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

Kouwenhoven, T.

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

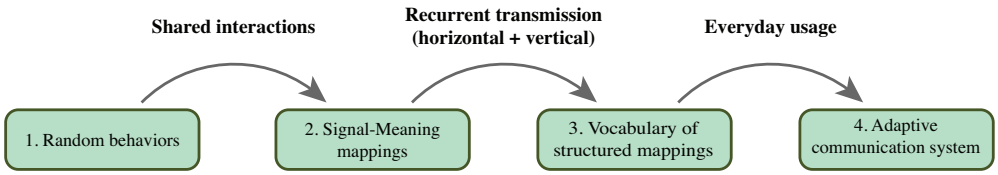

Figure 1.1: The various steps of evolving natural communication systems. First, initially random behaviours obtain meanings and become more structured through recurrent horizontal and vertical transmission. Everyday usage facilitates the continuous evolution of communication systems, which will adapt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

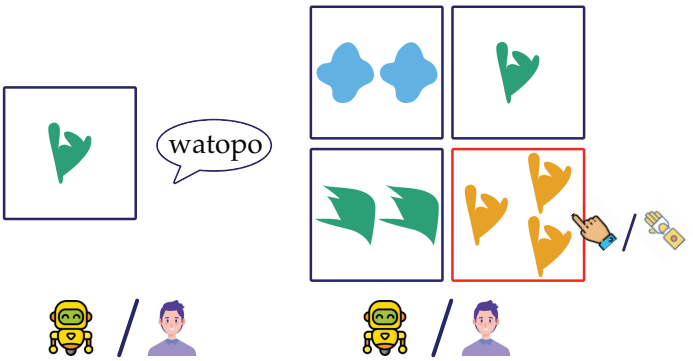

Figure 1.2: An exemplar setup in which humans and machines collaborate in a referential game as used in Chapter 6 and 7. In these games, a speaker (human or machine) utters a non-existing word which the listener uses to guess the target. Repeated interactions offer a means to establish initial conventions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

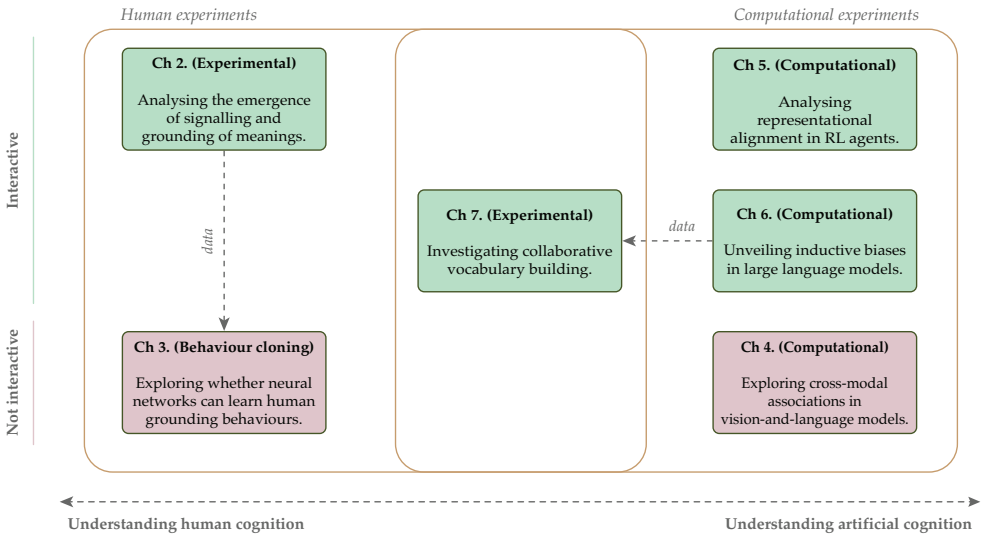

Figure 1.3: The structure of this dissertation as constituted by six chapters and their themes. Block colours indicate whether the Chapter incorporated interactions (green) or constituted a non-interactive approach (pink).

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.