Machine Learning and the Evolution of Language

Building a bridge between communities

Workshop organizers: *Mathieu Rita, Lukas Galke, Dr. Florian Strub, Prof. Olivier Pietquin, Prof. Emmanuel Dupoux, Prof. Bart de Boer, Dr. Limor Raviv*

Contact Email: ml4evolang@mpi.nl

Schedule

Morning session (10am–12pm local time)	
10:00	**Welcome & quick intro - (Organizing committee)**
10:05	**Introduction to machine learning methods - Florian Strub**
10:40	**Invited Talk: Machine Learning and Natural Language Processing - Douwe Kiela**
11:05	**Invited Talk: Designing agent-based models of sign language communities - Katie Mudd**
11:25	***Flash talks for spotlight posters (5x 5+2 minutes each)***
Afternoon session (1:30pm–3:30pm local time)	
13:30	**Invited Talk: Emerging linguistic universals in communicating neural network agents - Rahma Chaabouni**
13:55	**Invited Talk: Agent-Based Modeling of the cultural evolution of linguistic complexity is really hard - Matt Spike**
14:15	**Machine Learning for Evolang: Lessons from the Past - Bart de Boer**
14:30	**Emergent Communication and Language Evolution: What's Missing? - Lukas Galke**
14:45	**Panel discussion: (chair: Limor Raviv)** - What are the benefits of AI compared to classic agent-based models? How can they contribute to the study of language evolution? - What are the current problems in using deep learning methods to study evolang? Where are the gaps? - How can AI researchers make their work more useful/tangible for linguists? - How can language experts inform the AI community working on emergent communication?
15:25	**Final remarks - Mathieu Rita**

Workshop Description

The goal of this workshop is to build a bridge between the language evolution community and the machine learning (ML) community. Although both areas of research have similar interests and work on similar questions, there has been little to no crosstalk between them so far. This is unfortunate, since the progress in ML and other areas of AI may allow language evolution researchers to model phenomena that they could not model before using classic agent-based computational models. At the same time, theoretical, computational, and experimental knowledge coming from the language evolution community may help focus and improve the emergent communication models used by the ML community. The goal of this workshop is therefore to relate these two areas by bringing together researchers from both backgrounds, establishing common ground, bootstrapping a mutual dialogue between them, and discussing the potential pitfalls of incorporating ML methods in the study of language evolution.

Invited Talks

Douwe Kiela

Douwe is the Head of Research at Hugging Face, the company that maintains the main platform for developing and publishing large language models. His works cover language embodiment, language evolution, and now focuses on large-scale language models, and the study of language biases. His presentation will showcase language evolution with deep reinforcement learning while drawing links with recent large scale language models.

Katie Mudd: Designing agent-based models of sign language communities

My research focuses on what linguistic features look like in language emergence and in the initial stages of a language, drawing inspiration from sign languages. One methodology that I use is computational modeling to assess theories about how social structure shapes linguistic features in the initial stages of a (sign) language. I design these models based on my analyses of linguistic variation in sign languages, my experience learning French Belgian Sign Language (LSFB), insights from sign language linguists and from sociolinguistic sketches about signing communities (e.g., Zeshan and de Vos, 2012). In this talk I will explain how I approached designing an agent-based model to study how community properties affect lexical variation in language emergence (Mudd et al., 2022). I will show how I try to design the simplest model possible for the question I am trying to answer, and how even such a model can lead to complex dynamics. I will then list model shortcomings and challenges that I face when designing and presenting such a model, with the aim of starting a discussion about complementary or alternative computational tools to study similar research questions.

Rahma Chaabouni: Emerging linguistic universals in communicating neural network agents

The ability to acquire and produce a language is a key component of intelligence. If communication is widespread among animals, human language is unique in its productivity and complexity. In this talk, I focus on works that build up on the emergent communication field to investigate the well-standing question of the source of natural language. In particular, these works use communicating deep neural networks that can develop a language to solve a collaborative task. Comparing the emergent language properties with human cross-linguistic regularities can provide answers to the crucial questions about the origin and evolution of natural language. Indeed, if neural networks develop a cross-linguistic regularity spontaneously, then the latter would not depend on specific biological constraints. Looking at neural networks as another expressive species can shed light on the source of cross-linguistic regularities – a fundamental research interest in cognitive science and linguistics.

