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

Lian, Y.

Chapter 5

Simulating the Emergence of Differential Case Marking with NeLLCom-X

Multi-agent reinforcement learning frameworks based on neural networks have gained significant interest to simulate the emergence of human-like linguistic phenomena. NeLLCom(-X) is a framework proposed in **Chapter 3** and **Chapter 4**, in which agents first acquire an artificial language before engaging in communicative interactions, enabling direct comparisons to human result. NeLLCom(-X) implements neural-network learners that have no prior experience with language or semantic preferences, and the framework uses a very generic communication optimization algorithm to model interactions between language learners. Previous chapters have demonstrated the success of NeLLCom(-X) in replicating the emergence of language universals, using the word-order/case-marking trade-off as a case study. This chapter demonstrates the adaptability of NeLLCom(-X) to another linguistic phenomenon. Concretely, we ask:

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

In natural language, marker use is influenced not only by word order but also by semantic and pragmatic properties of arguments, a phenomenon known as differential case marking (DCM). The emergence of DCM has been studied in artificial language learning experiments with human participants, which were specifically aimed at disentangling the effects of learning from those of communication (Smith and Culbertson, 2020). In this chapter, we use DCM as another case study to further evaluate NeLLCom(-X). We follow the language design of (Fedzechkina et al., 2012) and (Smith and Culbertson, 2020). Specifically, we use individual agent supervised learning to simulate the human learning phase, and let agents play language games to simulate the human interaction phase. With NeLLCom(-X), we succeed in replicating the emergence of DCM after agents communication, supporting Smith and Culbertson (2020)'s findings highlighting the critical role of communication in shaping DCM.

Chapter adapted from:
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)*

5.1 Introduction

Human language is not a static entity but a dynamic system undergoing continuous change and evolution. The development of agent-based models is a productive approach to studying the emergence and change of linguistic systems, which has a long-standing tradition in the study of language evolution (Hurford, 1989; Hare and Elman, 1995; Steels, 1997; De Boer, 2006). Recent advancements in computational linguistics and deep learning have reinvigorated interest in such simulations, providing the opportunity to model increasingly realistic phenomena (Chaabouni et al., 2021; Lian et al., 2021, 2023). These models simulate the spontaneous development of communication systems through repeated interactions among individual neural-network agents (Lazaridou et al., 2017; Havrylov and Titov, 2017; Chaabouni et al., 2020; Boldt and Mortensen, 2024). A key challenge in this area is that the languages developed by these agents, when initialized from scratch, often lack human-like characteristics (Chaabouni et al., 2019a; Lian et al., 2021; Galke et al., 2022). An alternative approach is employed in the Neural agent Language Learning and Communication (NeLLCom) framework promoted in **Chapter 3** (Lian et al., 2023), where agents first learn a predefined miniature artificial language and then use it for communication, with the goal of studying the emergence of specific linguistic properties.

For instance, NeLLCom has been used to investigate the emergence of a trade-off between case-marking and word-order strategies (shown as in **Chapter 3** (Lian et al., 2023) and **Chapter 4** (Lian et al., 2024)), a phenomenon commonly observed in natural languages. Case marking and word order are both strategies to indicate who does what to whom in a sentence and languages often rely more heavily on one strategy than the other. This trade-off had previously been found to emerge in artificial language learning (ALL) experiments with humans (Fedzechkina et al., 2017), where participants dropped the use of markers more often if they were learning artificial languages with fixed word order than in the case of flexible word order. Having adapted the experimental design and artificial languages of Fedzechkina et al. (2017) to train neural network agents, Lian et al. (2023)(**Chapter 3**) found human-like patterns of language change only when agents actively attempted to be understood by a communication partner. Communication provided a pressure for the language to develop towards a form that makes it maximally efficient without losing communicative success. Simplifying the language by dropping the markers was possible when word order provided enough cues to derive the meaning correctly.

