Learning and communication pressures in neural networks: Lessons from emergent communication

Lukas Galke
Centre for Machine Learning, Department of Mathematics and Computer Science (IMADA), University of Southern Denmark (SDU), Odense, Denmark
LEADS group, Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands

Limor Raviv
LEADS group, Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands
Centre for Social, Cognitive and Affective Neuroscience, University of Glasgow, Glasgow, UK

Abstract: Finding and facilitating commonalities between the linguistic behaviors of large language models and humans could lead to major breakthroughs in our understanding of the acquisition, processing, and evolution of language. However, most findings on human–LLM similarity can be attributed to training on human data. The field of emergent machine-to-machine communication provides an ideal testbed for discovering which pressures are neural agents naturally exposed to when learning to communicate in isolation, without any human language to start with. Here, we review three cases where mismatches between the emergent linguistic behavior of neural agents and humans were resolved thanks to introducing theoretically-motivated inductive biases. By contrasting humans, large language models, and emergent communication agents, we then identify key pressures at play for language learning and emergence: communicative success, production effort, learnability, and other psycho-/sociolinguistic factors. We discuss their implications and relevance to the field of language evolution and acquisition. By mapping out the necessary inductive biases that make agents' emergent languages more human-like, we not only shed light on the underlying principles of human cognition and communication, but also inform and improve the very use of these models as valuable scientific tools for studying language learning, processing, use, and representation more broadly.

Keywords: language acquisition; language evolution; emergent communication; large language models; learning biases; learning pressures; neural networks; neural language models; multi-agent systems

Corresponding author: Lukas Galke, Centre for Machine Learning, Department of Mathematics and Computer Science (IMADA), University of Southern Denmark (SDU), Campusvej 55, DK-5230 Odense M, Denmark. Email: galke@imada.sdu.dk

ORCID ID: https://orcid.org/0000-0001-6124-1092

Citation: Galke, L. & Raviv, L. (2024). Learning and communication pressures in neural networks: Lessons from emergent communication. *Language Development Research* *5*(1), 116–143. http://doi.org/10.34842/3vr5-5r49

Introduction

Using neural language models for language development research dates back to Elman (1993) simulating language acquisition with recurrent neural networks and conceiving the theory of "the importance of starting small". Similarly, Harris (1954)'s distributional structure has motivated word embeddings – a seminal work showing that the semantic relationship between words can be learned without supervision from text data alone (Goth, 2016; Mikolov et al., 2013). These are just some examples of where machine learning has already influenced the development and testing of linguistic theories, showcasing a thriving relationship between the two disciplines (Baroni, 2021; Contreras Kallens et al., 2023; De Seyssel et al., 2023; Dupoux, 2018). The unprecedented success of language models in recent years (Bahdanau et al., 2015; Brown et al., 2020; Devlin et al., 2019; Raffel et al., 2020; Vaswani et al., 2017) provides many opportunities to further advance our understanding of human language learning.

A growing body of work has found similarities between large language models and humans (Dasgupta et al., 2022; Schrimpf et al., 2021; Srikant et al., 2022; Webb et al., 2023; Wei et al., 2022), showing that approximate representations of the outside world can be learned from statistical patterns found in linguistic input alone (Abdou et al., 2021; B. Z. Li et al., 2021; K. Li et al., 2023; Patel & Pavlick, 2022), and manifesting the usefulness of large language models for other disciplines such as psychology (Demszky et al., 2023). However, a so far open issue is the fact that language models are exposed to different input modalities (i.e., mainly text) and have much more data available for training than humans (De Seyssel et al., 2023; Warstadt & Bowman, 2022). Resolving the discrepancy by which language models require much more data than a human child is of high interest to both cognitive science (with the goal of more representative models) and natural language processing researchers (with the goal of more efficient models). Notably, there are ongoing efforts to train language models from similar input as available to a human child, e. g., as in BabyBERTa (Huebner et al., 2021), and the BabyLM challenge[1] (Warstadt et al., 2023).

To promote a deeper understanding of how large language models may be useful for language development research, we suggest to take inspiration from the field of emergent machine-to-machine communication – where two or more neural network agents without exposure to an existing language need to engage in a communication game with the goal of successfully understanding each other (Foerster et al., 2016; Kottur et al., 2017; Lazaridou & Baroni, 2020; Lazaridou et al., 2017). Specifically, emergent communication simulations explore what happens when artificial neural networks (on which also large language models are based) need to create their own languages from scratch, i.e., without first being pre-trained on natural language corpora: do they create human-like languages by-default, or are there specific biases and constraints that

[1] https://babylm.github.io

need to be introduced in order to replicate human behavior? By attempting to simulate phenomena previously observed in humans, research on emergent communication has provided valuable insights into the processes and pressures that shape the evolution of human language, and has allowed researchers to effectively scrutinize, identify, and tease apart the relevant learning biases and conditions that underlie the communicative behaviors of artificial neural networks when they are made to communicate by themselves.

Although the setting of emergent communication is typically motivated for studying the evolution of language (see Lazaridou & Baroni, 2020; Lian et al., 2023, inter alia), language learning and language evolution are intrinsically linked: As languages are passed from generation to generation in a repeated cycle of transmission, imitation, and use, their structure is continuously shaped by the pressures and biases introduced by learners during the process of language acquisition – with such learning biases effectively shaping the evolution of languages on a longer timescale (Chater & Christiansen, 2010; Kirby et al., 2014; Smith, 2022). As such, constraints and pressures associated with learning can causally affect (and, in fact, create) the universal properties of languages, including their most fundamental structural features (Kirby, 2002, 2017; Kirby et al., 2004). As such, we believe that the field of emergent communication provides an ideal testbed for exploring the learning pressures neural networks are exposed to in the process of language learning and use, and can help shed light on (some of) the criticial inductive biases needed for replicating human linguistic behavior.

