Emergence of linguistic universals in neural agents via artificial language learning and communication

Lian, Y.

Chapter 1

Introduction

Human language is not a static entity but a dynamic system undergoing continuous change and evolution. Its linguistic structure is shaped by mechanisms operating across different time-scales (Elman, 1995; Steels, 2000; Beckner et al., 2009). On shorter time scales, interaction and communication facilitate the negotiation of new meanings, while on longer time scales, processes such as learning and transmission across generations give rise to emergent linguistic patterns and enhance learnability (Smith, 2022).

While the languages of the world exhibit vast diversity, they also reveal universal common patterns (Greenberg, 1963). For instance, words with high frequency are commonly short compared to low-frequency words with longer word lengths. This statistical reverse relationship between word length and usage frequency is generalized as Zipf's law of abbreviation (Zipf, 1949). For example, the most common words in English, such as *the*, *a*, and *of*, are typically very short, while rare, highly specific words like *hemidemisemiquaver* (i.e. a sixty-fourth music note) or *prestidigitation* (i.e. the art of performing magic tricks) are long and rare. Beyond this lexical pattern, universal tendencies also manifest in syntactic and morphological structures.

It has been suggested that these shared linguistic features can be understood as adaptations to the contexts in which language is used and transmitted (Christiansen and Chater, 2008; Kirby et al., 2015; Smith, 2022).

1.1 Studying the Emergence of Language Universals

To analyze these universal patterns, **typological studies** analyze language data gathered from diverse time periods and geographical locations. Although this approach has led to the discovery of numerous linguistic universals (Greenberg, 1963; Croft, 2003), it does not reveal the underlying mechanisms involved in the emergence of these patterns (Fedzechkina et al., 2016).

Experiments with human participants can address this shortcoming by testing in the lab the precise mechanisms that may contribute to the emergence of commonly observed linguistic features (Fedzechkina et al., 2016; Smith, 2022). Such experiments allow researchers to observe the emergence of novel communication systems through social coordination when participants play language games, or through cultural transmission via an **artificial language learning (ALL)** paradigm. During these games, participants engage in controlled language learning tasks that model various transmission dynamics like cultural transmission over generations, or communicative interaction between individuals. The language design in these experiments is typically guided by specific hypotheses about certain linguistic features and how they may arise in language as an adaptation to, for example, communicative needs or learning constraints.These methods allow for a degree of experimental control and may reveal causal relationships between certain mechanisms and the emergence of common patterns.

Kirby et al. (2008), for example, simulated the emergence of compositionality through cultural transmission, where initially unstructured artificial languages were repeatedly learned and transmitted to the next participant, resulting in more regular and learnable languages. Words that could not be remembered easily by the participants were harder to reproduce during transmission to the next generation, and therefore had a higher chance to undergo changes. Words that survived after transmission over multiple generations tended to be more compositional, as being more structured implied better learnability. Besides such 'vertical' transmission, experimental ALL approaches have also been used to simulate 'horizontal' transmission in which signals are transmitted through

communication within a group (Raviv et al., 2019; Smith, 2024). Raviv et al. (2019), for example, investigated the effects of community size on language structure, and found that languages developed in larger groups are more systematic than those developed in smaller groups.

Beyond compositionality, a wide range of linguistic aspects has been explored, including syntactic patterns like word order or morphology (Christensen et al., 2016; Saldana et al., 2021b; Motamedi et al., 2022), a tendency to reduce dependency lengths (Fedzechkina et al., 2018; Saldana et al., 2021a), colexification patterns 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), and combinatorial organisation of basic building blocks (Roberts and Galantucci, 2012; Verhoef, 2012; Verhoef et al., 2014). Thus, this well-established experimental approach has proven to be both convincing and reliable, effectively filling the gap in providing direct causal inferences for the emergence of language universals (see Fedzechkina et al. (2016) and Culbertson (2023) for a detailed survey of this experimental ALL paradigm).