In naturally occurring human languages, word order is not the only factor that may influence the efficient use of markers. Even in languages with flexible word order, case-marking is frequently employed selectively rather than universally (De Hoop and Malchukov, 2008; Sinnemäki, 2014; Witzlack-Makarevich and Seržant, 2018; Levshina, 2021). This paper focuses on Differential Case Marking (DCM), a widespread phenomenon observed across many languages, where the morphological marking of a grammatical case varies depending on semantic, pragmatic or other factors. For example, in many languages, animate objects are explicitly marked to clarify their role in a sentence (García, 2018; Levshina, 2021) since animate entities typically appear in the subject instead of the object role (e.g. in an event involving *eating*, *cake* and *Alice*, we typically assume Alice to be doing the eating). The emergence of this phenomenon has been explored through various ALL experiments with human participants (Smith and Culbertson (2020), henceforth S&C; Fedzechkina et al. (2012), henceforth FJN), particularly in relation to the roles of learning and communication which S&C explicitly sought to disentangle. This makes the DCM phenomenon highly suitable for investigation using the NeLLCom framework, where effects of communication and learning can be simulated with a generic communication protocol and linguistically naïve learners, while closely replicating the experimental setups and language design of FJN and S&C. Our results provide additional evidence supporting S&C's proposal that communication plays a crucial role in shaping the emergence of DCM.

5.2 Differential Case Marking

Differential case marking, or differential argument marking, refers to a widespread cross-linguistic phenomenon in which the formal marking strategy for an argument differs according to its semantic, pragmatic, or other properties (De Hoop and Malchukov, 2008; Sinnemäki, 2014; Witzlack-Makarevich and Seržant, 2018; Levshina, 2021). More specific instances of DCM are Differential Object Marking (DOM) and Differential Subject Marking (DSM). An example (adapted from García (2018); Levshina (2021)) of DOM is the marking of human objects in Spanish:

(1) 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.'

where the atypical animate/human object in (b) is marked using *'a'*. Similarly, in DSM, typical subjects can remain unmarked while atypical subjects are more likely to be marked.

There is a long-standing debate about the mechanisms that cause this phenomenon to develop. Typological studies (Aissen, 2003; Croft, 2003; Levshina, 2021), artificial language experiments (Fedzechkina et al., 2012; Smith and Culbertson, 2020; Tal et al., 2022), and computational simulations (Lestrade, 2018) have been conducted to explore potential explanations. Levshina (2021) broadly contrasts two explanations for the emergence of DCM, which have been previously discussed in the literature. The first considers this phenomenon to be a result of efficient communication strategies, where markers are used more in cases where the probability to be misunderstood without them is higher. The second invokes markedness theory, where unmarked linguistic forms (the default or neutral forms that are more frequent and simpler in a given context) tend to be used for typical events, while (more atypical and complex) marked forms are iconically associated with atypical events. Corpus-based quantitative analyses suggest that the first (efficient communication) is a better predictor of cross-linguistic patterns observed in DCM (Levshina, 2021). Lestrade (2018) simulated the emergence of DCM with relatively simple interacting agents in a model where (in contrast to our work) marking strategies and grammaticalization principles were explicitly built-in to see what their combined or separate impact was. Their results suggested that argument marking can evolve gradually as languages adapt to usage.