Since the theoretical usefulness of a model is dependent on its resemblance to the target entity (Zeigler et al., 2000), identifying the relevant learning pressures and biases that govern language creation in neural network models can in turn make neural language models more behaviorally plausible, and consequentially a more robust scientific tool for the language sciences. Here, we review the emergent communication literature and identify underlying learning pressures, while contrasting those with the learning pressures at play when training large language models. Thereby we shed new light on the learning dynamics of neural language models and contribute to the development of more behaviorally plausible language models for language acquisition research.

In the following, we offer a comparative perspective on humans, large language models, and deep learning agents engaging in communication games by reviewing similarities and differences in observed phenomena, discussing how mismatches in the behavior of humans and neural agents can be resolved through appropriate inductive biases, and determining the underlying learning pressures at play. We first provide a brief overview of the emergent communication literature, and then showcase initial mismatches between neural agents and humans with respect to multiple linguistic phenomena: Zipf's law of abbreviation, the benefits of compositional structure, and social factors shaping linguistic diversity (e.g., population size effects). For each of these phenomena, we describe how the initial mismatch between humans and neural network models has been

resolved, and identify the underlying learning pressures giving rise to these patterns. In particular, we identify four cognitive and communicative pressures underlying both language acquisition and language evolution, and discuss whether they are inherent to the training objective (i.e., present by default given the learning environment and objective) or whether they need to be artificially incorporated into the models as inductive biases to elicit the desired outcome. We then contrast the identified pressures and biases with those present in the training of large language models, with the goal of promoting knowledge transfer between machine learning and language sciences. We conclude with concrete suggestions for future directions, aimed at developing more cognitively plausible language models for both language development and language evolution research.

Emergent communication, initial mismatches, and their resolution

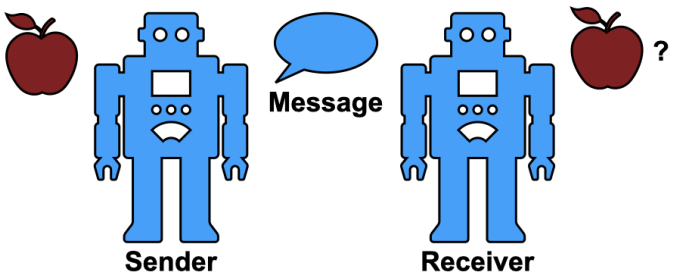

Figure 1. ***Schematics of a simple communication game. The sender sees an object and has to compose a message to describe it. The receiver only sees the message and has to discriminate the object against distractors, or fully reconstruct it.***

Computational modeling has long been used to study language evolution by simulating the process of communication and transmission between artificial agents, typically Bayesian learners (Dale & Lupyan, 2012; Gong et al., 2008; Kirby, 2002; Kirby et al., 2004; Kirby et al., 2015; Perfors & Navarro, 2014; Smith et al., 2003; Smith & Kirby, 2008; Steels, 2016). The emergence of new communication systems is similarly studied using deep neural network models (Lazaridou & Baroni, 2020), and in experimental work with human participants (Kirby et al., 2008; Raviv et al., 2019b; Selten & Warglien, 2007; Winters et al., 2015). Regardless of whether the subjects of these experiments are humans, Bayesian agents, or deep neural networks, they all share the same methodological framework, namely, sender-receiver communication games: One agent describes an input (e. g., an object or a scene), and transmits a message to another agent, that then has to guess or fully reconstruct the sender's input (see Figure 1). The agents in emergent communication experiments are typically based on deep neural networks, similar to those used in large language models.

Table 1: Observed phenomena from humans in agents from emergent communication simulations

Phenomenon in Humans	**Mismatch in Emergent Communication agents**	**Resolution**
Zipfian distribution in utterance length (frequent meanings are described by shorter utterances)	Sender agents exploit the full channel capacity because longer messages are easier to distinguish by receiver agents.	Introducing a penalty on long utterances (simulating "laziness") restores the Zipfian distribution on utterance length.
Compositional structure reliably emerges during communication and cultural transmission, and is beneficial for language learning and generalization	Inconsistent emergence of compositional structure in neural agents, and seemingly no advantage of more compositional protocols for generalization	Periodically resetting agents' parameters (simulating generational turnover) gives rise to compositional protocols, which are easier to learn for neural network agents
Population size affects the emergence of compositional structure (larger communities create more systematic languages)	Larger populations of neural agents do not create more compositional protocols	Introducing population heterogeneity (simulating individual differences) or production-comprehension symmetry (simulating role alternation in language use) leads to larger populations creating more systematic protocols

In a typical communication game, the sender acts as a conditioned-generation model, taking a target input (for example, an image or a set of attribute values) and produces a message consisting of multiple symbols. The symbols of the message are generated one by one without any pre-defined vocabulary. The generated message is then transmitted to the receiver. The receiver is trained to infer the sender's input based on the message, by selecting the correct object among distractors or by fully reconstructing it.

Emergent communication models start with randomly-initialized parameters, without any pre-defined list of words or look-up table. Thus, the messages start out as random, and only over the course of training and interaction do the models develop a communication protocol. In fact, it is the central assumptions of *emergent* communication that the agents are not seeded with some initial language or communication protocol, but that they develop the communication system on their own during interaction. Thus, agents start from scratch and are guided primarily by communicative success. Yet, there is room for inductive biases, i. e., additional biases that are imposed on the learning system to promote desired behaviours (Mitchell, 1980). While cognitive biases in biological learning systems occur naturally, inductive biases in machine learning are artificially introduced to guide the learning dynamics. For a profound overview of the emergent communication literature, we refer to recent review and survey papers by Lazaridou and Baroni (2020), Galke et al. (2022), and Brandizzi (2023).

Notably, methods from the field of emergent communication and from the closely related field of reinforcement learning (see Kosoy et al., 2020; Kosoy et al., 2022, inter alia) have already been used for language development research (see Ohmer et al., 2020; Portelance et al., 2021, inter alia), for example, to study the emergence of a mutual exclusivity bias with pragmatic agents.