However, this approach has several shortcomings. First, the selection of participants and their prior language knowledge is likely to influence the findings obtained from these experiments (Culbertson, 2023). As most studies predominantly recruit native English speakers, results may generalize poorly to participants of other language backgrounds. Second, scaling up these experiments is particularly challenging. Due to the nature of the lab training procedure, the languages learned by participants are often quite small and do not match the complexity of real human languages. In addition, time and budget constraints limit the scope of ALL experiments to short-term learning, few rounds of interactive communication, and small numbers of participants compared to real linguistic communities.

The use of **agent-based modeling** techniques has been proposed as a suitable solution that enables direct causal inference and facilitates scalability. As a productive approach to studying the emergence and evolution of linguistic systems, agent-based modeling has a long-standing tradition in language

evolution research (Hurford, 1989; Hare and Elman, 1995; Steels, 1997). In this context, agents are typically modeled as individual language users, with their capabilities, linguistic knowledge, and interaction behaviors designed and updated according to the specific research objectives. These agents typically maintain an evolving lexicon and shared knowledge about the environment, updating their understanding based on predefined interaction rules (see De Boer (2006) for a survey of early models on vertical and horizontal transmission of simple languages). Within this line of research, iterated learning, a well-known paradigm introduced by Kirby (2001), is used to simulate language evolution over generations on a longer timescale. In this paradigm, a child agent learns from its parent agent in the same way the parent learned from its predecessor. Similarly to its counterpart with human learners in the lab, this work demonstrated that compositionality can also emerge from initially unstructured languages in populations of agents through the process of language learning and transmission.

These traditional agent-based methods allow direct verification of the underlying language feature emergence hypothesis, by modeling the language change dynamics from reality to a high-level abstraction. Research with these models was highly productive in the 1990's and early 2000's, and led to numerous important insights for the field of language evolution. However, these methods were limited by computational resources available at that time and were often criticized for lacking realism (Cangelosi and Parisi, 2002).

Recent advances in deep learning have significantly enhanced the capabilities of these agent-based methods to deal with more complex, larger-scale communication tasks. Moreover, the impressive linguistic abilities displayed by deep-learning language models trained on the full complexity of natural language corpora have led to a renewed interest in agent-based simulations of emergent communication and language evolution.

1.2 Neural Network-Based Emergent Communication

The rapid advancements in deep learning have driven remarkable success in modern large-scale natural language processing (NLP). These achievements underscore the strong learning and generalization capabilities of neural networks. This has triggered a renewed interest in adopting neural network-based agents to simulate the emergence of novel linguistic protocols from scratch (Havrylov and Titov, 2017; Kottur et al., 2017; Lazaridou et al., 2017; Lazaridou and Baroni, 2020).

On the one hand, this line of work draws inspiration from the interactive nature of human language to enhance AI systems (Mikolov et al., 2018). In this context, emergent communication is used to facilitate interactions within an environment of agents using more flexible, non hand-crafted task-solving protocols (Foerster et al., 2016; Taniguchi et al., 2024). Additionally, the use of multi-agent communication techniques has been explored to improve the ability of deep learning chatbots pre-trained on human language corpora to interact with human users (Lazaridou et al., 2017; Das et al., 2017; Lazaridou et al., 2020).

On the other hand, neural-agent emergent communication paradigms contribute to advance the exploration of the underlying mechanisms of human language evolution, which are the focus of this thesis. Within this line of research, pairs of agents are often simulated to play language games, where a speaking agent attempts to help a listening agent recover an intended meaning by generating a message that the listener can interpret (Lazaridou et al., 2017). These agents typically invent and negotiate their languages **from scratch**, that is, starting from a set of random symbols with no pre-defined meaning or structure. Studies in this domain typically investigate the relationship between the emergence of human-like language properties in these successfully communicated agents' languages, and the simulated factors hypothesized to shape human languages.

In this later context, compositionality has been by far the most widely studied language feature (Li and Bowling, 2019; Kottur et al., 2017; Ren et al., 2020; Mordatch and Abbeel, 2018; Choi et al., 2018; Lazaridou et al., 2017; Resnick et al., 2020). Specifically, the previously established influence of iterated learning on the emergence of compositionality has been replicated with these modern agent setups (Ren et al., 2020; Li and Bowling, 2019; Cogswell et al., 2019; Zheng et al., 2024). Besides the process of language transmission over generations, constraints such as model capacity (Resnick et al., 2020) and memory bottlenecks (Kottur et al., 2017) have also been shown to be key factors in inducing compositionality in neural agent emergent communication.