In a laboratory setting, FJN conducted a set of ALL experiments, where human participants watched computer-generated videos and heard their descriptions in a novel artificial language. After 4 days of learning, researchers found that the learners' productions deviated from their input language towards more efficient case-marking systems (using markers more often for atypical arguments like animate objects or inanimate subjects, than in typical situations). The authors therefore conclude that language learners restructure their linguistic input so that it increasingly facilitates efficient communication. As pointed out by S&C, this interpretation is surprising, since it is more typically assumed that language *use*, and not *learning*, drives its evolution towards communicative efficiency (Kirby et al., 2015; Kemp et al., 2018; Gibson et al., 2019). Perhaps even more surprising, the language design by FJN was such that in the 10 animate nouns in the lexicon of the object marking language for example, 5 only occur as subjects and the other 5 as objects. The meaning was therefore potentially unambiguous regardless of the presence of a marker, which conflicts with the efficiency account (where disambiguation for a listener is assumed to drive the effect). Notably, these results were also consistent with an explanation based on markedness theory and iconicity instead of efficient communication (S&C). S&C adapted and extended the experiments of FJN, in a large-scale study (n $>$ 300), and introduced an interaction phase after the last day of learning where participants use the language to communicate with a simulated interlocutor implemented as a simple chatbot. Their findings do not replicate those of FJN. Instead, they suggest that learning alone cannot reliably explain the emergence of DOM, but actual communicative interaction is key to the emergence of a communicatively-efficient case marking system. Complementing these findings, we simulate neural-agent learning and communication with agents that do not have any iconic preference, linguistic knowledge or sense of animacy. This allows us to investigate whether typicality alone can lead to a DCM effect, and whether communicative pressures are a necessary factor for the emergence of DCM in neural learners, like in humans.

5.3 NeLLCom Framework

Artificial language learning has been widely used in experiments with humans (Fedzechkina et al., 2016; Culbertson, 2023), as it provides a means to isolate specific linguistic phenomena and study cause-and-effect relationships in a controlled setting. These human-based studies can serve as valuable inspiration for the design of emergent communication simulations, allowing for direct comparisons between human and agent behaviors. The NeLLCom framework (**Chapter 3** (Lian et al., 2023) and **Chapter 4** (Lian et al., 2024)) is a multi-agent communication framework designed to simulate ALL experiments for the study of language change and evolution. After being trained on an initial artificial language through supervised learning, agents in this framework start interacting via meaning reconstruction games in which they optimize a shared communicative reward through reinforcement learning.

NeLLCom (**Chapter 3**) enables researchers to scale ALL experiments in ways that are difficult to achieve with human participants. While Fedzechkina et al. (2017) focused solely on individual learning by human participants, **Chapter 4** (Lian et al., 2024) expanded it on the word-order/case-marking trade-off using NeLLCom-X, incorporating more realistic role-alternating agents and group communication. This extension demonstrated that the trade-off also emerges in populations of communicating individuals, which is something that would be rather difficult and expensive to achieve with human participants in a lab. Here we use the most recent version of the framework, NeLLCom-X.

The Task In NeLLCom-X, **meanings** describe simple scenes using triplets $m = \{Action,\ agent,\ patient\}$ (e.g., EAT, ALICE, CAKE). An artificial **language** is defined by a set of grammatical rules generating utterances u from a fixed-size vocabulary to convey meaning m. Utterances can be of variable length and multiple u candidates can be valid for the same m. In the meaning reconstruction game, a speaker conveys a meaning m by generating an utterance $\hat{u}$, which the listener then maps to meaning $\hat{m}$. The game is successful if $m = \hat{m}$.

Agent Architecture

The structures of listening and speaking networks are symmetric with meanings represented by unordered tuples while utterances are generated/processed sequentially. This results in a **linear-to-RNN** (Recurrent Neural Network) speaking network $\mathcal{S} : m \mapsto u$ and a **RNN-to-linear** listening network $\mathcal{L} : u \mapsto m$. An agent then includes two sets of parameters $\alpha_i = (N_i^{\mathcal{S}}, N_i^{\mathcal{L}})$ tied together through parameter sharing of their meaning and word embeddings.