I will focus on four cross-linguistic regularities related to word length, word order, semantic categorization, and compositionality. Across the different studies, we find that some of these regularities arise spontaneously while others are missing in neural networks' languages. We connect the former case to the presence of shared communicative constraints such as the discrete nature of the communication channel. On the latter, we relate the absence of human-like regularities to the lack of constraints either on the learners' side (e.g., the least-effort constraints) or language functionality (e.g., the transmission of information).

Matt Spike: Agent-Based Modeling of the cultural evolution of linguistic complexity is really hard

Questions of linguistic complexity, whether it varies within and across languages - and if so why - have a long and sometimes regrettable history. This probably shouldn't be a big surprise: it's hard enough to find agreement on what language actually is, and "complexity is so general a term that it seems to mean something different to everyone" (Adami, 2002). Despite the various conceptual and ethical issues, a growing empirical literature has driven a 'new consensus' (Joseph & Newmeyer, 2012) that at least *some* aspects of language are more or less complex in *some* languages than in others, and that this seems to be linked to various cultural and/or demographic factors.

This has, in turn, made room for agent-based modelers to test different *cultural evolutionary* theories of language, often with great success. What turns out to be really hard, on the other hand, is finding a way to reconcile all these positive accounts with the fact that, when it comes to their implementations as agent-based models, they represent a rather diverse range of interpretations of language, cultural evolution, and their basic units and mechanisms. Since an integrated approach would be a nice thing to have, I will sketch out a modeling framework for charting out different theories of the cultural evolution of language complexity, and thinking about how they might interact with each other.

Talks from the Organizers

Introduction to Machine Learning Methods

Florian Strub

In this presentation, we will go step-by-step in the machine learning basics, with a gentle introduction to deep neural networks. The goal is to provide all the necessary intuitions to understand the deep machine learning methods that may be introduced during the day. To ease intuition, we will explain the different approaches by using some examples from recent language modeling techniques and Lewis Game simulation.

Machine Learning for Evolang: Lessons from the Past

Bart de Boer

This talk will explore the past of computer modeling in language evolution and of neural network research in order to draw lessons from it. The talk will focus on two main lessons: 1) the past is deeper than many researchers are aware of and filled with interesting ideas that were abandoned at some point, but that could provide inspiration for new research with more modern tools. 2) Even though computers were many orders of magnitude less powerful, insightful models were built; this was done through a careful focus on what was essential complexity and what was not. The talk will briefly touch upon a number of models from the last 75 years to illustrate these lessons.

Emergent Communication and Language Evolution: What's missing?

Lukas Galke

Emergent communication protocols among humans and artificial neural network agents do not yet share the same properties and show some critical mismatches in results. In this talk, I present three important phenomena with respect to the emergence and benefits of compositionality: ease-of-learning, generalization, and group size effects (i.e., larger groups create more systematic languages). The latter two are not fully replicated with neural agents, which hinders the use of neural emergent communication for language evolution research. One possible reason for these mismatches is that key cognitive and communicative constraints of humans are not reflected in machine learning experiments. Specifically, in humans, memory constraints and the alternation between the roles of speaker and listener influence the emergence of linguistic structure, yet these constraints are typically absent in neural simulations. Introducing such communicative and cognitive constraints might promote more linguistically plausible behaviors of neural network agents.

Spotlight Posters

Competition exacerbates Language Drift

Michael Noukhovitch, Aaron Courville, Issam H. Laradji

Language drift is when a pretrained model diverges from its initial protocol after finetuning on an external, non-linguistic reward. This is a common problem in emergent communication, where pretrained language models diverge from natural language when trained in self-play. This phenomena is also connected to the linguistic drift in natural language and Wittgenstein's language games that form group-specific languages. The first paper on the subject in machine learning (Lewis et al, 2017) found significant drift when learning dialogue agents for negotiation. We propose that the drift found was more pronounced because negotiation is a competitive game and competition exacerbates language drift. To investigate, we use follow Noukhovitch et al. (2021) in creating a game with a smooth parameter that can set competition between fully cooperative and fully competitive. We evaluate the drift in the sender's protocol using the Jensen-Shannon divergence and look at changes between epochs over the course of training. To evaluate the overall drift, we use the area under the curve. Running 100 random hyperparameter searches of 5 seeds each, we find that a game's competitiveness does influence how much agents drift. Agents drift least in the fully cooperative game and the fully competitive game, and most in the middle of the two. This is logical as drift will occur when agents have something to gain from sharing information (cooperative reward) but must make sure that their information isn't fully exploited (competitive reward) and these pressures are balanced at the equally cooperative-competitive case. Since agents are usually initialized from pretrained models, we try initializing from fully cooperative pretrained agents (as opposed to random initialization) and find that similar results hold. When we initialize from agents pretrained on the equally cooperative-competitive game, we find the same results again. This demonstrates that competitive games drift more than fully cooperative regardless of pretraining, and drift is most pronounced when games are equally cooperative and competitive.