While emergent communication simulations hold a great potential for advancing our understanding of how languages emerge, we can only expect insights gained with deep neural networks to inform language evolution research if the resulting languages actually show the same properties as natural languages (Galke et al., 2022). Consequently, most emergent communication simulations try to compare the properties of their emerging communication protocols to the properties found in natural languages (Havrylov & Titov, 2017; Kottur et al., 2017; Lazaridou et al., 2017). By following this approach, the field has unveiled substantial differences between humans and machines in how they learn to communicate and what kinds of languages they develop.

Crucially, although the emergent languages of neural networks initially did not exhibit many of the linguistic properties typically associated with human languages, most of these differences could be reconciled by adding adequate inductive biases, such as laziness and impatience – which, when introduced, recovered the effects found in humans. Notably, some linguistic phenomena such as the word-order/case-marking trade-off seem to occur in communicating neural networks without specific inductive

biases (Lian et al., 2023). Below we review selected properties of human languages in which initial mismatches between humans and neural network agents were resolved and discuss the inductive biases that were necessary for their recovery. Table 1 provides an overview of the three phenomena and their occurrence in neural simulations.

Zipfian distribution in utterance length

Perhaps the most illustrative example of mismatches between the languages developed by humans and machines was the initial absence of Zipf's law of abbreviation in machine learning simulations. According to Zipf's law of abbreviation, the relationship between word frequency and word length follows a power law distribution, such that more frequent words are typically shorter while less frequent words are typically longer (Newman, 2005; Zipf, 1949). Zipf (1949) suggested that this effect is caused by the principle of least effort, i.e., since frequent words are produced often, and shorter words are easier to produce. Critically, Zipf's law has important implications for language evolution (Kanwal et al., 2017) and language acquisition (Ellis & Collins, 2009), with active restructuring of lexicon towards more efficient communication (Gibson et al., 2019).

Initial findings in emergent communication showed that Zipf's Law of Abbreviation is absent from the languages developed by neural agents, which was dubbed as 'anti-efficient coding' (Chaabouni et al., 2019). This was because neural senders were not under any pressure to communicate efficiently or to reduce effort. In fact, longer messages were easier for the receiver agent to process because they allowed for more opportunities to differentiate between meanings: for a 1-symbol utterance, the sender can select only 1 item from the alphabet of size k, but for a n-symbol utterance, the sender can produce k^n different combinations. The more distinct utterances are from another, the easier it is for the receiver to distinguish the target meaning from other possible meanings. Thus, longer utterances are advantageous for conveying the meaning correctly – especially when there is no penalty for utterance length.

The mismatch with human language was resolved by adjusting the optimization objective in a direction that made sender agents "lazy" (i.e., longer messages were penalized) and receiver agents "impatient" (i.e., receivers tried to infer the meaning as early as possible in a sequential read) (Rita et al., 2020). This inductive bias, which aims at mimicking real human behavior during language production and comprehension, has recovered Zipf's Law of Abbreviation in emergent communication simulations – showing that when such biases for efficiency are introduced, communication protocols developed by neural agents do show a similar frequency–length relationship as found in natural languages.

The emergence of compositional structure and its benefits for learning and generalization

Compositional structure is considered a hallmark feature of human language (Hockett, 1960; Szabó, 2022): there is a systematic mapping between linguistic forms (e.g., words, morphemes) and their meanings (e.g., concepts, grammatical categories), such that the meaning of a complex expression can be typically derived from the meanings its constituent parts. For example, the meaning of the phrase "small cats" is directly derived from the meanings of the words "small", "cat", and the marker "-s" (denoting plurality). The presence of such compositional structure underlies the infinite expressive and productive power of human languages, allowing us to describe new meanings in a way that is transparent and understandable to other speakers (Kirby, 2002; Zuidema, 2002).

In experiments simulating the evolution of languages in the lab using sender-receiver communication games, the need to communicate over a growing number of different items or in an open-ended meaning space leads to the emergence of compositional languages (Nölle et al., 2018; Raviv et al., 2019a). Crucially, the degree of compositional structure in linguistic input then predicts adults' learning and generalization accuracy, such that, compared to languages with little to no compositionality, languages with more compositional structure are learned better and faster and result in better (i.e., more transparent and systematic) generalizations to new meanings, which are also shared across different individuals who never interacted before (Raviv et al., 2021). Thus, the evolution of more compositional and systematic linguistic structure allows for more productive generalization and facilitates communication and convergence between strangers.

The learning advantage of more compositional structure for adult participants is also echoed in numerous iterated learning studies, which have shown that artificial languages become more compositional and consequently easier to learn over the course of cross-generational transmission (Beckner et al., 2017; Carr et al., 2017; Kirby et al., 2008; Kirby et al., 2014).

Testing the limits of our imagination, neural networks seemed to generalize well even without compositional communication protocols (Chaabouni et al., 2020; Lazaridou et al., 2018). Specifically, Chaabouni et al. (2020) found that, after many repetitions of an emergent communication experiment, all compositional languages generalized well, but so did non-compositional languages. This finding spurred numerous follow-up studies that aimed at improving the learning dynamics through inductive biases or by making the communication game more difficult (more complex stimuli, larger alphabet, longer messages, more agents) to successfully promote the emergence of compositional structure (Chaabouni et al., 2022; Rita, Tallec, et al., 2022). However, the lack of correlation between the degree of compositional structure – as measured by

topographic similarity (Brighton & Kirby, 2006) – and generalization performance had remained.