Training

Before communication, each agent is first trained by Supervised Learning (**SL**). Using a set of reference meaning-utterance pairs $D = (m, u)$ and teacher forcing, this phase minimizes the cross-entropy loss between u and the words generated by the speaker given m. Conversely, for the listener, SL minimizes the loss between m and the meaning tuple generated by the listener given u. Then, two (or more[1]) trained agents α_0 and α_1 learn to communicate with each other via Reinforcement Learning (**RL**). During this phase, agents maximize a shared communication reward $r(m, \hat{u})$ which captures how close the listener's prediction $\mathcal{L}(\hat{u})$ given the speaker-generated utterance $\hat{u} = \mathcal{S}(m)$ is to m. See more details in **Chapter** 4 (Lian et al., 2024).[2]

5.4 Experimental Setup

We use NeLLCom-X to simulate the emergence of DCM in neural agents, following the language design of FJN and S&C as explained in this section.

Meaning Space

As previously mentioned, DCM implies that marker production can differ depending on the typicality of the entities in a sentence. Mirroring human languages where animate agents (e.g. *Alice*) and inanimate patients (e.g. *cake*) are more typical, the **Object-Marking** condition in FJN and S&C defines a meaning space where agents are always animate entities, while patients can be either animate or inanimate. Conversely, in the **Subject-Marking** condition, patients are always inanimate, while agents can be either animate or inanimate. Note that, in the human experiments, animacy referred to a property of the entities depicted in the stimuli, which were concepts previously

[1] We only consider two-agent communication in this work. However NeLLCom-X can model group communication with more than two individuals by iteratively sampling pairs of two agents from the group to proceed with an interaction.

[2] In each interaction turn, each agent is assigned to a role (speaker or listener) with equal probability. To ensure self-understanding is maintained, rounds of interaction between different agents are interleaved at regular intervals with rounds of *self-communication* where an agent's speaking network sends messages to its own listening network.

known to the participants (e.g. animate *artist, baker*, etc. versus inanimate *ball, cake*, etc.). By contrast, neural networks are trained from scratch and have no previous world knowledge. In our setup, all entities are encoded in the same way (as entries of the meaning embedding table, all randomly initialized), and the typicality of an entity's role is inferred from the statistical properties of the observed meaning space (e.g. 'entity-3' occurring half of the times as agent and the other half as patient, versus 'entity-5' occurring always as patient). Working with neural agents therefore allows us to tease apart the effect of typicality as a purely statistical property from prior animacy associations, which was not possible in the setup of S&C's or FJN's human experiments. We call *Amb* (ambiguous) the subset of entities that can have two roles, and $\neg Amb$ (unambiguous) the subset of entities that can only occur in one role. Thus, possible meaning structures are $\{A, a_{Amb}, p_{\neg Amb}\}$ and $\{A, a_{Amb}, p_{Amb}\}$ in the Object-Marking language; $\{A, a_{\neg Amb}, p_{Amb}\}$ and $\{A, a_{Amb}, p_{Amb}\}$ in the Subject-Marking language.[3]

In each language condition, 20 entities (10 ambiguous and 10 unambiguous) and 8 actions are included, resulting in a total of $10 * (10 + (10 - 1)) * 8 = 1520$ possible meanings. This expanded meaning space results in a better model convergence in preliminary experiments (Zhao et al., 2018; Chaabouni et al., 2020), compared to the relatively small space used in human experiments (10 entities and 4 verbs).

Artificial Languages

Following FJN and S&C, we adopt verb-final languages allowing SOV or OSV orders in varying proportions. The token *'mk'* serves as a case marker and is optionally assigned to either the subject or object based on the language type. For example, given the meaning m=$\{A$: EAT, a: ALICE, p: CAKE$\}$, flexible-order object-marking languages admit four utterances: *'Alice cake eat'*, *'Alice cake mk eat'*, *'cake Alice eat'*, and *'cake mk Alice eat'*.

A specific language is defined by four factors: whether it is object- or subject-

[3] Note this setup corresponds to the 'Subjects Can Be Objects' condition introduced by S&C.