Challenges in Simulating the Word Order vs. Case Marking Trade-off with Neural Agents

Yuchen Lian, Arianna Bisazza, Tessa Verhoef

Natural languages commonly display a trade-off among different strategies to convey constituent roles (Sinnemäki, 2008; Futrell et al., 2015): flexible order typically correlates with the presence of case marking (e.g. in Russian, Tamil, Turkish), while fixed order is often observed in languages with little or no case marking (e.g. in English or Chinese). Chaabouni et al. (2019) designed a neural agent model with iterated language learning to study the emergence of such strategies, but failed to find a clear preference to avoid redundant coding strategies as natural languages do. Our recent work (Lian et al., 2021) re-evaluated this finding in light of three factors known to play an important role in comparable experiments and simulations in the Language Evolution field, namely:(i) a speaker bias towards efficient

messaging (i Cancho and Solé, 2003), (ii) unpredictable variation in the initial languages (Smith and Wonnacott, 2010; Fedzechkina et al., 2017), and (iii) the exposure of learners to a relatively small set of example utterances, also known as ‘learning bottleneck’ (Kirby et al., 2014). In all cases, our agents proved to be accurate learners, but strived to maintain the distribution of utterance types observed during learning instead of introducing generalizations or making the language more systematic. Specifically, the efficiency bias combined with highly unpredictable input led to a collapse of the communication system, whereas moderate variability combined with a learning bottleneck led to a stable language distribution. We concluded that the current neural-agent iterated learning framework is not yet ready to simulate language evolution processes in a human-like way. In real language use, a pressure for efficiency is balanced with communicative needs (Kirby et al., 2015; Regier et al., 2015), and this would normally not lead to a severe language degradation. Currently, we focus on leveraging advantages of both setups, i.e., to pre-train agents with predefined languages via supervised learning, and then let them play a cooperative reconstruction game where the likelihood of a message to be understood by the listener is used as a shared reward signal to update both agents via reinforcement learning.

References

Rahma Chaabouni, Eugene Kharitonov, Alessandro Lazaric, Emmanuel Dupoux, and Marco Baroni. 2019. Word-order biases in deep-agent emergent communication. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5166–5175, Florence, Italy. Association for Computational Linguistics.

Maryia Fedzechkina, Elissa L Newport, and T Florian Jaeger. 2017.Balancing effort and information transmission during language acquisition: Evidence from word order and case marking. Cognitive science, 41(2):416–446.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Quantifying word order freedom in dependency corpora. In Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015), pages 91–100, Uppsala, Sweden. Uppsala University, Uppsala, Sweden.

Ramon Ferrer i Cancho and Ricard V Solé. 2003.
Least effort and the origins of scaling in human language. Proceedings of the National Academy of Sciences, 100(3):788–791. Simon Kirby, Tom Griffiths, and Kenny Smith. 2014. Iterated learning and the evolution of language. Current opinion in neurobiology, 28:108–114.

Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. 2015. Compression and communication in the cultural evolution of linguistic structure. Cognition, 141:87–102.

Yuchen Lian, Arianna Bisazza, and Tessa Verhoef. 2021. The effect of efficient messaging and input variability on neural-agent iterated language learning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10121–10129, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Terry Regier, Charles Kemp, and Paul Kay. 2015. Word meanings across languages support efficient communication. The handbook of language emergence, 87:237.

Kaius Sinnemäki. 2008. Complexity trade-offs in core argument marking. Language complexity: Typology, contact, change, 67:88.

Kenny Smith and Elizabeth Wonnacott. 2010. Eliminating unpredictable variation through iterated learning. Cognition, 116(3):444–449.