The most reliable way to promote the emergence of compositional languages is periodically resetting the parameters of the neural network agents (Chaabouni et al., 2022; F. Li & Bowling, 2019; Zhou et al., 2022), similar to Kirby et al. (2014)'s iterated learning paradigm – leading to the hypothesis that compositional languages have a learnability advantage (Chaabouni et al., 2020; Chaabouni et al., 2022; Guo et al., 2019; F. Li & Bowling, 2019). However, these attempts did not directly test language learnability in a purely supervised fashion.

Recently, Conklin and Smith (2022) have re-analyzed the setting of Chaabouni et al. (2020) and found that, in fact, the lack of correlation between compositionality and generalization performance in the original simulation was caused by a fallacy of the topographic similarity metric that had been used to measure compositionality. For instance, homonyms (different forms for same meaning) obscure compositionality under the topographic similarity measure. When taking this variation into account, compositional structure does reliably emerge and is beneficial for generalization. In other words, it is probably the case that there was not really a mismatch between humans and neural agents in the first place.

Supporting this view, Galke et al. (2023) have replicated a large-scale language learning study originally conducted with human participants (Raviv et al., 2021) with deep neural networks and have confirmed the advantage of compositional structure for learning and generalization in neural networks. The results showed similar pattern across three learning systems – humans, small-scale recurrent neural networks trained from scratch, and the large pre-trained language model GPT-3 – with compositional structure being advantageous for all types of learners. Specifically, the results showed that neural networks benefit from more structured linguistic input, and that their productions become increasingly more similar to human productions when trained on more structured languages. This structure bias can be found in the networks' learning trajectories and their generalization behavior, mimicking previous findings with humans: although all languages can eventually be learned, languages with a higher degree of compositional structure were led to better and more human-like generalization to new, unseen items.

Population size effects

Socio-demographic factors such as population size have long been assumed to be important determinants of language evolution and variation (Lupyan & Dale, 2010; Nettle, 2012; Wray & Grace, 2007). Supporting this idea, global cross-linguistic studies report that bigger communities tend to have languages with more regular and transparent structures (Lupyan & Dale, 2010). Similarly, in experimental work, larger groups of interacting

participants generally develop languages with more systematic (i.e., compositional) grammars (Raviv et al., 2019b). These findings are typically attributed to compressibility pressures arising during communication: remembering partner-specific variants becomes increasingly more challenging as group size increases and shared history decreases, which lead larger groups to prefer easier-to-learn-and-generalize variants and thus converge on more transparent and systematic languages.

Tieleman et al. (2019) has investigated populations of autoencoders. Autoencoders are neural network models composed of an encoder module and a decoder module that learn to "good" representations (the code) by reconstructing their own input. Now Tieleman et al. (2019) have decoupled encoder and decoders and exchanged them throughout training – while communicating in a continuous channel. There, larger communities produced representations with less idiosyncrasies and lead to better convergence among different agents. While a promising starting point, the communication was modeled as exchanging continuous vectors and training the encoder decoder modules together, as if they were one model. This is arguably natural communication paradigm for neural networks because it is optimized in the same way as the communication between layers in a single neural network. However, this continuous channel stands in contrast with the discrete nature of human communication (Hockett, 1960). Most other approaches in emergent communication, however, do consider a discrete channel (Galke et al., 2022).

While Chaabouni et al. (2022) argued that it is necessary to scale up emergent communication experiments in different aspects including population size in order to better align neural emergent communication with human language evolution, they have not found a consistent advantage of population size in generalization and ease-of-learning (in contrast with (Tieleman et al., 2019)). Similarly, Rita, Strub, et al. (2022) found that language properties are not enhanced by population size alone.

While emergent communication in populations of agents has been investigated earlier (Fitzgerald, 2019; Graesser et al., 2019; Lowe et al., 2019, e.g.), the effect of population size on structure with groups of more than two agents has only recently been analyzed (Chaabouni et al., 2022; Michel et al., 2023; Rita, Strub, et al., 2022). Out of these, two studies aimed to recover the group size effect in populations of neural network agents by introducing population heterogeneity (Rita, Strub, et al., 2022) and manipulating sender-receiver ties (Michel et al., 2023).The first study by Rita, Strub, et al. (2022) modeled population heterogeneity by giving each agent a different random learning rate While previous simulations used populations of identical agents, Rita et al. modeled population heterogeneity by giving each agent a different random learning rate. Results showed that in this scenario, group size effects could be partially recovered. Notably, the authors found that it is important to give sender agents having (much) higher learning rates than receivers.

Secondly, while most emergent communication simulations keep senders and receivers distinct (i.e., agents that produce never comprehend, and vice versa), there is also work that emphasizes linking production and comprehension components within the agents (e.g., by sharing some of the model parameters) (Graesser et al., 2019; Portelance et al., 2021). Galke et al. (2022) argue that this naturalistic property of alternating between sending and receiving (i.e., engaging in both production and comprehension in typical language use) may be a crucial ingredient to ensure more linguistically plausible learning dynamics – and could lead to recovering the group size effect. Subsequently, Michel et al. (2023) have introduced sender-receiver ties via gradient blocking, such that a sender and a receiver together form a single agent and each receiver is only optimized for its corresponding sender. This change indeed led to a recovery of the group size effect, with larger population of agents creating more compositional protocols. Another promising approach is to have agents model other agents' knowledge, allowing them to communicate differently with different agents - something that has been implied to underlie group size effects in humans (Lutzenberger et al., 2021; Meir et al., 2012; Mudd et al., 2020; Thompson et al., 2020). While such "theory of mind" is generally absent from emergent communication simulations in populations, the ability to infer other agents' beliefs has been successfully implemented in various reinforcement learning setups, e.g., (Filos et al., 2021; Ohmer et al., 2020).

