Communication and Linguistic Structure in Collaborative Human-Machine Language Evolution

Tom Kouwenhoven*[1], Neval Kara[2], and Tessa Verhoef[1]

*Corresponding Author: t.kouwenhoven@liacs.leidenuniv.nl
[1]Creative Intelligence Lab, Leiden University, Leiden, The Netherlands
[2]Department of Psychology, Çankaya University, Ankara, Turkey

The compositionality of human language allows us to combine meaningful words into more complex meanings. The emergence of compositionality is extensively examined through human experiments (e.g., Kirby et al., 2008, 2015; Raviv et al., 2019) and (agent-based) simulations (e.g., Kirby, 1998; Brighton, 2002; Vogt, 2005; Lazaridou & Baroni, 2020). The latter is seeing increased attention due to computational advances and a rising interest into large language models (LLM). Although the behaviour of LLMs is fundamentally different from humans, their linguistic abilities are unprecedented, rendering them the first close comparators of language users. Moreover, LLMs are capable of in-context learning, i.e., having the model tackle a novel task based on a few examples in the prompt (Brown et al., 2020). In this pilot study, we examine whether LLMs can act as controlled variables in experiments on language evolution. Specifically, we assess if linguistic structure evolves when participants ($n = 10$) communicate with an LLM (*text-davinci-003*, temperature 0.0) in the Lewis signalling game.

The setup of our pilot is based on that of Kirby et al. (2008) and Raviv et al. (2019). Participants go through an exposure phase to learn an initially holistic artificial language, followed by a labelling phase in which they type labels for each object. The LLM learns artificial languages through the in-context learning method used by Galke et al. (2023), who showed that LLMs can learn artificial languages and that, similar to experimental findings, systematic generalisation was higher for more structured languages. Following the labelling phase, the participants and LLM alternate roles (speaker, listener) and use the learned language to communicate for six rounds in a referential task with four objects and one target. Here, the LLM must generalise labels for unseen stimuli based on its context, i.e., the vocabulary. Similar to Raviv et al. (2019), the meaning space expands by three objects after each communication round, starting with 12 objects. Finally, the participant labels each scene in the naming phase. Accuracy is the percentage of correctly identified objects. Identical to Kirby et al. (2008), z-scores of the Mantel test (Mantel, 1967) indicate the degree of structure in the vocabulary.

Results

Although learning the initial holistic language proves difficult, participants can identify objects based on the labels to some extent ($\approx 60\%$ accuracy) after the exposure phase. However, using this language for communication is challenging (figure 1, left) since accuracy is not better than chance and does not improve over the rounds. Across the four phases of the experiment, we observe an increase in structure in the labels produced by human participants. Yet, the overall structure does not increase as radically as previously found in dyadic interactions, even though in principle the LLM can generalise over such artificial languages (Galke et al., 2023). We expect this is due to the computational mechanisms used to generate responses. While humans update their beliefs following interactions, LLMs only have access to the prompt containing the vocabulary and question at hand but do not integrate past experiences (e.g., (un)successful rounds) or reason about the others' state of mind. Our preliminary findings suggest that linguistic structure does not radically change over time, thereby not following findings from earlier experiments (Kirby et al., 2008; Raviv et al., 2019) with humans alone.

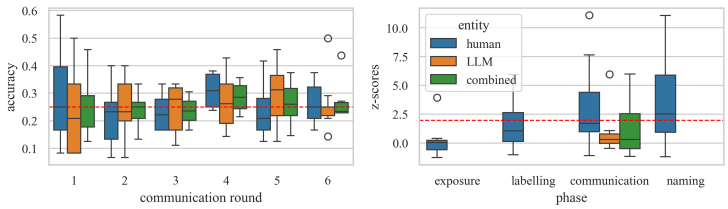

Figure 1. Accuracy over communication rounds (left) and the degree of structure (right) in each phase (only round 6 for the communication phase). The colour indicates whether predictions are from the human (blue), the LLM (orange), or both (green) entities. The red line indicates *chance* performance (left) and the threshold for which structure is likely to be caused by the entities instead of chance (right). Structure scores can be compared between exposure-labelling and communication-naming.

These results show that, although communication is difficult, the vocabularies do not completely collapse when used in human-machine communication. This is promising given that experiments with human-human dyads have shown that interaction dynamics and personal differences affect how language evolves (Verhoef et al., 2022; Kouwenhoven et al., 2022). Moreover, it may suggest that mere presence of a communicative partner–even one that is bad–can push the participant to increase the systematicity of its productions. While this work is preliminary and only a first step in human-machine language evolution, it opens possibilities for future work in which one communicative partner can be relatively fixed and kept under experimental control. For example, in more complex setups like iterated learning or by comparing languages from human-human with human-computer experiments to isolate biases of a single human in cooperative settings.

References

Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial life*, *8*(1), 25–54.

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

Galke, L., Ram, Y., & Raviv, L. (2023). What makes a language easy to deep-learn? *arXiv preprint arXiv:2302.12239*.

Kirby, S. (1998). Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681-10686.

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

Kouwenhoven, T., Kleijn, R. de, Raaijmakers, S., & Verhoef, T. (2022). Need for structure and the emergence of communication. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Lazaridou, A., & Baroni, M. (2020). Emergent multi-agent communication in the deep learning era. *arXiv preprint arXiv:2006.02419*.

Mantel, N. (1967). The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research*, *27*, 209-220.

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

Verhoef, T., Walker, E., & Marghetis, T. (2022). Interaction dynamics affect the emergence of compositional structure in cultural transmission of space-time mappings. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Vogt, P. (2005). The emergence of compositional structures in perceptually grounded language games. *Artificial Intelligence*, *167*(1), 206-242. (Connecting Language to the World)