Table 5.1: The miniature languages used in this study.

language	mark	SOV	mk|SOV	mk|OSV
dominant-order (S&C, exp.1)	obj	60%	67%	50%
neutral-order	obj	50%	67%	67%
	subj	50%	67%	67%

marking, its order profile $p(SOV)$, marking proportion in the SOV order $p(mk|SOV)$, and marking proportion in the OSV order $p(mk|OSV)$. We consider two types of languages. The first **dominant order language** replicates one of S&C's target languages, which was in turn designed following FJN. This is an object-marking language with $p(SOV) = 60\%$, $p(mk|SOV) = 67\%$, and $p(mk|OSV) = 50\%$, resulting in an overall 60% marking proportion. This language was designed by FJN to simulate real flexible-order languages, where one order is typically dominant. However, a limitation of this design is that neural agents may amplify the initial bias towards using more SOV order and marking SOV utterances more often than OSV ones, driven by a generic pressure to regularize their input. We thus expect agents to drift towards a strongly SOV and strongly SOV-marking solution, just because those were the most frequently observed patterns in the training data.

To disentangle input bias amplification from the actual agents' preferences towards different DCM strategies, we also experiment with a **neutral order language**, where SOV and OSV are evenly distributed, and marking proportion is 67%.[4] We implement both an object-marking and a subject-marking version of this language. Table 5.1 summarizes the three languages used in our experiments.

Following the previous **Chapter** 4 (Lian et al., 2024), each entity corresponds to a word, resulting in the fixed-size vocabulary = 20 + 8 + 1 = 29.

[4] In preliminary experiments starting from 50% case marking, we observed a strong tendency of the agents to drop markers altogether, making it impossible to explore the DCM effect.

5.4.1 Evaluation

Accuracy and production preference evaluation are adopted from the previous **Chapter** 3 and **Chapter** 4. All evaluations are based on an unseen meaning set M_{test}. In the SL phase, performance is measured by listening and speaking accuracy against the reference dataset. In the RL phase, communication success is evaluated by meaning reconstruction accuracy, where $acc(m, \hat{m})$ equals 1 iff the entire meaning is matched. Production preferences are visualized as the proportion of markers and different orders in a set of utterances generated by an agent for M_{test}, after discarding utterances that are not well-formed according to the language grammar.[5] Furthermore, we split M_{test} into ambiguous $M_{test,\ Amb}$ and unambiguous $M_{test,\neg Amb}$ meanings, and evaluate the two sets separately.

We use generalized linear mixed-effects models (GLMMs) in the lme4 package version 1.1-35 (Bates et al., 2015) with R Version 4.4.2 (R Core Team, 2024) to evaluate how the marking proportion and word order preferences are influenced by ambiguity after communication and whether marking use is conditioned on word order.

5.4.2 Model Configuration and Training Details

We apply the same configuration as the previous **Chapter** 4 (Lian et al., 2024). The sequential layer in speaking and listening networks consists of 16-dimensional Gated Recurrent Units (Chung et al., 2014). The shared meaning embeddings and shared word embeddings have 8 and 16 dimensions, respectively. Utterance length for the speaker is limited to 10 words.

We first split the dataset D into 80/20% training/test samples. The test split (304 meanings) is used throughout the whole evaluation. During SL, we resample 66.7% meanings out of the first train set following the all-seen-entities/actions rule as in **Chapter** 3 (Lian et al., 2023). SL iterates 60 times

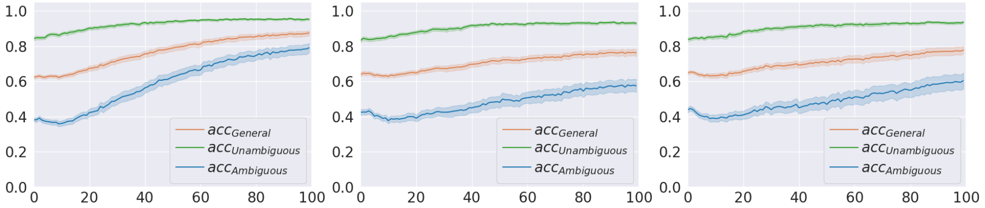
[5] In our experiments, the ratio of non well-formed utterances averaged over seeds is around 10% before RL and 21-25% (depending on the initial language) after RL, which is overall comparable to the results in **Chapter** 4 (Lian et al., 2024).

with a 0.01 learning rate using the default Adam optimizer. During RL, 320 meanings are sampled from the first train set and used as the training samples for each communication turn. RL iterates 200 inter_turns with a 0.005 learning rate using the REINFORCE algorithm (Williams, 1992). Batch size is set to 32 in both SL and RL training. We repeat each language setup with 50 pairs of agents (i.e. 100 random seeds).

