Chapter 6

Conclusions

As a complex adaptive dynamical system, human language is constantly evolving, with the individual behaviors of language users driving linguistic emergence and change at the population level. Inspired by the interactive and dynamic nature of human language, the development of AI has increasingly focused on simulating the emergence of human-like languages with neural network agents (Mikolov et al., 2018; Galke and Raviv, 2025; Rita et al., 2024). Early frameworks have been progressively expanded to display important aspects of human language and communication. Within this body of work, most studies initialize their agents on *sets of random symbols*, which makes it intrinsically difficult to analyze the emergent agent protocols and to compare them to human preferences at the level of specific language properties.

This thesis extended this line of work by introducing the NeLLCom framework, designed to study the emergence of specific language universals. Specifically, our agents start by learning a pre-defined artificial language, inspired by experimental research in artificial language learning (ALL) with human participants. The interactive nature of language systems is modeled by letting agents participate in meaning reconstruction games while optimizing a shared communication success reward. The use of pre-defined artificial languages makes the communication learning process interpretable and directly comparable to human experimental results. Using the word-order/case-marking trade-off and differential case marking as our use cases, we examined how language productions of neural agents evolve during learning and repeated communication. Specifically, we answered four progressive research questions:

RQ-A Can the introduction of more realistic simulation factors lead to the emergence of a word-order/case-marking trade-off in neural-agent iterated language learning?

Natural languages commonly display a trade-off, using either word order or case marking to convey constituent roles. A similar trade-off, however, had not been observed in previous simulations of iterated language learning with neural network based agents (Chaabouni et al., 2019b). In **Chapter 2**, we re-evaluated Chaabouni et al. (2019b)'s findings in light of three factors known to play an important role in comparable experiments and simulations from the language evolution field, namely: (i) the speaker bias towards efficient messaging (**RQ-A.1**), (ii) the variable and unpredictable nature of input languages (**RQ-A.2**), and (iii) the learning bottleneck (**RQ-A.3**).

Our simulations showed, under different conditions, that neural agents tend to maintain the distribution of utterance types observed during learning instead of displaying behaviours we see in similar experiments with humans, like introducing structure or making the language more systematic. Specifically, introducing the least-effort bias (§2.4) and exposing the agents to highly unpredictable input languages (§2.5.2) resulted in the collapse of the communication system, whereas moderate input language variability (§2.5.1) and the presence of a learning bottleneck (§2.6) led to a stable maintenance of variable strategies, matching the input distribution, instead of a gradual regularization of marker usage or word order. This aligns with prior findings by Chaabouni et al. (2019b) whereby redundant coding strategies persist in the neural-agent iterated learning framework. Only combining least-effort bias with moderate language variability (§2.5.1) led to a temporary optimization of the language, but that was again followed by communication failure due to the continued influence of the hard-coded least-effort bias, causing utterances to become dramatically shorter over time.

In summary, we found that the existing neural-agent iterated learning framework is inappropriate to simulate the emergence human-like language universals. Simply hard-coding cognitive biases is insufficient to yield human-like results in this framework. In natural language use, the pressure to communicate efficiently must be balanced against the need to maintain a stable and expressive communication system —a key insight that motivates our next research question.

RQ-B Does a human-like word-order/case-marking trade-off emerge in communicative neural agents?

As reviewed by Chaabouni et al. (2019a); Galke et al. (2022); Rita et al. (2022), and confirmed in **Chapter 2**, artificial learners often behave differently from human learners in the context of neural agent-based simulations of language emergence and change. We proposed that more naturalistic settings of language learning and use could lead to more human-like results. Specifically, we studied the effect of combining the standard supervised learning objective with a measure of communicative success. To this end, we introduced a new Neural-agent Language Learning and Communication framework (NeLLCom), where pairs of speaking and listening agents learn a given artificial language through supervised learning, and then use it to communicate with each other, optimizing a shared reward via reinforcement learning.

We used NeLLCom to replicate the experiments of Fedzechkina et al. (2017), where two groups of human participants were asked to learn a fixed- and a flexible-order artificial language, respectively, and tested after training. Our results confirmed previous findings in neural agent studies and showed that SL is sufficient for perfectly learning the languages, but does not lead to any human-like regularization. By contrast, communication learning leads agents to modify their production in a human-like way (**RQ-B.1**): Firstly, optional markers are dropped more frequently in the redundant fixed-order language than in the ambiguous flexible-order language. Moreover, in the flexible-order language one of the two word orders becomes clearly dominant and an asymmetric case marking strategy arises. Agent productions after communication showed a clear correlation between effort and uncertainty, which strongly matches the core finding of Fedzechkina et al. (2017) (**RQ-B.2**). Besides the similarity, some interesting differences compared to human results were also observed. For instance, NeLLCom agents showed a slightly stronger tendency to reduce effort rather than uncertainty.

In summary, we found that the word-order/case-marking trade-off, as a specific realization of the efficiency/informativity trade-off, can indeed emerge in neural network learners when these are equipped with a need to be understood.

RQ-C What are the necessary ingredients to scale up NeLLCom to larger populations?

The previous **Chapter 3** introduced the NeLLCom framework, allowing agents to first learn an artificial language and then use it to communicate. However, the way agents were modeled to fulfill separate, complementary roles (i.e. one agent always speaks, the other always listens) restricted the scenarios where NeLLCom could be applied. To scale this up, in the following **Chapter 4**, we extended the vanilla NeLLCom agent to act as both listener and speaker (i.e. role alternation) using parameter sharing and a self-play procedure. This enabled us to simulate group communication via a turn scheduling algorithm.