Additionally, numerous studies explore other influencing factors, such as community level dynamics (Harding Graesser et al., 2019; Tieleman et al., 2019; Kim and Oh, 2021; Michel et al., 2023) and multi-modal perception (Lazaridou et al., 2018; Choi et al., 2018). Some research has also examined various syntactic and pragmatic linguistic features, including word-order preferences (Chaabouni et al., 2019b; Kuribayashi et al., 2024) and communication efficiency (Chaabouni et al., 2019a; Lowe et al., 2019; Kharitonov et al., 2020). For a comprehensive overview of this domain, we refer to the survey of Lazaridou and Baroni (2020) and the more recent survey of Boldt and Mortensen (2024).

Since the agents fully invent their own language from scratch in the typical emergent communication setup, there is a key challenge in this line of research: analyzing the emergent protocols developed by agents is inherently difficult. The languages they invent are only comprehensible to the agents involved in the game. Therefore, the majority of current evaluations for the agents' productions still focus on very general high-level features like compositionality or generalizability (Lazaridou et al., 2018; Chaabouni et al., 2020), with metrics like topographic similarity (Brighton and Kirby, 2006) being often used. Thus, a major obstacle lies in the need to decrypt the protocols and manually ground them into understandable natural language or identifiable linguistic features — a process constrained by the absence of a standardized methodology, making systematic comparisons challenging (Boldt and Mortensen, 2022, 2024). A fur-

ther limitation lies in the fact that the typical emergent communication setup requires agents to negotiate a set of symbols to refer to a set of world entities, akin to the emergence of a vocabulary. This makes the 'from-scratch' approach unsuitable to study the emergence of more structured language properties, such as in the realm of syntax.

To address this challenge and increase the applicability of neural-agent communication techniques to study the origins of more language universals, this thesis introduces a novel framework where **neural network-based agents learn to communicate using pre-defined artificial languages**, directly inspired by ALL experiments with human participants.

1.3 Neural-Agent Language Learning and Communication Framework (NeLLCom)

As the main contribution of this thesis, we develop a novel neural-agent framework to study the emergence of language universals. In NeLLCom, agents start from learning a pre-defined artificial language before interacting with each other. This ALL paradigm is inspired by experimental research in human language learning, where the design of the artificial languages focuses on specific linguistic properties of interest. For example, an artificial language may be designed to convey simple actions involving two entities, with variants of this language displaying different word orders. We then model the interactive nature of language systems by letting those agents participate in meaning reconstruction games. During the game, a listener is asked to reconstruct the input meaning according to the speaker's utterance which is the description of that input meaning. As both listener and speaker are pre-trained, the communication protocol does not need to be built up from scratch. Instead, symbols of the utterances already have a pre-defined mapping towards the world entity they represent (in other words, the vocabulary has already been established), but crucially an aspect of the language syntax (e.g. its word order) may be in a suboptimal state of unpredictable ambiguity.

Consequently, our work examines **how the structure of agent productions**

evolve during interaction under different influencing factors, and starting from slightly different initial languages. The use of pre-defined artificial languages distinguishes our approach from models that initiate communication from scratch or rely on pre-trained models with large-scale natural language corpora. By closely simulating human experimental setups, the productions of agents can be directly compared to those of the human participants, effectively addressing the shortcomings of the prevailing emergent communication paradigm. Simulating language learning and use under a unified framework aligns with modern approaches to the study of language evolution, which center on the strong interplay between processes of language acquisition and communicative need in shaping human languages Christiansen and Chater (2008); Kirby et al. (2015); Smith (2022); Verhoef et al. (2022)

In this thesis, we first introduce a basic version of NeLLCom, where agents have complementary roles, i.e., one always acts as speaker and the other always as listener (**Chapter 3**). To make the framework more scalable and better resemble real human interaction, we then modify the agent design to support role alternation, resulting in full-fledged agents which can both speak and listen (**Chapter 4**). The resulting NeLLCom-X framework thus extends its simulation scope to include group communication and interactions among different language learners.

