Emergent Communication in Multi-Agent Systems: Compositionality and Interpretability of Emergent Languages

Author: Bastien Vanderplaetse

Supervisors: Prof. Stéphane Dupont & Prof. Xavier Siebert

Institution: University of Mons, Faculty of Engineering

The field of emergent communication in multi-agent systems (MAS) explores how autonomous agents develop unique communication protocols without human intervention. This research field intersects artificial intelligence, computational linguistics, and robotics, addressing the challenge of enabling autonomous agents to develop and utilize complex communication protocols. These protocols refer to a set of rules, conventions, and structures that governs how information is exchanged between agents in a MAS. They define the format, timing, sequencing, and semantics of messages, ensuring that the agents can effectively interpret and respond to the information they receive.

More specifically, emergent communication protocols focus on the functional aspects of communication, such as how messages are encoded, transmitted, and decoded. These protocols govern the mechanics of interaction, ensuring that agents can exchange information efficiently and reliably. Communication protocols require a language, which refers to the symbolic system or vocabulary that agents develop to represent and convey meaning. It encompasses the creation of symbols, grammar, and shared understanding, allowing agents to express complex ideas and concepts. While emergent communication protocols deal with the "how" of communication, emergent language deals with the "what" and the shared understanding that emerges from interactions.

Emergent communication uses controlled interactions to observe how communication protocols emerge from scratch. These artificial settings allow researchers to systematically study phenomena that would be difficult to observe in real-world systems. Therefore, emergent communication serves as a simulation tool to study and understand the origins and the evolution of human language, which is challenging to examine directly due to its ancient roots [1, 2]. Most research in this area focuses on knowledge transmission over generations, i.e. introducing new agents into an already trained system to learn an already emerged language [3 – 8]. However, studying how emergent language evolves during the training of MAS requires to focus on two important components of emergent languages: compositionality and interpretability. These insights could help to design better environments and tasks for training agents to communicate effectively.

The compositionality of the language is the principle that complex meanings of utterances are derived from the meanings of their components [2]. For example, the message “red car” means “a car that is red”, combining the meanings of “red” and “car”. Compositionality is an important aspect in multi-agent communication since three important phenomena in human language are related to the emergence of a compositional structure, all three being discussed theoretically and confirmed experimentally in the literature [9]:

- Compositional languages are easier to learn [10 – 12].
- Compositional languages enhance the agents' ability for generalization [12].
- Larger populations typically develop rule-based languages [12, 13].

To evaluate compositionality, we usually use topographic similarity (TopSim) [14, 15]. This metric measures the correlation between distances in the referent feature space and distances in the message space. However, this metric relies on strong assumptions [1], such as the selection of a specific distance metric and the use of linear correlation. While most studies evaluate compositionality in controlled settings with artificial data, these assumptions may not hold when transitioning to real-world data. When establishing particular assumptions, TopSim could be low, but it does not imply that there is no compositionality in the emergent language. The agents could use others properties of the input data that were not considered in the assumptions.

This leads us to the first research question of this thesis:

RQ1: What influences the compositionality of emergent languages?

Recent advances have highlighted the flexibility of these emergent protocols, showcasing their ability to improve coordination and collaboration across diverse tasks and scenarios [16, 17]. Current research primarily focuses on improving the overall performance of multi-agent systems, with additional emphasis on the resilience and adaptability of emergent protocols under various conditions [17, 18]. While emergent protocols improve agent collaboration, a notable gap in the literature lies in the interpretability of emergent languages and their alignment with human understanding, which is a critical aspect for ensuring the usability and trust of such systems in real-world contexts [19]. Some research works propose to design adaptive languages that enhance the clarity and sparsity of messages to meet the needs of diverse human-agent teams [20]. Others works explore the development of communication strategies more aligned with human language [21], specifically on its compositionality property [22]. However, the interpretation of language emerging from such techniques requires a lot of manual analysis, which leads to the second research question of this thesis:

RQ2: How to easily interpret emergent languages?