BotTown: A platform for agent-based simulations of naturalistic conversations

Roberta Rocca, Rokas Maksevičius, Kenneth Christian Enevoldsen, Kristian Tylén

Dialogue and interaction are the primary modes of language use. We acquire language through interaction, and language itself evolves as a function of contextualized use. Language acquisition and change have traditionally been studied in experimental paradigms relying on simplified symbolic systems – with limited generalizability to complex symbolic phenomena –, or by analyzing large-scale diachronic corpora – with no affordances for controlled manipulation and causal inference. Here, we present an open-source platform for agent-based simulation of naturalistic conversations, supporting computational investigations of emergent linguistic phenomena in the context of complex linguistic behavior. The platform wraps generative models available through HuggingFace's transformers into a framework that allows users to easily define, through high-level commands: - an agent population, varying in size (number of agents = number of generative models), internal diversity (type of model), and network structure (likelihood to interact); - a set interactional constraints, e.g., stopping rules; - update rules, i.e., whether and how agents (i.e., generative language models) will be "updated" or fine-tuned on the output of interactions with other agents. Users can run multiple epochs of conversations within a population, and easily log text outputs and model checkpoints at each epoch to analyze the causal impact of parameter choices on characteristics of the outputs and of agents' linguistic behavior and internal representations. We envisage applications of interest for both language evolution, cognitive science and NLP researchers, for instance for the study of semantic change, opinion dynamics, synchronic and emergent properties of dialogue, and dynamic approaches to fine-tuning of LLMs in interactive settings. An alpha version of the platform is available (https://github.com/centre-for-humanities-computing/chatbot-conversations). We welcome input and contributors on both technical and theoretical aspects of the project.

Signalling signalhood in machine learning agents

Edward Hughes, Abhinav Gupta, Ekaterina Tolstaya, Thom Scott-Phillips

Background. Experimental language evolution investigates how pairs or groups interacting individuals create communicative conventions, and how those conventions develop some of the properties characteristic of natural languages. Here, we report ongoing work aiming to replicate a key task in this literature, commonly known as the Embodied Communication Game (Scott-Phillips et al., 2009; Kouwenhoven et al., 2022)—but with artificial agents rather than human participants. The distinctive feature of the Embodied Communication Game is that there is no a priori difference between communicative and non-communicative behaviour. This poses a boot-strapping challenge that may be important for artificial agents. Methods & Results. We characterize the Embodied Communication Game as a decentralized partially observable Markov decision process, and implement it as a multi-agent reinforcement learning environment. Training state-of-the-art reinforcement learning agents in one-shot episodes via self-play or population-play produces a fixed default color strategy that is not only not communicative, but also cannot generalize to new co-players in the 'other-play' setting (Hu et al. 2020). In few-shot episodes, which is a

memory-based meta-learning paradigm (Duan et al. 2016), agents find a local optimum of maximizing the number of rounds played, but do not achieve communication. Discussion. Comparison between these results and existing experimental results with human participants reveals deep and serious challenges for replicating human communicative competence in artificial agents. Speculatively, we suggest these tasks could be performed by agents who have (i) goals with respect to others' internal ('mental') states, and (ii) models of others' goals and the means by which those goals might be satisfied. Some limited progress has been made in this direction using Bayesian approaches (e.g. Ho et al., 2021), but such work is in its infancy and there is an enormous amount to be done. Whether these capacities can emerge via multi-agent reinforcement learning is presently unknown.

Emergent Communication as Decentralized Representation Learning: Metropolis-Hastings Naming Game and Beyond

Tadahiro Taniguchi, Yuto Yoshida, Kazuma Furukawa, Jun Inukai, Ryota Okumura, Yoshinobu Hagiwara, Akira Taniguchi

Semiotic communication is a crucial part of human cognitive capabilities. Recently, generative views of cognition, e.g., predictive processing, world modeling, and a free-energy principle, are attracting attention. Therefore, it is worth modeling emergent communication from the generative viewpoints. The authors proposed a Metropolis-Hastings naming game that enables multiple agents to form categories and share signs through the language game. Notably, the game can be regarded as decentralized Bayesian inference and representation learning from a mathematical viewpoint. In this poster presentation, we introduce the Metropolis-Hastings naming game and the outcomes of the prior works. Also, we discuss the possible further extensions of the approach, including a Metropolis-Hastings naming game among N agents and an experimental-semiotics approach based on the naming game.

References

[1] Taniguchi, T., Yoshida, Y., Taniguchi, A., & Hagiwara, Y. (2022). Emergent Communication through Metropolis-Hastings Naming Game with Deep Generative Models. arXiv preprint arXiv:2205.12392.
[2] Hagiwara, Y., Furukawa, K., Taniguchi, A., & Taniguchi, T. (2022). Multiagent multimodal categorization for symbol emergence: emergent communication via interpersonal cross-modal inference. Advanced Robotics, 36(5-6), 239-260.
[3] Hagiwara, Y., Kobayashi, H., Taniguchi, A., & Taniguchi, T. (2019). Symbol emergence as an interpersonal multimodal categorization. Frontiers in Robotics and AI, 6, 134.
[4] Taniguchi, T., Ugur, E., Hoffmann, M., Jamone, L., Nagai, T., Rosman, B., Matsuka, T., Iwahashi, N., Oztop, E., Piater, J., & Wörgötter, F. (2018). Symbol emergence in cognitive developmental systems: a survey. IEEE transactions on Cognitive and Developmental Systems, 11(4), 494-516.