Underlying learning pressures and inductive biases

In general, there are two types of learning biases and pressures. First, some biases and pressures seem to be present naturally, or universally, across all different learning systems investigated here, including deep learning agents. An example for this is the structure-bias, i.e., the learnability and generalization advantage of more compositional communication protocols (Galke et al., 2023) (see above). This structure-advantage seems to be present for both humans and neural networks, even without specific inductive biases. In contrast, some biases need to be artificially introduced in order to recover the effects found in humans. These include, for example, adding a length-penalty for senders, which effectively makes agents "lazy". In the above examples, we demonstrated the flexibility and adaptive nature of neural simulations and how they can be tweaked to replicate human behavioral patterns. While many features associated with natural languages were initially absent from such simulations, these mismatches have been fully or partially resolved by introducing theory-driven and human-inspired cognitive biases and learning pressures to the learning system – and these inductive biases have consequentially led to better alignment between neural agents and humans. Below, we outline on a more fine-grained level what pressures are relevant for language learning and evolution in neural networks, contrasting them with the pressures to which current large language models are exposed, and to what extent incorporating the pressures may promote the relevance of large language models for developmental research. Table 2 provides an overview of the comparison of learning pressures in emergent communication agents and large language models. Notably,

Table 2: Pressures derived from emergent communication simulations and their operationalization in neural agents and large language models

Derived Pressure	**Emergent Communication Agents**	**Large Language Models**
Pressure for successful communication	The main training objective in communication games	Absent in pre-training and fine-tuning. Only introduced when learning from human preferences in RLHF.
Pressure for learnability	Can be artificially introduced through parameter reset and iterated learning	Neural networks underlying large language models have a tendency to find the simplest solution first
Pressure to reduce production effort	Can be artificially introducing, e.g., through a penalty term for long messages	Production length is learned from LLM's training data and human feedback in RLHF.
Memory constraints	Absent because the high capacity of neural agents is sufficient to memorize even unstructured mappings	Huge capacity due to extremely high amount of parameters, yet "working memory" for in-context learning is limited by context window (how many tokens the models can process at a time)
Production-comprehension symmetry	Can be artificially introduced by linking sender and receiver modules	By design – LLMs employ the same neural network modules and parameters for comprehension and production
Modeling other agents' internal states	Can be modeled explicitly, e.g., for pragmatic reasoning	In the RLHF training stage, a reward model is trained and consulted to estimate human preferences.

this is not an exhaustive list – it focuses on the specific pressures that underlie the phenomena described above, but do not consider many other important aspects that govern natural language learning, such as grounding, a noisy environment, multi-modal communication, or referential and iconic signs.

Pressure for successful communication

In order to achieve successful communication, language users need distinguish between a variety of meanings. This expressivity pressure is hypothesized to underlie human language evolution, and serves as a "counter pressure" for simplicity/compressibility (i.e., the idea that languages should be as simple and as learnable as possible) (Kirby et al., 2015). The pressure for communicative success, e. g., to accurately reconstruct the meaning of referents from a message during interaction, is the most straight-forward pressure found in collaborative communication games (and, arguably, in real-world interaction). In emergent communication with deep neural networks, this pressure is encoded right in the optimization objective of the neural networks.

In contrast, for large language models such as GPT-3.5, the main objective during pre-training is not communication success. The standard language modeling objective used during pre-training of large language models instead optimizes for utterance completion (i. e., learning to predict words from their context). While this language modeling objective leads to tremendous success regarding language competence other emergent abilities (Devlin et al., 2019; Wei et al., 2022), it is clearly a different training objective than optimizing for communicative success, as in emergent communication simulations. After large-scale pre-training, large language models are fine-tuned using small datasets of human-generated pairs of instructions and their corresponding responses, usually with the same training objective as in pre-training. In other words, the models are made to learn from interactions by completing utterances from human-generated interactions, but not by interacting themselves. Only during the last stage of training, the models are trained via Reinforcement Learning from Human Feedback (RLHF), where a reward model estimates human preferences based on human ratings of different machine-generated responses (Ouyang et al., 2022; Schulman et al., 2017). Only in this final RLHF training stage of LLMs, the models are optimized for successful communication. Yet, this stage is important to turn base models into chat assistants that engage in conversations with humans (OpenAI, 2023; Ouyang et al., 2022).

In general, while emergent communication simulations are tuned for communicative success by design, this is in fact an extra step in large language models after pre-training on utterance completion. Thus, the learning paradigms of fine-tuning and subsequent learning from human feedback are worth further exploration for the goal of having language models being more representative of human behavior. For instance, a recent study has showcased that fine-tuning large language models on data from psychological tests turns them into useful cognitive models (Binz & Schulz, 2023).

Pressure to reduce production effort

Humans constantly strive to reduce effort during interaction (Gibson et al., 2019). For instance, this is demonstrated by our tendency to shorten or erode highly frequent words (Kanwal et al., 2017; Zipf, 1949). However, the pressure to communicate with least effort is absent in neural networks, and is usually not reflected in their training objective. In other words, it simply does not cost more "effort" for a neural network to generate a longer message. By introducing a bias for more efficient communication, Rita et al. (2020) have shown that typical human behavior can be recovered. Since language models similarly don't have an 'innate' pressure to reduce effort, it may be worth considering integrating such a pressure for efficient communication into these models for the sake of mimicking human behavior with respect to language development. However, one needs to strike a balance, as imposing a least-effort bias could also lead to communication failure in emergent communication scenarios (Lian et al., 2021), calling for further investigation of how a least-effort bias is best incorporated.

In large language models, there is no pressure to reduce production effort: LLMs are trained on next-token production over large corpora of text data, which is being piped through the model in a batched fashion to maximize throughput (see for instance Brown et al., 2020; Touvron et al., 2023, inter alia). Thus, the main driver for production length is simply the utterance length in data, and the placement of specific separator tokens, e.g., at the end of each unit of consecutive text during training. Moreover, the RLHF stage of training large language models (Ouyang et al., 2022; Schulman et al., 2017), which is supposed to align LLMs with human preferences, even promotes the generation of longer utterances, as they are deemed to be more "helpful" by (instructed) human annotators (Singhal et al., n.d.).