5.5 Results

5.5.1 Dominant Order Language

Results for the language adopted from S&C are presented in the Figure 5.1a and **first row** of Figure 5.2. **Before the start of RL**, communication accuracy is already around 60% (Figure 5.1a) reflecting a relatively high speaking and listening accuracy acquired by the agents at the end of SL (81% and 78% respectively; results not shown in the plots). When analyzing agent's performance conditioned on meaning ambiguity, we find that communication accuracy before RL is much higher for $M_{test,\neg Amb}$ than for $M_{test,\ Amb}$, which was expected and matches the human results of S&C. Additionally, production preferences before RL (columns 2 and 3, pink cluster around the solid diamond) closely align with the original proportions of the artificial language, reflecting the typical post-SL probability-matching behavior observed in previous work (Chaabouni et al., 2019b; Lian et al., 2023).



(a) dominant-order **(b)** neutral (OBJ) **(c)** neutral (SUBJ)

Figure 5.1: Meaning reconstruction accuracy across communication rounds, computed on the whole test set (orange line), as well as split by non-ambiguous (green) and non-ambiguous (blue) meanings. Each experiment is repeated with 50 agent pairs.

Effects of communication The increase in overall communication accuracy (orange line) indicates that agents optimize their language during interaction. This is confirmed by the clear shift in production preferences shown in columns 1 and 2 (first row). More specifically, the average preferences after interaction (solid purple circle), indicate a decrease in marker proportion alongside a significant shift towards fully using SOV order.

We further analyze changes in production preferences conditioned on ambiguity. For $M_{test,\neg Amb}$, column 1 reveals a general decrease in marker usage and an increase in the preference for the SOV order. Additionally, we observe a linear relationship between marker proportion and SOV proportion: the more frequently the SOV order is used, the more markers are generated ($b = 3.62,\ SE = 0.31,\ p < 0.001$). This could be due to the fact that the input language also has more markers in SOV utterances than for OSV.

For the remaining meanings, $M_{test,\ Amb}$, an even stronger preference for the SOV order is observed, together with a decrease in marker usage, as shown in column 2. This suggests that, after communication, agents resolve ambiguity by regularizing the word order to SOV, instead of increasing marker proportion, which in turn may reflect a general tendency to amplify biases in the original language.

To investigate whether a DOM effect emerges in the productions of neural agents, we visualize the individual-level production differences between ambiguous and unambiguous patients (column 3). On average, we observe only a small, but significant, difference in marker usage. While there is a general decrease in marker use, agents retain more markers ($b = 0.25,\ SE = 0.08,\ p < 0.01$) for $M_{test,\ Amb}$. A more noticeable difference in word order preference is observed: after communication, agents regularize more towards SOV ($b = 3.43,\ SE = 0.22,\ p < 0.001$) on $M_{test,\ Amb}$ as compared to $M_{test,\neg Amb}$. Typicality therefore has significant effects on both case marking and word order.

Comparison to human results

While human participants in S&C tend to increase the use of markers during communication, we observe a general decrease in marker use in our agent interactions. Instead of producing more markers, the agents tend to regularize towards one consistent word order to disambiguate the meanings. Even though human learners in S&C also frequently started over-producing one order versus the other, they still introduced more markers in communication to increase the chance of being correctly understood. Despite these differences, we do see a human-like DOM effect appear during agent interactions, where markers are used significantly more frequently for $M_{test,\ Amb}$, just like the increased marker usage of human participants for animate objects in S&C.