Regular Posters

Creating a Baseline to Evaluate Correlations Between Language and Environment

Catherine Arnett, Maho Takahashi

Previous studies have proposed a correlation between environmental features such as elevation, humidity, and temperature and language features such as ejective consonants, complex tone systems, and consonant-vowel ratio (Everett, 2013; Everett et al., 2015; Everett, 2017), though the findings are contested (Urban and Moran, 2021; Roberts, 2018). Through a series of statistical analyses, we echo the concern that the correlation is not robust and could even be coincidental. After taking into account the random effects of language family and region with mixed-effects models (Bates, 2007), we failed to replicate two of the three correlations previously proposed to be significant; the only correlation that was replicated was the case of high humidity and the presence of tone (Estimate = 176.22, SE = 58.44, $p < 0.01$). Furthermore, we found significant correlations between environmental and linguistic features where the latter is highly unlikely to be influenced by the former, such as adposition-noun order ~ temperature (Estimate = -0.002, SE = 0.0005, $p < 0.01$) and the existence of numeral classifiers ~ elevation (Estimate = 0.98, SE = 0.29, $p < 0.001$). The agent-based model provides support for a more nuanced view of these relationships, where the language-environment effect may be modulated by other factors (Janssen and Dediu, 2018), such as diet (Blasi et al., 2019), community size (Raviv et al., 2019), articulator anatomy and genetic differences (Dediu and Moisik, 2019; Dediu et al., 2019, 2021; Butcher, 2018; Everett and Chen, 2021), interaction and cultural evolution (N¨olle et al., 2020), language contact (Moran et al., 2021), frequency (Macklin-Cordes and Round, 2020), and other physical and social factors (Bentz et al., 2018). To do this, we will find the correlation coefficients between the three environmental features mentioned above and a variety of language features, such as presence of certain sounds (e.g. lateral consonants) or relative order of verb and object. We will identify the distribution of these correlation coefficients, which will establish a baseline to compare the tone-humidity correlation to and allow us to test whether the correlation is significantly stronger than chance.

Long-term frequency shift in blue whale songs could be explained using flocking models and dynamical social network analysis

Franck Malige, Julie Patris, Pascale Giraudet, Maxime Hauray, Hervé Glotin

Blue whales emit complex and low frequency sounds, repited rythmically for hours, called songs, which are probably linked to reproduction (Oleson et al. 2007). Several distinct types of songs have been registered worldwide for this species (McDonald et al. 2006). All these blue whale song types undergo an almost linear decrease of the emitted frequencies over time which has been documented for more than sixty years for some of them (McDonald et al. 2009, Rice et al. 2022). This slow but constant evolution is an unexplained phenomenon. These last ten years, various hypothesis have been emitted to explained it, from recovering from hunting to an increase of anthropogenic noise (McDonald et al. 2009, Leroy et al. 2018). However none of the hypothesis presented seem decisive to explain this global and very stable phenomenon and it remains an open question. To model this frequency shift, we propose a cultural explanation, based on very few biological hypothesis of conformity and sexual competition. We then build a theoretical framework applying technics from dynamical social network analysis and from flocking behaviour. The same type of model has already been used to model langage evolution (Cucker et al. 2003). We then compare the results given by this theoretical tool to the measured shifts. We finally draw the conclusion that

these models are compatible with all the data available, including yearly variations in the frequency (Gavrilov et al. 2012). Finally, we propose a way to test these models in the future.

Mutual influence between language and perception in multi-agent communication games

Xenia Ohmer, Michael Marino, Michael Franke, Peter König

Language interfaces with many other cognitive domains. This work explores how interactions at these interfaces can be studied with deep learning methods, focusing on the relation between language emergence and visual perception. To model the emergence of language, a sender and a receiver agent are trained on a reference game. The agents are implemented as deep neural networks, with dedicated vision and language modules. Motivated by the mutual influence between language and perception in cognition, we apply systematic manipulations to the agents' (i) visual representations, to analyze the effects on emergent communication, and (ii) communication protocols, to analyze the effects on visual representations. Our analyses show that perceptual biases shape semantic categorization and communicative content. Conversely, if the communication protocol partitions object space along certain attributes, agents learn to represent visual information about these attributes more accurately, and the representations of communication partners align. Finally, an evolutionary analysis suggests that visual representations may have evolved in part to facilitate the communication of environmentally relevant distinctions. Aside from accounting for co-adaptation effects between language and perception, our results point out ways to modulate and improve visual representation learning and emergent communication in artificial agents.

