centric AI
04 Rianne Schouten (TU/e), Exceptional Model Mining for Hierarchical Data
05 Nele Albers (TUD), Psychology-Informed Reinforcement Learning for Situated Virtual Coaching in Smoking Cessation
06 Daniël Vos (TUD), Decision Tree Learning: Algorithms for Robust Prediction and Policy Optimization
07 Ricky Maulana Fajri (TU/e), Towards Safer Active Learning: Dealing with Unwanted Biases, Graph-Structured Data, Adversary, and Data Imbalance
08 Stefan Bloemheuvel (TiU), Spatio-Temporal Analysis Through Graphs: Predictive Modeling and Graph Construction
09 Fadime Kaya (VUA), Decentralized Governance Design - A Model-Based Approach
10 Zhao Yang (UL), Enhancing Autonomy and Efficiency in Goal-Conditioned Reinforcement Learning
11 Shahin Sharifi Noorian (TUD), From Recognition to Understanding: Enriching Visual Models Through Multi-Modal Semantic Integration
12 Lijun Lyu (TUD), Interpretability in Neural Information Retrieval
13 Fuda van Diggelen (VUA), Robots Need Some Education: on the complexity of learning in evolutionary robotics
14 Gennaro Gala (TU/e), Probabilistic Generative Modeling with Latent Variable Hierarchies
15 Michiel van der Meer (UL), Opinion Diversity through Hybrid Intelligence
16 Monika Grewal (TU Delft), Deep Learning for Landmark Detection, Segmentation, and Multi-Objective Deformable Registration in Medical Imaging
17 Matteo De Carlo (VUA), Real Robot Reproduction: Towards Evolving Robotic Ecosystems
18 Anouk Neerincx (UU), Robots That Care: How Social Robots Can Boost Children's Mental Wellbeing
19 Fang Hou (UU), Trust in Software Ecosystems
20 Alexander Melchior (UU), Modelling for Policy is More Than Policy Modelling (The Useful Application of Agent-Based Modelling in Complex Policy Processes)
21 Mandani Ntekouli (UM), Bridging Individual and Group Perspectives in Psychopathology: Computational Modeling Approaches using Ecological Momentary Assessment Data
22 Hilde Weerts (TU/e), Decoding Algorithmic Fairness: Towards Interdisciplinary Understanding of Fairness and Discrimination in Algorithmic Decision-Making
23 Roderick van der Weerdt (VUA), IoT Measurement Knowledge Graphs: Constructing, Working and Learning with IoT Measurement Data as a Knowledge Graph
24 Zhong Li (UL), Trustworthy Anomaly Detection for Smart Manufacturing
25 Kyana van Eijndhoven (TiU), A Breakdown of Breakdowns: Multi-Level Team Coordination Dynamics under Stressful Conditions
26 Tom Pepels (UM), Monte-Carlo Tree Search is Work in Progress
27 Danil Provodin (JADS, TU/e), Sequential Decision Making Under Complex Feedback
28 Jinke He (TU Delft), Exploring Learned Abstract Models for Efficient Planning and Learning
29 Erik van Haeringen (VUA), Mixed Feelings: Simulating Emotion Contagion in Groups
30 Myrthe Reuver (VUA), A Puzzle of Perspectives: Interdisciplinary Language Technology for Responsible News Recommendation
31 Gebrekirstos Gebreselassie Gebremeskel (RUN), Spotlight on Recommender Systems: Contributions to Selected Components in the Recommendation Pipeline
32 Ryan Brate (UU), Words Matter: A Computational Toolkit for Charged Terms
33 Merle Reimann (VUA), Speaking the Same Language: Spoken Capability Communication in Human-Agent and Human-Robot Interaction
34 Eduard C. Groen (UU), Crowd-Based Requirements Engineering
35 Urja Khurana (VUA), From Concept To Impact: Toward More Robust Language Model Deployment
36 Anna Maria Wegmann (UU), Say the Same but Differently: Computational Approaches to Stylistic Variation and Paraphrasing
37 Chris Kamphuis (RUN), Exploring Relations and Graphs for Information Retrieval
38 Valentina Maccatrozzo (VUA), Break the Bubble: Semantic Patterns for Serendipity
39 Dimitrios Alivanistos (VUA), Knowledge Graphs & Transformers for Hypothesis Generation: Accelerating Scientific Discovery in the Era of Artificial Intelligence
40 Stefan Grafberger (UvA), Declarative Machine Learning Pipeline Management via Logical Query Plans
41 Mozhgan Vazifehdoostirani (TU/e), Leveraging Process Flexibility to Improve Process Outcome - From Descriptive Analytics to Actionable Insights
42 Margherita Martorana (VUA), Semantic Interpretation of Dataless Tables: a metadata-driven approach for findable, accessible, interoperable and reusable restricted access data
43 Krist Shingjergji (OU), Sense the Classroom - Using AI to Detect and Respond to Learning-Centered Affective States in Online Education
44 Robbert Reijnen (TU/e), Dynamic Algorithm Configuration for Machine Scheduling Using Deep Reinforcement Learning
45 Anjana Mohandas Sheeladevi (VUA), Occupant-Centric Energy Management: Balancing Privacy, Well-being and Sustainability in Smart Buildings