Within this extended framework, NeLLCom-X, we experimented with a novel setup where pairs of agents interact after having been exposed to different initial languages. We showed that agents with different languages realistically adapt their utterances to each other to increase communicative success (**RQ-C.1**). Focusing on the effect of group size (**RQ-C.2**), we successfully extended our key findings from **Chapter 3** and demonstrated that a word-order/case-marking trade-off emerges not only in individual agents but also at the group level. Additionally, languages used by agents in larger groups become more optimized and less redundant, which is in line with previous hypotheses on the effect of population size on language structure (Lupyan and Dale, 2010; Raviv et al., 2019). Importantly, an experiment at this scale could not easily be done with human participants in the lab.

In summary, we successfully extended the original NeLLCom to support more realistic groups of role-alternating agents, and showed that group size has an important role on the emergence of our studied language universal.

RQ-D Can the NeLLCom-X framework be used to simulate the emergence of another case marking universal?

Chapter 3 and **Chapter 4** demonstrated the success of NeLLCom(-X) in replicating the emergence of the word-order/case-marking trade-off. In this research question, we use another case study to further evaluate our newly developed framework, namely differential case marking (DCM). DCM refers to a natural language phenomenon where marker use is influenced not only by word order but also by semantic and pragmatic properties of arguments. Once again, our experimental setup and language design draw direct inspiration from human experiments previously conducted by Fedzechkina et al. (2012) and Smith and Culbertson (2020). Specifically, focusing on an object-marking condition, Smith and Culbertson (2020) found that human participants exhibited a DCM effect (i.e., using markers more often for animate than inanimate objects) after communicating with a chatbot.

In our neural-agent simulations, we did see a human-like DCM effect appear during agent interactions. However, agents were more sensitive to specific patterns in the input language than humans, and had a greater tendency to drop markers and disambiguate meanings using word order (**RQ-D.1**). Agents were also more sensitive to, and often tended to amplify, the initial language biases. Thus, the original artificial language designed by Fedzechkina et al. (2012), with its uneven word order distribution (60%SOV-40%OSV) and case marking conditioned on word order (67% on SOV, 50% on OSV), likely influenced the agents' production regularization. To control for language input bias, we further experimented with a neutral-order language where SOV and OSV are evenly distributed with a unified case marking proportion (67%). In this setting, we observed a more pronounced differential case marking phenomena (**RQ-D.2**).

Taken together, these results support Smith and Culbertson (2020)'s findings highlighting the critical role of communication in shaping DCM and showcase the potential of neural-agent models to complement experimental research on language evolution. We take this as an encouraging indication that NeLLCom can be used to study different language phenomena that have already been explored with human artificial language learning experiments.

We publicly released our framework[1] to foster future research on the emergence of different language universals in communicative neural agents.

Limitations and Future work

This thesis represents an important step towards developing a neural-agent framework that replicates patterns of human language change without the need to hard-code ad-hoc biases. We are also aware of several limitations, which we discuss here along with possible solutions as future work.

First, the current artificial languages are overly simplistic, with a small language scale, low structural complexity, and a meaning space that is strongly abstracted from reality. For the next steps, more complex languages with larger vocabularies, more realistic (e.g. Zipfian) lexical distributions, as well as less constrained meaning spaces (e.g. pixel-level image input) could enhance the generality of our current findings.

Second, all experiments with NeLLCom(-X) in this thesis used a layer of Gated Recurrent Units to model input and output sequences. However, different neural network architectures exhibit distinct inductive biases (Kuribayashi et al., 2024) and generalization abilities (Shiri et al., 2024; Fukushima and Tani, 2023). Replicating our experiments with other architectures would therefore be an important step to assess the possible impact of architecture-specific structural biases.

Third, the language transmission dynamics explored so far within NeLLCom(-X) can be expanded. In the horizontal dimension, we have only considered

[1] https://github.com/Yuchen-Lian/NeLLCom-X

a fully-connected group scenario, while a more realistic simulation could include different community structures and connectivities, with agents' identity awareness. Regarding communication between speakers of different languages, future work could also expand from pair-wise to group-wise, to investigate the emergence of dialect during language contact between isolated communities (Harding Graesser et al., 2019). In the vertical dimension, transmission over generations may amplify the small inductive biases of individual agents (Kirby et al., 2015; Thompson et al., 2016). Therefore, augmenting NeLLCom(-X) with iterated learning presents another promising research direction.

Lastly, while the experiments in this thesis focused on the interplay between word order and case marking as their use case, our proposed framework simulates general language learning and communication processes, and can be adapted to study many other language phenomena. Since the writing of this thesis, NeLLCom has been successfully adapted by Zhang et al. (2024) to study dependency length minimization, i.e. the widely observed tendency of natural languages to reduce the overall linear distance between syntactically related words. Other linguistic aspects previously explored in Artificial Language Learning experiments with human participants —such as colexification and the role of iconicity or metaphor in the emergence of new meanings (Verhoef et al., 2015, 2016; Tamariz et al., 2018; Karjus et al., 2021; Verhoef et al., 2022), or the combinatorial organisation of basic building blocks (Roberts and Galantucci, 2012; Verhoef, 2012; Verhoef et al., 2014)— could also be suitable candidates for investigation within NeLLCom.

Concluding remarks

In this thesis, we introduced a novel neural-agent language learning and communication framework combining language learning and transmission processes, both of which have been proven to play an important role in the evolution of human language. We see NeLLCom as a useful approach to complement experimental research on language evolution, allowing us to precisely control and compare various aspects of language systems and population dynamics while at the same time revealing ways in which neural-agent language learning and use differ from those of humans. We hope our work will facilitate future simulations of language evolution with the end goal of explaining why human languages look the way they do.