By exploring various settings and language phenomena, this thesis will demonstrate that NeLLCom agents replicate human-like linguistic patterns when subject to a communicative pressure. Additional experiments will show that our simulations can be scaled up to larger-scale agent populations, and that the framework is adaptable to study different language phenomena, making NeLLCom a useful tool for language evolution researchers interested in scaling up their ALL experiments and in refining their hypotheses before carrying out costly human experiments.

1.4 Word Order and Case Marking

While NeLLCom simulates general language learning and communication procedures and can be adapted to study many other language phenomena, this thesis focuses on the interplay between word order and case marking as a use case. Specifically, we investigate the origins of two widely attested language phenomena: (i) the trade-off between word order flexibility and case marking, and (ii) differential case marking.

Word order, as one of the essential syntactic features of languages, has long been studied through linguistic typology and experimental research. Cross-linguistic typological studies (Dryer, 2005) show that the large majority of world languages have either Subject-Object-Verb (SOV) or Subject-Verb-Object (SVO) as the dominant constituent order, yet most languages permit some degree of word order variation. For example, both Russian and English use SVO as their basic word order. However, Russian allows a much greater flexibility in word order than English, accompanied by a richer morphological system (Gell-Mann and Ruhlen, 2011). More specifically, case marking refers to the use of morphological markers, such as suffixes, to indicate the grammatical function of pronouns, nouns and their modifiers within a sentence.

While both word order and case marking can be key features of a language, each describing different aspects of its typological properties, they are largely redundant strategies for conveying the syntactic roles of sentence constituents, leading to a well-known **trade-off** (Sinnemäki, 2008; Futrell et al., 2015): flexible order typically correlates with the presence of case marking (e.g. in languages like Russian or Japanese), while fixed order is often observed in languages with little or no case marking (e.g. English or Chinese). This is illustrated by Example 1, where the order of two noun phrases in a Japanese sentence is switched without affecting the meaning ('Mary reads the book'), as the case marker 'を' indicates '本'(book) is the object and case marker 'が' indicates 'マリー'(Mary) is the subject. However, in English (Example 2) and Chinese (Example 3), the subject 'Mary' can only be placed before the verb, while the object ('book') can only be placed after it as the fixed Subject-Verb-Object

order is the only strategy to assign semantic roles in these languages.

(1) a. マリーが その 本を 読んだ.
Mary the book read.
'Mary reads the book.'[✓]

b. その 本を マリーが 読んだ.
the book Mary read.
'Mary reads the book.'[✓]

(2) a. *Mary reads the book.* [✓]

(3) a. 玛丽 读了 这本 书
Mary read the book.
'Mary reads the book.' [✓]

An important aspect of variation among case marking languages concerns the extent to which case markers can be omitted depending on semantic and pragmatic features of the arguments, a phenomenon known as **differential case marking** (De Hoop and Malchukov, 2008; Sinnemäki, 2014; Witzlack-Makarevich and Seržant, 2018; Levshina, 2021). Example 4 below (adapted from García (2018); Levshina (2021)) illustrates differential case marking in Spanish. In this language, subjects (*'Pepe'*) are not marked while objects are only marked if referring to a human (or animate) entity (*'actriz'*) but not if referring to an inanimate entity (*'película'*).

(4) a. *Pepe ve la película.*
Pepe sees the film.
'Pepe sees the film.'

b. *Pepe ve a la actriz.*
Pepe sees TO the actress.
'Pepe sees the actress.'

Beyond typological distributions, these two common phenomena have also been extensively investigated through experiments based on artificial language learning (ALL) paradigms.

Among these, Fedzechkina et al. (2017) focus on the trade-off between word order and case marking. In their study, two groups of participants learned artificial languages with optional markers but different word orders (fixed vs. flexible). After training, learners of the fixed-order language reduced the case marking proportion, whereas learners of the flexible-order language used case marking more frequently and asymmetrically, favoring its use with less common word orders, indicating the successful replication of this trade-off.

