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

Kouwenhoven, T.

Summary

Language is a uniquely human trait that enables us to collaborate and exchange ideas on a daily basis. Although it is now taken for granted that we understand each other's language, the way this shared understanding came to be remains a mystery. A prominent theory in language evolution proposes that repeated interactions provide anchor points where signals and meanings become linked. Cultural processes can then spread these meaningful signals across populations. Together, these processes help shape a language such that it aligns with users' cognitive preferences, such as a preference for simple and structured signals. Sustained pressure from such preferences results in a shared language that is structured, expressive, and learnable.

A new type of language user is playing an increasingly prominent role in our society. Artificially intelligent systems, such as Large Language Models, can now be considered mature language users. However, the way they make decisions fundamentally differs from how humans do. Therefore, it is essential to investigate these largely unknown forms of cognition. This dissertation does so from the perspective of language evolution. It uses methods from that field, which are not only applicable to humans but are also suitable for unravelling cognitive patterns and preferences in non-human systems. The resulting findings contribute to our understanding of language evolution and inform us about how non-human cognition processes linguistic and visual information. In doing so, this dissertation addresses the question of how human and non-human cognition can complement each other in research on the evolution of language.

First, it examines how rudimentary signals emerge and the role that neural networks can play in this process. Experiments with humans confirm that shared rudimentary signals arise from repeated interactions and that it can help when conversation partners differ in their need for structure. This latter point nuances existing theory, which suggests that shared preferences are advantageous. By simulating human behaviour in this task using computational models, we see that bidirectional mechanisms best explain this behaviour–meaning that focusing on the surrounding context (both before and after) of a communicative signal is essential, just as it is in language processing.

Next, the dissertation investigates whether multimodal models display human-like cross-modal associations and studies, using self-learning models, the evolution of structured language

built from rudimentary elements. Where humans consistently name a rounded object "bouba" instead of "kiki," four multimodal models barely exhibit this pattern. This confirms that modern multimodal models struggle to link visual and textual information. Even when these models play a communication game where they develop their own language, they use different visual features than humans do. As a result, they are unable to distinguish between two images containing the same types of objects arranged differently–even when explicitly trained to learn the correct features.

Finally, the dissertation explores whether modern language models can serve as participants in language evolution experiments and whether humans and machines can collaboratively learn a language. When language models repeatedly interact in a communication game, a language emerges with compositional properties in which syllables are recombined to describe different objects. This bears striking similarities to human language evolution, even though the underlying mechanisms differ fundamentally. Simulations involving cultural transmission–where successive generations learn and use the language–also show that the language adapts to the user preferences of the language models in a way similar to how it does with humans. Still, there are differences: language models more frequently produce a single signal with multiple meanings (homonyms), and longer signals, whereas humans tend to prefer short, expressive utterances. The final contribution of this dissertation is an experiment in which participants collaborate with a Llama3-70B language model to develop a language. Despite the fundamentally different mechanisms for language acquisition and use, expressive and meaningful languages emerge. These languages contain compositional properties and show more human-like traits than languages developed without human involvement. These results support the idea that language adapts to its users and suggest that interactions, where communicative success matters, can play a role in natural language acquisition by machines.

Through this interdisciplinary approach to language evolution, this dissertation contributes to a line of research that studies both human and non-human cognition from a human-centred perspective. It shows that both forms of cognition can complement each other precisely because they differ. This finding offers promising possibilities for the development of communication systems adapted to the cognitive capacities of both humans and machines, potentially enabling more natural interactions between them.