Multi-agent collaboration through emergent communication based on MLDA and Metropolis-Hastings

Tomoaki Nakamura, Tadahiro Taniguchi, Akira Taniguchi

In this poster, we propose a probabilistic generative model for multi-agent collaboration based on emergent communication. The leader agent selects an action and sends a low-dimensional message to the follower agent. The follower agent selects an action based on the received message, and both agents conduct their actions and obtain the reward from the environment. To achieve the task, the leader needs to generate appropriate messages, and the follower needs to interpret them appropriately. To do so, relationships between actions, rewards, and messages in each agent are modeled by multimodal latent Dirichlet allocation (MLDA). The leader agent infers the message to likely obtain the high reward based on MLDA and sends it to the follower. The follower agent infers the action to likely obtain the high reward based on the received message based on MLDA. However, messages are not always interpretable for the follower agent; in this case, the agent needs to reject the message. We formulate this exchange of the messages using Metropolis-Hastings and make it possible to learn appropriate messages through reinforcement learning. We conduct a simple guessing game of two agents in the preliminary experiment. Through the experiment, we show that our formulation of multi-agent

collaboration based on emergent communication can achieve higher performance than other formulations. Furthermore, we show the leader agents can generate meaningful messages, and the follower agent can appropriately interpret them.

Learning inflection classes using Adaptive Resonance Theory

Peter Dekker, Bart de Boer

In this study, we investigate how humans process verbal morphology, using Adaptive Resonance Theory as a cognitive model. We specifically study the role of generalization, which happens both between different persons within a verb (e.g. I walk, you walk) and across different verbs (e.g. it walks, it falls). In this last type of generalization, inflection classes play an important role: groups of verbs that are inflected in the same way (e.g. sing, sang sung/ ring, rang, rung). Recently, a range of computational models, specifically sequence-to-sequence neural networks, have been developed to study processing of morphology (Elsner et al., 2019). To model how humans infer inflection classes from raw data, and to evaluate the role of generalization, we propose the task of unsupervised inflection class clustering: given a list of verb forms for different paradigm cells of different lemmas, the task is to cluster these verb forms together into inflection classes. We would like to test if a cognitively inspired computational model is able to perform this clustering task. We apply Adaptive Resonance Theory (ART) (Carpenter & Grossberg, 1987), a neural network architecture in which generalization plays a central role, represented by the vigilance parameter. The network consists of an input layer, where new stimuli come in, and a recognition layer, which represents learned categories. We performed experiments on a dataset of verb forms for Romance languages, augmented with inflection classes for evaluation (Beniamine et al., 2020). Our results point in the direction that an ART model is able to learn a system of inflection classes from word forms. In the future, we would like to evaluate if ART learns inflection classes better when data is supplied in batches more akin to human conversation, with forms of the same verb more likely to be in the same batch. Also, it would be interesting to analyze the ‘critical feature patterns’ inside the ART model, which explain which features in the input data a learned category attends to (Grossberg, 2020).

Seeing the bigger picture: Can deep neural agents learn higher-level concepts in crossmodal referential games?

Radu Alexandru Cosma, Lukas Knobel, Marianne De Heer Kloots, Oskar van der Wal

Referential games are a popular setup for studying the emergence of communication systems that are functionally optimised for task success between (human or artificial) players. In modern computational implementations of such games, deep neural agents respectively play the sender or receiver role in communicating about some symbolic or perceptual data (Lazaridou et al., 2017). However, earlier work with this setup found that agents prefer communicating about low-level features rather than high-level concepts when the data consists of raw pixel input (Bouchacourt & Baroni, 2018), and produce less structured messages compared to settings with symbolic input data (Lazaridou et al., 2018). In this study, we explore the contribution of visual and textual modalities in agents'

higher-level concept formation, using images and accompanying captions from the MS-COCO dataset (Lin et al., 2014). In different conditions, we vary the sender and receiver target modalities (image vs. caption), aiming to investigate whether crossmodal setups may drive agents to focus on mode-independent concepts in communication. This is unlike previous work, which used captions to supervise message generation directly (Havrylov and Titov, 2017; Lee et al., 2018). We analyse the obtained messages with respect to message statistics and employ probing to evaluate to what degree real-world concepts such as object classes are encoded. In preliminary results, we find that probe models indeed more easily recognize the presence of objects in a crossmodal (image-to-caption) setup, while message statistics are mostly preserved. Additionally, we plan to evaluate the word-level relevance of generated messages using feature attribution methods on the object-recognizing probe models and the receiver. We hope that our work will contribute useful analysis methods for finding what features are encoded in emergent communication systems and answer whether a crossmodal setup facilitates the emergence of a concept-based language.