At inference time, when the LLM is prompted to generate text, a hard cut-off on the number of tokens or a soft length penalty may be introduced – the details of these techniques, however, are often not publicly available. Regardless, the training procedure itself does usually not include a length penalty, which needs to be taken into account when planning to use large language models for language development research.

Pressure for learnability

Based on our review, a pressure for learnability (or continual re-learning) also governs the development of communication protocols between neural network agents. That is, agents should prioritize communication protocols (or single variants) that are easier to learn, and such protocols should in turn boost performance. This learnability pressure is strongly connected to the fact that languages must be transmitted, learned, and used by multiple individuals, often from limited input and with limited exposure time (Smith et al., 2003). Yet, there is a subtle difference to strict transmission chains of iterated learning, as it is sufficient with neural networks to reset only some of the

agents (F. Li & Bowling, 2019), or only parts of a single agent (Zhou et al., 2022). In numerous different settings, it has been shown that learnability pressures are crucial for compositional structure to emerge (Chaabouni et al., 2022; F. Li & Bowling, 2019; Zhou et al., 2022).

This also suggests that under repeated learning, either in Iterated Learning with human participants or with parameter reset in neural networks, weak learning biases can get amplified in the process of cross-generational transmission (Reali & Griffiths, 2009). But what are these learning biases exactly? How can they be operationalized? And how do they actually translate into language learning in the real-world? For example, do these biases differ between children and adults, or between different levels of linguistic analyses (e.g., vocabulary vs. syntax)? At the moment, these are still open questions. However, they highlight the need to seriously consider the meaning and implications of different modeling choices when simulating language acquisition using language models and deep neural networks.

As for large language models, Chen et al. (2024) have made relevant findings by analyzing the learning dynamics: language models pick up grammar as the simplest explanation for the data very early on during training (structure onset), and only shortly thereafter, general linguistic capabilties arise. In addition, when suppressing grammar as a possible way to explain the data, the models learn other strategies, but do not go back to grammar when the constraint is removed later in training.

This finding connects well with more general findings of simplicity bias in neural networks (Geirhos et al., 2020). In addition, it also connects with the findings of emergent communication in emphasizing that re-learning (e. g., through parameter reset) is important for compositional structure to emerge (F. Li & Bowling, 2019). Our hypothesis is that, if there was no pressure for re-learning, then agents would fall for the earliest successful strategy and do not consider alternatives – stressing the importance of the learnability pressure.

Memory constraints

Human language learning is governed by cognitive constraints such as a limited memory capacity. These, in turn, affect processes of language evolution and promote greater convergence to a common language within a community: once groups become too big, it becomes hard to maintain unique communication protocols with different partners (i.e., idiolects) (Wray & Grace, 2007).

Such constraints have been shown to underlie patterns of cross-linguistic diversity, whereby larger populations develop more structured and less variable languages (Raviv et al., 2019b). Yet, neural networks have virtually no memory constraints because they are commonly heavily over-parametrized. Due to this over-parametrization, neural

networks have no problem to keep a large number of different partner-specific variants in their memory, and have little need to converge on a single shared language. However, simply reducing the number of model parameters to the theoretical minimum is not feasible either, as explored in emergent communication by Resnick et al. (2020). This is because over-parametrization is, in fact, a critical ingredient for the success of deep neural networks (Arora et al., 2019; Cybenko, 1989; Nakkiran et al., 2021; Zhong et al., 2017). But given the importance of such memory constraints for human language learning and evolution, it may be worth considering how such pressures can nonetheless be mimicked or introduced as inductive biases when employing deep neural networks as models for language development research.

While large language models have even higher model capacity with billions of learnable parameters, there is an interesing conceptual connection with working memory: As the model parameters are not updated at inference time (when the model is prompted with a specific input), the model can only base its generation on what is available in the prompt, which is limited by the LLMs' context window of how many tokens can be processed at a time. Although also these context windows grow larger and larger with the development of new models (OpenAI, 2023), it allows researchers to explicitly control what information is available to the model at a specific point in time.

Production-comprehension symmetry

In addition, in naturalistic settings with proficient language users, every person capable of producing a language is also capable of understanding it (Hockett, 1960) – a property that was typically absent from emergent communication simulations (Galke et al., 2022). Indeed, introducing an inherent connection between production and comprehension in neural networks has led to an increase in the desirable properties of emergent languages (Michel et al., 2023). Interestingly, comprehension and production are intrinsically linked in autoregressive large language models as the same model parameters are used for processing and for generation (Radford et al., 2019). Such results again underscore the importance of keeping seemingly basic psycholinguistic features in mind when using large language models and neural networks as models for human language learning and use.

Modeling other agents' internal states

Furthermore, another intriguing direction is to explicitly model other agents' internal states. For instance, Ohmer et al. (2020) integrates pragmatic reasoning into the agents, leading to accelerated learning – an effect that is even stronger with Zipfian input distributions compared to uniform input distributions. Explicitly modeling other agents internal states and social learning has been shown to be successful in other reinforcement learning scenarios, where agents can cooperate or compete about resources (Filos et al., 2021; Ndousse et al., 2021). Interestingly, these ideas of explicitly modeling the

internal state of the interlocutor are already present in the final training stage of large language models, when optimizing for human preferences via RLHF (Ouyang et al., 2022; Schulman et al., 2017): the common procedure is to learn a specific reward model that estimates human preferences on new data, which is then be employed for steering the generations of the language model in a particular direction – here the reward model is specifically designed to estimate to what extent humans would prefer one generation over the other, which is closely resembles the idea of modeling other agents' (or humans') internal states.