Contributions of this thesis:

This thesis employs reinforcement learning (RL) and deep neural networks to train agents in a cooperative settings, focusing on the Lewis Game, a benchmark where a *Speaker* generates messages about images and a *Listener* identifies the correct image based on these messages. The agents are trained on multiple datasets:

- Multi-Object Positional Relationship Dataset (MOPRD) [23]: this dataset aims to train agents in the communication of object relationships. Each image features two shapes arranged in a predefined relationship, with a random position and rotation. MOPRD incorporates five distinct shapes and four differet relationships.
- Colored Multi-Object Positional Relationship Dataset (C-MOPRD): this dataset is an extension of MOPRD that we conceive in this thesis. C-MOPRD adds a color to the shapes (red, green, blue or white). The goal is to propose a dataset with more properties available for the agents to convey in their messages.
- Visual Genome Human-Animal-Circular (VGHAC): VGHAC is a dataset we constituted based on a specific sampling over the Visual Genome (VG) dataset. VG is designed to enhance image understanding by providing detailed annotations for a large number of real world pictures [24]. Based on specific criteria, we developed VGHAC by selecting a restraint number of objects from VG, which resulted in the selection of six objects grouped into three categories:
    - Humans: man, woman, person;
    - Animals: bear, cat, giraffe;
    - Circular objects: pizza, plate, wheel.

  With VGHAC, we propose a dataset with more realistic data with a rich annotation, giving more information for interpretation of emergent languages.

- Manga109s Dataset [25]: this dataset is a collection of 109 manga volumes annotated for computer vision tasks. It focuses on semantic segmentation, providing pixel-level annotations. Based on these data, we create a new dataset where each image is a manga panel (a box). This dataset provides a unique structure combining text, graphics, and layered storytelling elements that challenges agent to develop more nuanced representations.

For RQ1, we examine how different image encoders and image feature processing methods influence agent performance and the compositionality of emergent languages. These experiments have highlighted that the choice of the processing applied to the image features influences the information that the agents use, and therefore it influences which metric distance should be used to get a TopSim that is more related to the emergent language compositionality.

Publication related to RQ1:

- B. Vanderplaetse, S. Dupont, and X. Siebert, “Influence of image encoders and image features transformations in emergent communication”, in *ESANN 2024 proceedings*. Ciaco – i6doc.com, 2024, pp. 685-690.

For RQ2, we propose the *Automated Semantic Rules Detection* (ASRD) algorithm. This algorithm is designed to help at the interpretation of emergent languages in MAS by automatically extracting semantic rules from messages exchanged by the agents. More specifically, the algorithm identifies patterns in messages and links them to specific attributes and hyperattributes of the image data. ASRD addresses a gap in the literature about interpretability of emergent languages, since it provides a tool for automated analysis of these languages, replacing the manual analysis done in previous works to try to understand the emergent languages.

Publication related to RQ2:

- B. Vanderplaetse, X. Siebert, and S. Dupont, "Automated Semantic Rules Detection (ASRD) for Emergent Communication Interpretation", in *EXTRAAMAS 2025* [Accepted]

Finally, through this thesis, we propose an open-source, plug-and-play framework for experimenting easily with emergent communication in multi-agent systems[1].

[1] A GitHub link will be provided.

Bibliography

[1] Rahma Chaabouni, Florian Strub, Florent Altché, Eugene Tarassov, Corentin Tallec, Elnaz Davoodi, Kory Wallace Mathewson, Olivier Tieleman, Angeliki Lazaridou, and Bilal Piot. Emergent Communication at Scale. January 2022.

[2] Brendon Boldt and David R. Mortensen. A Review of the Applications of Deep Learning-Based Emergent Communication. *Transactions on Machine Learning Research*, August 2023.

[3] Niko Grupen, Daniel Lee, and Bart Selman. *Curriculum-Driven Multi-Agent Learning and the Role of Implicit Communication in Teamwork*. June 2021.

[4] Fushan Li and Michael Bowling. Ease-of-Teaching and Language Structure from Emergent Communication. In *Advances in Neural Information Processing Systems*, 2019.

[5] Yi Ren, Shangmin Guo, Matthieu Labeau, Shay B. Cohen, and Simon Kirby. Compositional Languages Emerge in a Neural Iterated Learning Model, February 2020. arXiv:2002.01365[cs].