Spontaneous sign emergence in humans and machines through an embodied communication game

Emily Cheng, Yen-Ling Kuo, Boris Katz, Ignacio Cases, Andrei Barbu

We study the emergence of symbolic communication in humans and machines with a communication game that focuses on the emergence of shared signs. In our experimental setup a teacher must communicate a first-order logic task to a student, only through continuous motions in a shared arena. In the human implementation of the experiments, subject pairs communicate via motions of car avatars, a communication channel where familiar conventions are absent. We also perform the parallel experiments for artificial agents in order to study the emergence of signs in a multi-agent reinforcement learning setting. We find that human subjects spontaneously develop a shared vocabulary of motions, and we observe a transition in the dominant sign category being developed from indices to icons to symbols. Moreover, we find that ambiguous game environments drive this transition. We now collect data for the artificial agent experiments, quantitatively compare the dynamics of sign establishment in humans and machines, and observe the extent to which this “symbolic transition” also occurs in artificial agents.

Softly Constrained Agent-Based Models

Marnix Van Soom, Bart de Boer

An agent-based model (ABM) can be thought of as a computer program that implicitly defines a probability distribution $q(x)$ over all of its possible outputs x. The density $q(x)$ cannot be evaluated directly, but is represented by a set of samples $\mathbf{x} = x_n$, where each $x_n \sim q(x)$ is the output of an independent run of the ABM. We are typically after the expectation value $F \equiv \langle f(x) \rangle$ of some interesting statistic $f(x)$, which can be estimated from the samples $\mathbf{x}$ in the usual way. To obtain the object of interest F, therefore, the ABM is run in what we define as the "forward" direction, schematically represented as $\mathbf{x} \rightarrow F$. It is also possible to run the ABM "backwards" $\mathbf{x'} \leftarrow F'$, where the object of interest is now the set of

samples **x**' = { x'_m } . In the backward direction the expectation < $f(x')$ > ≡ F' is *constrained* to take a given value $F' \neq F$ and now we solve for the probability distribution $p(x')$ which satisfies that soft constraint while still as close as possible to the prior $q(x)$. It turns out that the optimal solution to this problem can be approximated by a simple reweighting of the original samples **x**, from which the **x'** can be obtained by standard resampling such that roughly each $x'_m \sim p(x')$. By the same logic used in the first paragraph, the obtained **x'** then represents the probability distribution $p(x')$ of a new computer program automatically derived from the original ABM, which we call a softly constrained agent-based model (SCABM). To show that SCABMs are computationally feasible, we investigate the influence of softly constraining the global clustering coefficient on the convergence of a simple language game played on different network types.

Capturing historical causality via game-theoretic interactions

Alexandra Simonenko

Capturing historical causality via game-theoretic interactions This project in-progress tackles the long-standing problem of causal relations in theoretical diachronic linguistics using a game-theoretic framework. Specifically, the quantitatively well-documented historical correlations between, on the one hand, shifts from pro-drop to obligatory subjects and verbal endings syncretisation, and, on the other, between the disappearance of particular word orders and the emergence of (in)definite determiners have been widely hypothesized to have causal underpinnings. However, attempts to relate these pairs of phenomena in purely grammar-theoretical terms have been empirically inadequate. More generally, it appears that causal connections should be modeled at the level of pragmatic interactions between speakers, rather than within the grammar proper. We build on Simonenko, Crabbé, Prévost (2020) who establish a relation between the rate of prop-drop and the rate of verbal agreement syncretism via a reinforcement learning algorithm, whereby the probability of a syncretic ending in the input data at a given period translates into the probability of the pro-drop grammar parsing failure. While Simonenko et al. (2020) make use of a convergence theorem to calculate the probability of the pro-drop grammar being used based on its (and its competitor's) failure probability, we implement their intuition in a game-theoretic set-up. Our design involves models of a speaker and a listener stochastically aware of the pragmatic consequences of their grammatical choices and learning from the results of their interactions. We also extend the game-theoretic design of Simonenko and Carlier (2020) that relates the use of certain word orders and (in)definite determiners via pragmatic reasoning. The premises for our models (e.g. the exact grammatical alternatives an agent considers) are informed by the attested data from historical French and English. We will also use quantitative historical data from the Penn treebanks of these languages in order to evaluate models' performance.