Discussion

Several important mismatches between humans and neural agents with respect to language emergence can be explained by the absence of key cognitive and communicative pressures, such as memory constraints and production-comprehension symmetry, which drive language evolution. Here, we demonstrated how including these factors in neural agents can resolve said mismatches, and lead to more accurate simulations that mimic the settings and pressures operating during human language learning and use – and consequentially resulting in emergent neural communication protocols that are more linguistically plausible. Notably, additional psycho- and sociolinguistic factors may affect language evolution and learning, and might also play a role in explaining further discrepancies in behavioral patterns across learning systems.

In the current paper we presented a number of initial mismatches between humans and agents engaging in communication games – and demonstrated how they could be resolved through inductive biases. So far, there is no unified approach that consolidates all of the resolutions mentioned above. We deem this a promising direction of future work – e. g., merging the techniques of population heterogeneity, laziness and impatience, and sender-receiver ties, which have so far only been evaluated independently.

As exemplified by recent work, it is promising to keep up and nourish the knowledge exchange between researchers working on human languages and those working on computational simulations of language, e. g., via theory diffusion from language studies into machine learning and vice versa. A famous example is cultural evolution (Tomasello, 2008) and the iterated learning paradigm (Kirby et al., 2008; Kirby et al., 2014), which sparked the idea of iteratively training neural networks while resetting some of the networks' parameters (Frankle & Carbin, 2018; F. Li & Bowling, 2019; Nikishin et al., 2022; Zhou et al., 2022). This idea has, for instance, advanced our understanding of neural networks (their reliance on sparse sub-networks) and led to favorable learning dynamics that cause better and more systematic generalization beyond the training distribution. Similarly, the discrete and compositional structure of natural languages inspired researchers to incorporate discrete representations into neural network architectures in order to advance the models' generalization performance and continual learning capabilities (Liu et al., 2021; Träuble et al., 2023).

In conclusion, The emergent communication literature provided the opportunity to assist in developing linguistic theories in the spirit of Elman (1993), while, conversely, reflecting on how phenomena and biases known from humans may ultimately enhance neural networks, as in lifelong and open-world learning, which is still a major open problem in machine learning. For making use of large language models in language development research, we consider it a promising direction for future work to take inspiration from the emergent communication literature, and see which inductive biases (such as the ones sketched here) have helped to recover patterns from human language learning. Concretely, this would entail ingesting a training objective for communicative success earlier in language model training, and integrating a pressure to keep utterances as short as possible. Integrating these biases into large language models may very well lead to more cognitively plausible models for gaining new insights on how children acquire their first language.

References

Abdou, M., Kulmizev, A., Hershcovich, D., Frank, S., Pavlick, E., & Søgaard, A. (2021). Can language models encode perceptual structure without grounding? a case study in color. *Proceedings of the 25th Conference on Computational Natural Language Learning*, 109–132. https://doi.org/10.18653/v1/2021.conll-1.9

Arora, S., Du, S., Hu, W., Li, Z., & Wang, R. (2019). Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks. *Proceedings of the 36th International Conference on Machine Learning*, 322–332.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. *Proceedings of ICLR*.

Baroni, M. (2021). On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. *arXiv preprint, abs/2106.08694*. https://arxiv.org/abs/2106.08694

Beckner, C., Pierrehumbert, J. B., & Hay, J. (2017). The emergence of linguistic structure in an online iterated learning task. *Journal of Language Evolution*, *2*(2), 160–176.

Binz, M., & Schulz, E. (2023). Turning large language models into cognitive models. *arXiv:2306.03917*.

Brandizzi, N. (2023). Towards More Human-like AI Communication: A Review of Emergent Communication Research. *arXiv:2308.02541*.

Brighton, H., & Kirby, S. (2006). Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artificial Life*, *12*(2), 229–242. https://doi.org/10.1162/artl.2006.12.2.229

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems 33*.

Carr, J. W., Smith, K., Cornish, H., & Kirby, S. (2017). The cultural evolution of structured languages in an open-ended, continuous world. *Cognitive science*, *41*(4), 892–923.

Chaabouni, R., Kharitonov, E., Bouchacourt, D., Dupoux, E., & Baroni, M. (2020). Compositionality and generalization in emergent languages. *ACL*, 4427–4442.

Chaabouni, R., Kharitonov, E., Dupoux, E., & Baroni, M. (2019). Anti-efficient encoding in emergent communication. *NeurIPS*, 6290–6300.

Chaabouni, R., Strub, F., Altché, F., Tarassov, E., Tallec, C., Davoodi, E., Mathewson, K. W., Tieleman, O., Lazaridou, A., & Piot, B. (2022). Emergent communication at scale. *ICLR*. https://openreview.net/forum?id=AUGBfDIV9rL

Chater, N., & Christiansen, M. H. (2010). Language Acquisition Meets Language Evolution. *Cognitive Science*, *34*(7), 1131–1157. https://doi.org/10.1111/j.1551-6709.2009.01049.x

Chen, A., Shwartz-Ziv, R., Cho, K., Leavitt, M. L., & Saphra, N. (2024). Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=MO5PiKHELW

Conklin, H., & Smith, K. (2022). Compositionality with Variation Reliably Emerges in Neural Networks. *The Eleventh International Conference on Learning Representations*.

Contreras Kallens, P., Kristensen-McLachlan, R. D., & Christiansen, M. H. (2023). Large Language Models Demonstrate the Potential of Statistical Learning in Language. *Cognitive Science*, *47*(3), e13256. https://doi.org/10.1111/cogs.13256

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Math. Control. Signals Syst.*, *2*(4), 303–314. https://doi.org/10.1007/BF02551274

Dale, R., & Lupyan, G. (2012). Understanding the origins of morphological diversity: The linguistic niche hypothesis. *Advances in complex systems, 15*(03n04), 1150017.

Dasgupta, I., Lampinen, A. K., Chan, S. C. Y., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2022). Language models show human-like content effects on reasoning. *arXiv:2207.07051*.