[6] Kenny Smith, Simon Kirby, and Henry Brighton. Iterated learning: a framework for the emergence of language. *Artificial Life*, 9(4):371–386, 2003.

[7] Simon Kirby and James R. Hurford. The Emergence of Linguistic Structure: An Overview of the Iterated Learning Model. In Angelo Cangelosi and Domenico Parisi, editors, *Simulating the Evolution of Language*, pages 121–147. Springer London, London, 2002.

[8] Simon Kirby, Hannah Cornish, and Kenny Smith. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences of the United States of America*, 105:10681–6, September 2008.

[9] Lukas Galke, Yoav Ram, and Limor Raviv. Emergent Communication for Understanding Human Language Evolution: What's Missing?, April 2022. arXiv:2204.10590 [cs].

[10] Simon Kirby, Tom Griffiths, and Kenny Smith. Iterated learning and the evolution of language. Current Opinion in Neurobiology, 28:108–114, October 2014.

[11] Jon Carr, Kenny Smith, Hannah Cornish, and Simon Kirby. The Cultural Evolution of Structured Languages in an Open-Ended, Continuous World. Cognitive Science, 41:892–923, May 2017.

[12] Limor Raviv, Marianne de Heer Kloots, and Antje Meyer. What makes a language easy to learn? A preregistered study on how systematic structure and community size affect language learnability. Cognition, 210:104620, May 2021.

[13] Gary Lupyan and Rick Dale. Language structure is partly determined by social structure. *PloS One*, 5(1):e8559, January 2010.

[14] Henry Brighton and Simon Kirby. Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artificial Life*, 12(2):229–242, 2006.

[15] Angeliki Lazaridou, Karl Hermann, Karl Tuyls, and Stephen Clark. Emergence of Linguistic Communication from Referential Games with Symbolic and Pixel Input. April 2018.

[16] Elías Masquil, Gautier Hamon, Eleni Nisioti, and Clément Moulin-Frier. Intrinsically-Motivated Goal-Conditioned Reinforcement Learning in Multi-Agent Environments. November 2022.

[17] Yuqi Wang, Xu-Yao Zhang, Cheng-Lin Liu, and Zhaoxiang Zhang. Emergence of Machine Language: Towards Symbolic Intelligence with Neural Networks. January 2022.

[18] Kalesha Bullard, Douwe Kiela, Franziska Meier, Joelle Pineau, and Jakob Foerster. Quasi-Equivalence Discovery for Zero-Shot Emergent Communication, June 2021. arXiv:2103.08067 [cs].

[19] Changxi Zhu, Mehdi Dastani, and Shihan Wang. A Survey of Multi-Agent Reinforcement Learning with Communication, March 2022. arXiv:2203.08975 [cs].

[20] Seth Karten, Mycal Tucker, Huao Li, Siva Kailas, Michael Lewis, and Katia Sycara. Interpretable Learned Emergent Communication for Human–Agent Teams. *IEEE Transactions on Cognitive and Developmental Systems*, 15(4):1801–1811, 2023. Conference Name: IEEE Transactions on Cognitive and Developmental Systems.

[21] Diane Bouchacourt and Marco Baroni. How agents see things: On visual representations in an emergent language game. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 981–985, Brussels, Belgium, 2018. Association for Computational Linguistics.

[22] Igor Mordatch and Pieter Abbeel. Emergence of Grounded Compositional Language in Multi-Agent Populations, 2018.

[23] Y. Feng, B. An, and Z. Lu, “Learning Multi-Object Positional Relationships via Emergent Communication”, 2023. [Online]. Available: http://arxiv.org/abs/2302.08084

[24] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, “Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations”, *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, May 2017. [Online]. Available: https://doi.org/10.1007/s11263-016-0981-7

[25] K. Aizawa, A. Fujimoto, A. Otsubo, T. Ogawa, Y. Matsui, K. Tsubota, and H. Ikuta, “Building a Manga Dataset “Manga109” With Annotations for Multimedia Applications”, *IEEE MultiMedia*, vol. 27, no. 2, pp. 8–18, Apr. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9069265