Learning to make sense out of ambiguous messages leads to language evolution, a simulation

Antonio Norelli, Emanuele Rodolà

The problem of replicating the human process of knowledge discovery in a machine –i.e. the task of automatically discovering the symbolic explanation that enables sensible predictions on a novel phenomenon– has been addressed by Program Synthesis, and more recently by Explanatory Learning. The fundamental difference between these two paradigms resides in the nature of the interpreter that should digest the explanation: PS assumes it to be given, leaving the burden of interpreting symbols to a rigid human-coded compiler, while EL calls for a learned interpreter, trained on a limited collection of existing explanations paired with observations of the corresponding phenomena. A remarkable property of a learned interpreter is its ability in handling ambiguous statements; even if trained only with strings generated through the production rules of a formal grammar, it can correctly interpret also strings close but not valid in that grammar. We demonstrate this feat using Critical Rationalist Networks on Odeen, a basic EL environment that simulates the scientific practice in a small flatland-style universe. Given a new phenomenon, sometimes our agent discovers an explanation that is not valid in the formal grammar we used to generate the world, yet is effective in dictating new correct predictions. Our current research involves an interpreter that continuously learns on a growing training dataset, enriched whenever a new explanation granting correct predictions is found. This causes the interpreter to learn also from explanation strings not valid in the initial formal grammar, generating a ripple effect that drifts the agent away from the starting production rules, while still continuing to improve its prediction performance thanks to the larger dataset. This dynamics makes Odeen a small simulation of a language that evolves over training iterations, and suggests that Explanatory Learning may be a useful computational model to understand some aspects of language evolution.

Emergence of Grounded Signal-Meaning Mappings in Human-Machine collaboration

Tom Kouwenhoven. Tessa Verhoef

Although machine learning of human language patterns has recently improved dramatically, language models still lack a real understanding of how language is related to the real world [1], known as the grounding problem. To enable natural conversations between humans and machines, we need to overcome this problem. A promising solution may be to let natural human-machine language emerge from frequent interactions between humans and machines to slowly build grounded vocabularies that are understood by both, humans, and machines [2, 3]. This process has been studied extensively in the field of Language Evolution and our work concerns a first step in this direction by computationally modeling human behavior in the Embodied Communication Game (ECG) [4]. This collaborative game investigates the emergence of shared signal-meaning mappings when participants have a shared goal but no conventional means of communication (i.e., a new communication system must emerge). While this task is not trivial, most participants are able to solve this by establishing an initial convention (i.e., common ground) and collaboratively bootstrap new signals upon this convention [4, 5]. Our work aims to develop Deep Reinforcement Learning agents that can collaborate with humans and solve this task too. This poses challenges such as dynamic role allocation, behavior policy adaptation, and efficient learning. We first identify architectures that are suitable to learn from individuals' human behavior captured whilst playing the ECG. To do so we trained deep networks to predict subsequent moves given

previous game states. First results show that bidirectional LSTM networks can model participant behavior (mean evaluation accuracy = .994, mean negative log likelihood = .402). However, two models trained on a pair of participants are not able to translate the captured behavior to unseen games, something that would be trivial for humans. To improve this deficit, future work involves investigating more flexible architectures that adapt behavior according to the behavior of their teammate.

References

[1] Igor Mordatch and Pieter Abbeel. 2018. Emergence of grounded compositional language in multi-agent populations. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32.

[2] Maartje ter Hoeve, Evgeny Kharitonov, Dieuwke Hupkes, and Emmanuel Dupoux. 2021. Towards interactive language modeling. arXiv preprint arXiv:2112.11911.

[3] Tom Kouwenhoven, Tessa Verhoef, Roy de Kleijn, and Stephan Raaijmakers. 2022. Emerging grounded shared vocabularies between human and machine, inspired by human language evolution. Frontiers in Artificial Intelligence, 5

[4] Thomas C Scott-Phillips, Simon Kirby, and Graham RS Ritchie. 2009. Signalling signalhood and the emergence of communication. Cognition, 113(2):226–233.

[5] Tom Kouwenhoven, Roy de Kleijn, Stephan Raaijmakers, and Tessa Verhoef. 2022b. Need for structure and the emergence of communication. In 44th Annual Meeting of the Cognitive Science Society (CogSci 2022). Cognitive Science Society