Smith and Culbertson (2020) ran a large-scale experiment to study the emergence of the differential case marking (DCM) phenomenon and focused on the influence of learning and communication pressures on language universals. Building on prior work by Fedzechkina et al. (2012), they conducted experiments where participants learned an artificial language with animate vs. inanimate entities. An interaction phase was introduced after the final learning session, where participants communicated with a chatbot. The results of the experiment showed that DCM did not emerge during learning, but only during the subsequent interaction phase, suggesting that learning alone is insufficient to explain the emergence of differential object marking; rather, communicative interaction plays a crucial role in shaping an efficient case-marking system.

In addition to these human experimental studies, a few studies have investigated word order and case marking universals through agent-based modeling approaches. Following a classical agent-based modeling approach, Lestrade (2018) simulated the emergence of DCM by designing a computational model in which relatively simple agents communicate with each other using words from an initial lexicon, modeled as a list of randomly generated vectors. Marking strategies, heuristics for interpreting messages and grammaticalization principles were explicitly built-in to examine their combined or separate impact on the emergence of DCM. Their simulation shows that argument marking can evolve gradually as languages adapt to usage.

In a shift towards deep learning approaches, Chaabouni et al. (2019b) investigated whether recurrent neural network-based agents have particular word order biases, and whether these resemble the tendencies observed in natural

languages. They implemented an iterated learning process (Kirby et al., 2014) using neural agents trained on hand-designed artificial languages, and examined their productions over generations. The results were mixed, showing a human-like tendency to avoid long-distance dependencies but no clear trend towards trading off between word order and case marking to avoid redundancy.

Our first study presented in **Chapter 2** re-evaluates the findings of Chaabouni et al. (2019b), in light of several key factors known to play important roles in comparable experiments and simulations within the field of language evolution. Focusing on the same word-order/case-marking trade-off, **Chapter 3** introduces a novel artificial language training paradigm for neural agents (NeLLCom) that combines supervised and reinforcement learning, and successfully replicates the trade-off in a pairwise communication setup. **Chapter 4** further improves the agent architecture and investigates whether a similar trade-off also emerges at the group level within the extended framework, NeLLCom-X. Finally, **Chapter 5** validates the applicability of our framework to simulate a related but different phenomenon, namely differential case marking.

1.5 Thesis Overview and Research Questions

In this dissertation, we set out to answer the following 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?

Specifically, in **Chapter 2**, we re-assess the findings of an existing supervised neural-agent iterated language learning framework (Chaabouni et al., 2019b), which failed to replicate the emergence of a word-order/case-marking trade-off. We investigate the role of specific factors known to affect human language evolution through the three following sub-questions:

RQ-A.1 *How does a least-effort bias affect the emergence of the word-order/case-marking trade-off?*

An efficiency-based account is widely accepted as a key factor in shaping natural languages (i Cancho and Solé, 2003; Kanwal et al., 2017; Fedzechkina et al., 2017). However, neural network learners are known to be different from humans in terms of biases. In this question, we investigate whether hard-coding an utterance-length penalty into the agents as an explicit pressure to minimize effort can lead to a human-like word-order/case-marking trade-off in agent simulations.

RQ-A.2 *How can input variability impact the emergence of the word-order/case-marking trade-off?*

We notice another possible discrepancy between human language evolution processes and agent language learning in the previous simulations. In artificial language learning experiments involving human participants, unpredictable variation is a common and crucial feature of the designed languages (for example, case marking is optional in (Fedzechkina et al., 2017) or ambiguous with two plural marker forms in (Smith and Wonnacott, 2010)). By contrast, in the languages of Chaabouni et al. (2019b) both word order and case marking systems are fully systematic, leaving little space for a neural network to make changes or optimize this system. We introduce two levels of variability into the languages and evaluate agent production preferences in response to these unpredictable variations. We also test the combined effect of input variability and least-effort bias.

