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

Kouwenhoven, T.

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).