Another notable difference concerns whether marker use is conditioned on word order. In the language design we adopted from S&C, the initial case marking proportion is higher for SOV than OSV sentences. Human participants did not maintain these differences while learning the 60%-SOV with 60% marking language, but as shown above our agents keep conditioning marker use on word order even during communication. Since the agents seem to be more sensitive to existing patterns present in the initial language, we continue our analyses with the two languages with neutral word order and no conditioning of marker use on word order.

5.5.2 Neutral Order Languages

Results for the languages with initial 50/50% SOV/OSV order are shown in the Figure 5.1 ((b) and (c)), and **second and third rows** of Figure 5.2 (object-marking and subject-marking version, respectively). **Before the start of RL**, communication accuracies for both languages are very similar to those of the dominant order language, and so are the production preferences, again reflecting probability matching behavior.

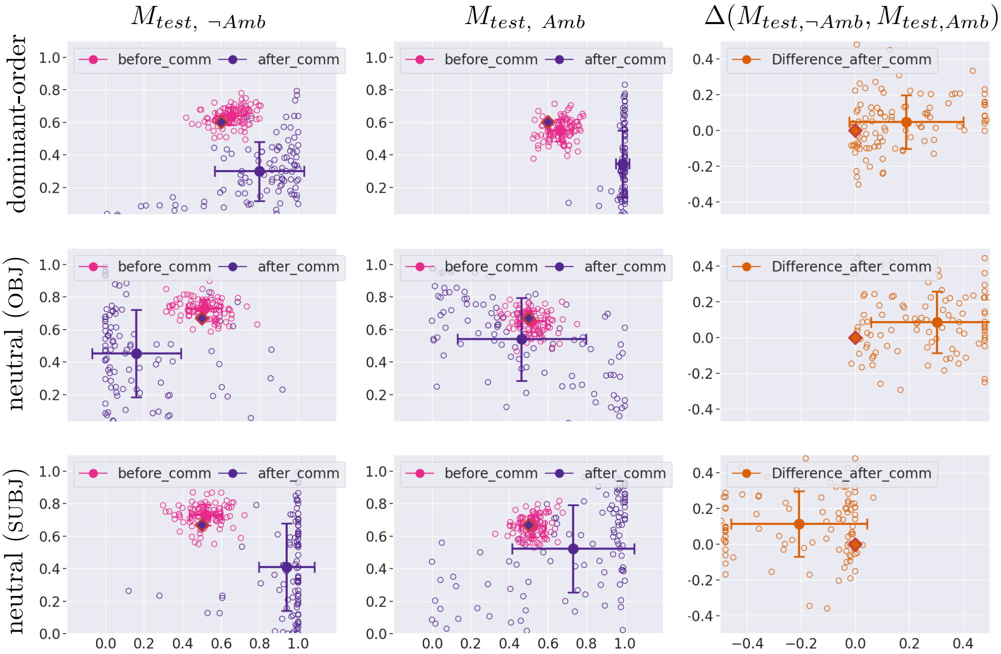


Figure 5.2: Production preferences (PP) in terms of order proportions and marker use. Specifically, Col. 1 and 2 show PP for non-ambiguous and ambiguous meanings respectively, before and after communication, and Col. 3 shows the difference in PP between $M_{test,\neg Amb}$ and $M_{test,\ Amb}$ after communication. Solid diamonds mark the initial language. Each empty circle is an agent and solid circles are the average of all agents, with error bars showing standard deviation. Each experiment is repeated with 50 agent pairs.