RQ-A.3 *How does a learning bottleneck influence the emergence of the word-order/case-marking trade-off?*

The learning bottleneck has been proposed as a key pressure driving language regularization (Smith et al., 2003; Brighton et al., 2005; Kirby et al., 2014). In the original iterated learning framework (Kirby, 2001; Kirby et al., 2008), this pressure is realized by transferring only partial utterances or incomplete sets of signals from a generation to the next one. However, in the neural agent training setup of Chaabouni et al. (2019b), the large majority of the meaning space (80%) was used to train the next generation. We study the role of the learning bottleneck by gradually reducing the proportion of meanings provided

to the agents during training.

We find that all three tested factors have visible effects on the agent productions. However, no factor or combination of factors lead the agents to optimize their language for efficiency without quickly incurring in a collapse of the communication system, suggesting the existing framework is not suitable to replicate the emergence of a human-like trade-off.

RQ-A is based on the following published research article:

> 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 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10121–10129. Association for Computational Linguistics

Building on the previous chapter's findings, we set out to design a new neural-agent framework combining the classical supervised learning objective with a communicative success objective. This leads to the next research question, addressed in chapter **Chapter 3**:

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

It has been proposed that more natural settings of agent language learning and use might also lead to more human-like patterns (Mordatch and Abbeel, 2018; Lazaridou and Baroni, 2020; Kouwenhoven et al., 2022; Galke et al., 2022). In line with this proposal, **Chapter 3** introduces a novel Neural-agent Language Learning and Communication framework (NeLLCom), which combines the standard supervised learning objective with a communication learning phase based on a meaning reconstruction game. To enable a direct comparison with human production preferences, we adopt artificial languages that were designed with inherent variability and used in previous human experiments on the trade-off by Fedzechkina et al. (2017). We then investigate the following

sub-questions:

RQ-B.1 *Does introducing communicative success lead to the regularization of word order and case marking?*

We first apply supervised learning to teach agents the meaning-to-utterance mapping defined by a given artificial grammar. To introduce a communication pressure, we further set up a meaning reconstruction game, where a speaking agent tries to convey a given meaning to a listening agent via an utterance. Both agents are rewarded based on task success, optimized through reinforcement learning. By analyzing how agent productions change over the course of communication learning, we uncover the agents' intrinsic preferences towards different strategies to convey argument roles.

RQ-B.2 *To what extent does the trade-off observed in the productions of individual communicative agents resemble that observed in human participants?*

Because our artificial languages are borrowed from Fedzechkina et al. (2017), we can compare agent productions directly to the productions of their human participants. Specifically, we look at word order and marker use preferences, both at the level of speaker-listener pair and at the level of a population of agent pairs.

As demonstrated by Fedzechkina et al. (2017), the specific strategy employed by each human participant reveals considerable variation at the individual level. A similar variation is found in the agents' production. At the population level, we find an inverse relationship between uncertainty and utterance length, which aligns with human results and confirms the key role of communicative pressure in replicating language universals.

RQ-B is based on the following published research article:

> Yuchen Lian, Arianna Bisazza, and Tessa Verhoef. 2023. Communication drives the emergence of language universals in neural agents: Evidence from the word-order/case-marking trade-off. *Transactions of the Association for Computational Linguistics*, 11:1033–1047

While the basic NeLLCom framework succeeds at replicating the emergence of a word-order/case-marking trade-off, the agents are still relatively simple, and only able to either speak or listen. By contrast, human language users are obviously able to act both as a speakers and listeners. In human ALL experiments, participants also usually take turns being the speaker and listener (Roberts and Galantucci, 2012; Verhoef et al., 2015; Kirby et al., 2015; Namboodiripad et al., 2016; Verhoef et al., 2022). Additionally, interaction in NeLLCom can only be simulated between pairs of agents, whereas human languages emerge in much larger populations. By means of both typological studies and human experiments, population size has been found to correlate with salient language properties, such as morphological complexity (Raviv et al., 2019) and systematicity (Lupyan and Dale, 2010). The importance of studying interaction in *larger groups of role-alternating agents* motivates our next research question:

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

We address this question in **Chapter 4** by extending NeLLCom in two ways: First, role alternation is achieved by parameter sharing between the speaker and listener networks and by introducing a self-play procedure during communication. Then, the resulting 'full-fledged' agents are made interact in larger groups using a turn scheduling algorithm. The extended framework, NeLLCom-X, enables us to investigate two new research questions:

RQ-C.1 *Do the new full-fledged agents adapt to each other when they start interacting after being trained on different languages?*

Role alternation in NeLLCom-X makes it possible to investigate communication between speakers of different languages, i.e. agents that have been initially trained on different languages. We consider a number of pairwise communication scenarios where one agent is always trained on a neutral language, while the other starts from languages with different word-order and case-marking properties. We expect the agent pairs to negotiate a mutually understandable language, and the neutral language to drift in different directions according to

the interlocutor's language.

RQ-C.2 *How does group size affect the emergence of the word-order/case-marking trade-off?*

Natural languages typically have more than two speakers, and the community size is proposed as a factor that can shape the language structure. Typological data indicate that languages in larger communities tend to be simpler than those in smaller, isolated groups (Wray and Grace, 2007; Lupyan and Dale, 2010), a pattern also confirmed by human experiments (Raviv et al., 2019). In this research question, we investigate whether the same can be seen in populations of neural agents and whether the word-order/case-marking trade-off also emerges at the group level.

In the scenario of two agents initially trained on different languages, we show that agent pairs succeed in negotiating a mutually understandable language whose properties largely depend on the language with stronger initial biases. In larger group communication scenarios, where all agents are initially trained on the same language, we see a larger entropy reduction in the languages used by larger groups as compared to the languages used by pairs of agents. This result aligns with experimental findings by Raviv et al. (2019), who found that larger groups of participants use more systematic languages.

RQ-C is based on the following published research article:

> Yuchen Lian, Tessa Verhoef, and Arianna Bisazza. 2024. NeLLCom-X: A comprehensive neural-agent framework to simulate language learning and group communication. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, pages 243–258. Association for Computational Linguistics

So far, we demonstrated the success of NeLLCom and NeLLCom-X in replicating the emergence of one particular language universal, using the word-order/case-marking trade-off as a case study. In the last research question we explore the possibility of applying our framework to study a related but

different linguistic phenomenon.

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

In natural languages, marker use is influenced not only by word order but also by the semantic and pragmatic properties of the arguments, a phenomenon known as differential case marking (DCM). In **Chapter 5**, we use DCM as another case study to validate the broader applicability of NeLLCom-X, and investigate the following sub-questions:

RQ-D.1 *To what extent does the DCM observed in communicative agents' production resemble that of human participants?*

The underlying mechanism of DCM remains debated. In human language experiments, Fedzechkina et al. (2012) propose that DCM arises from learning. However, Smith and Culbertson (2020) found different results, suggesting that DCM emerges in real communication rather than through learning. The two-fold experimental design of Smith and Culbertson (2020), including a learning phase followed by interaction, aligns well with the general idea of NeLLCom, making the agent-human comparison particularly relevant in this context. We follow their setup and adopt their artificial language, specifically one that is designed to simulate real flexible-order languages, where one order is typically dominant over the others.

RQ-D.2 *How does the order distribution of the initial language affect the emergence of DCM in neural agents?*

We hypothesize that an initially uneven word order distribution, while typical in natural languages, may constitute a confounder in the simulation of DCM. Namely, neural agents may amplify input biases in general as a form of regularization, and in turn this may complicate the interpretation of the results. To disentangle input bias from the emergence of DCM, we also experiment with a neutral-order language where SOV and OSV are evenly distributed.

Aligning with the claims of Smith and Culbertson (2020), we find that neural

agents develop a human-like DCM pattern after interaction in both dominant-order and neutral-order setups, highlighting the critical role of communication in shaping DCM. Additionally, we observe that the initially neutral-order language leads to a more pronounced differential marking of objects and subjects.

RQ-D is based on the following research article:

> Yuchen Lian, Arianna Bisazza, and Tessa Verhoef. 2025. Simulating the emergence of differential case marking with communicating neural-network agents. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*