Effects of communication

Starting with the object-marking language (second row), we again find a drop in the marking proportion, which is present for both ambiguous and unambiguous meanings (columns 1 and 2), but again markers are retained more for $M_{test,\ Amb}$ ($b = 0.49,\ SE = 0.10,\ p < 0.001$). The development of word order is very different from what was observed for the dominant order object-marking language. Instead of amplifying the already present majority of SOV in the dominant language, agents exposed to the neutral object-marking language develop a clear preference for OSV, which is significantly stronger for $M_{test,\neg Amb}$ ($b = 2.75,\ SE = 0.17,\ p < 0.001$) as compared to $M_{test,\ Amb}$. In sum, we again see that typicality has a significant effect on both order and case marking. A linear relation between word order and marker use appears for $M_{test,\ Amb}$, where less markers are used when SOV is more frequent ($b = -1.98,\ SE = 0.15,\ p < 0.001$).

As expected, results for the subject-marking language (third row) show a symmetric trend where, again, markers persist significantly more for $M_{test,\ Amb}$ ($b = 0.70,\ SE = 0.10,\ p < 0.001$), but the order preference is reversed, where SOV is used for $M_{test,\neg Amb}$ significantly more ($b = 3.05,\ SE = 0.25,\ p < 0.001$) than for $M_{test,\ Amb}$. These contrasting order preferences between the neutral object-marking and subject-marking languages seem to indicate an agent preference to put the marked entity first. In addition, the relation between word order and marker use for $M_{test,\ Amb}$ is reversed for the subject-marking language, with more markers used when SOV is more frequent ($b = 1.72,\ SE = 0.17,\ p < 0.001$). Interestingly, FJN similarly found that markers were used more frequently with SOV in their subject-marking language in the early stages of learning, while this was the case for OSV in the object-marking language. Since there is no order-conditioned case marking in the neutral input languages for our agents, these linear relationships could suggest that generating an utterance for $M_{test,\ Amb}$ in the majority order creates a need for an added marker to be reliably understood, while using the other order serves, in itself, as a way to disambiguate.

5.6 Discussion and Conclusion

We used NeLLCom-X to study the emergence of Differential Case Marking, employing previous experimental set-ups of human studies by FJN and S&C. Neural agents do not have the same biases in learning and signal production as humans, so differences in preferences between agents and humans after learning and communication are expected. Indeed, we saw that our 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. While our agents learned about typicality of entities solely based on statistical properties in the artificial language, human participants in FJN and S&C already had knowledge about animacy in addition to this. Moreover, human participants have existing preferences to iconically relate marked forms with atypical events (Aissen, 2003; Haspelmath, 2008), while agents have no such bias. Finally, humans have a preference to place human and animate entities before inanimates in a sentence (Aissen, 2003), while our agents are not aware of these distinctions. The interacting effects of all these biases can make it difficult to tease apart causal mechanisms contributing to the emergence of DCM when working with human participants. As discussed in the introduction, FJN and S&C indeed found conflicting results when looking at the role of learning. Complementing these previous findings, and supporting S&C's conclusions, our simulations demonstrate that DCM does not arise as the result of learning, but does emerge when agents start communicating in pairs. Importantly, our agent set-up allowed to study these factors in the absence of prior language experience and sense of animacy or iconicity in the learners.

Beyond replicating human artificial language learning results with linguistically naïve neural learners, employing NeLLCom also offers advantages in scalability. Using neural agents, we can conduct numerous iterations and explore diverse language conditions, which would be costly and time-consuming in human laboratory experiments. For example, studying real communication between two or more interacting participants would have been hard to coordinate with the large number of (online) participants included in S&C's study, which may explain their use of a chatbot. In our setup, we could easily model pairs of interacting agents, and this can just as easily be extended to groups. Additional directions for future work include experimenting with a less clear-cut distinction between entity-role distributions (e.g. 55/45% and 5%/95%, rather than 50/50% and 0/100%), which would more closely resemble real-language distributions. Another way to possibly achieve more human-like patterns would be to endow agents with a notion of animacy by initializing them with meaning embeddings pre-trained on large text corpora.

To conclude, NeLLCom-X can be used 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 learning and language use differs from that of humans.