De Seyssel, M., Lavechin, M., & Dupoux, E. (2023). Realistic and broad-scope learning simulations: First results and challenges. *Journal of Child Language*, 1–24. https://doi.org/10.1017/S0305000923000272

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*. https://doi.org/10.1038/s44159-023-00241-5

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dupoux, E. (2018). Cognitive Science in the era of Artificial Intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition, 173*, 43–59. https://doi.org/10.1016/j.cognition.2017.11.008

Ellis, N., & Collins, L. (2009). Input and Second Language Acquisition: The Roles of Frequency, Form, and Function Introduction to the Special Issue. *The Modern Language Journal, 93*(3), 329–335. https://doi.org/10.1111/j.1540-4781.2009.00893.x

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition, 48*(1), 71–99. https://doi.org/10.1016/0010-0277(93)90058-4

Filos, A., Lyle, C., Gal, Y., Levine, S., Jaques, N., & Farquhar, G. (2021). Psiphi-learning: Reinforcement learning with demonstrations using successor features and inverse temporal difference learning. *International Conference on Machine Learning*, 3305–3317.

Fitzgerald, N. (2019). To populate is to regulate. *EmeCom workshop at NeurIPS*.

Foerster, J., Assael, I. A., de Freitas, N., & Whiteson, S. (2016). Learning to Communicate with Deep Multi-Agent Reinforcement Learning. *Advances in Neural Information Processing Systems, 29*.

Frankle, J., & Carbin, M. (2018). The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. *International Conference on Learning Representations*.

Galke, L., Ram, Y., & Raviv, L. (2022). Emergent communication for understanding human language evolution: What's missing? *Emergent Communication Workshop at ICLR 2022*. https://openreview.net/forum?id=rqUGZQ-0XZ5

Galke, L., Ram, Y., & Raviv, L. (2023). What makes a language easy to deep-learn? *arXiv:2302.12239*.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence, 2*(11), 665–673.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How Efficiency Shapes Human Language. *Trends in Cognitive Sciences, 23*(5), 389–407. https://doi.org/10.1016/j.tics.2019.02.003

Gong, T., Minett, J. W., & Wang, W. S.-Y. (2008). Exploring social structure effect on language evolution based on a computational model. *Connection Science, 20*(2-3), 135–153.

Goth, G. (2016). Deep or shallow, NLP is breaking out. *Communications of the ACM, 59*(3), 13–16. https://doi.org/10.1145/2874915

Graesser, L., Cho, K., & Kiela, D. (2019). Emergent linguistic phenomena in multi-agent communication games. *EMNLP/IJCNLP (1)*, 3698–3708.

Guo, S., Ren, Y., Havrylov, S., Frank, S., Titov, I., & Smith, K. (2019). The emergence of compositional languages for numeric concepts through iterated learning in neural agents. *arXiv preprint, abs/1910.05291*.

Harris, Z. S. (1954). Distributional Structure. *WORD, 10*(2-3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Havrylov, S., & Titov, I. (2017). Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. *NeurIPS*, 2149–2159.

Hockett, C. F. (1960). The origin of speech. *Scientific American, 203*(3), 88–97.

Huebner, P. A., Sulem, E., Cynthia, F., & Roth, D. (2021). BabyBERTa: Learning more grammar with small-scale child-directed language. *Proceedings of the 25th Conference on*

Computational Natural Language Learning, 624–646. https://doi.org/10.18653/v1/2021.conll-1.49

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, *165*, 45–52.

Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax.

Kirby, S. (2017). Culture and biology in the origins of linguistic structure. *Psychonomic bulletin & review*, *24*(1), 118–137.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current opinion in neurobiology*, *28*, 108–114.

Kirby, S., Smith, K., & Brighton, H. (2004). From ug to universals: Linguistic adaptation through iterated learning. *Studies in Language*, *28*(3), 587–607.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Kosoy, E., Collins, J., Chan, D. M., Huang, S., Pathak, D., Agrawal, P., Canny, J., Gopnik, A., & Hamrick, J. B. (2020). Exploring exploration: Comparing children with rl agents in unified environments. *Bridging AI and Cognitive Science workshop at ICLR*.

Kosoy, E., Liu, A., Collins, J. L., Chan, D., Hamrick, J. B., Ke, N. R., Huang, S., Kaufmann, B., Canny, J., & Gopnik, A. (2022). Learning causal overhypotheses through exploration in children and computational models. In B. Schölkopf, C. Uhler, & K. Zhang (Eds.), *Proceedings of the first conference on causal learning and reasoning* (pp. 390–406). PMLR. https://proceedings.mlr.press/v177/kosoy22a.html

Kottur, S., Moura, J., Lee, S., & Batra, D. (2017). Natural language does not emerge 'naturally' in multi-agent dialog. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2962–2967. https://doi.org/10.18653/v1/D17-1321

Lazaridou, A., & Baroni, M. (2020). Emergent multi-agent communication in the deep learning era. *arXiv preprint, abs/2006.02419*.

Lazaridou, A., Hermann, K. M., Tuyls, K., & Clark, S. (2018). Emergence of linguistic communication from referential games with symbolic and pixel input. *ICLR*.

Lazaridou, A., Peysakhovich, A., & Baroni, M. (2017). Multi-agent cooperation and the emergence of (natural) language. *ICLR*.

Li, B. Z., Nye, M., & Andreas, J. (2021). Implicit representations of meaning in neural language models. *Proc. of ACL*, 1813–1827. https://doi.org/10.18653/v1/2021.acl-long.143

Li, F., & Bowling, M. (2019). Ease-of-teaching and language structure from emergent communication. *NeurIPS*, 15825–15835.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2023). Emergent world representations: Exploring a sequence model trained on a synthetic task. *Proc. of ICLR*. https://openreview.net/forum?id=DeG07_TcZvT

Lian, Y., Bisazza, A., & Verhoef, T. (2021). The Effect of Efficient Messaging and Input Variability on Neural-Agent Iterated Language Learning. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10121–10129. https://doi.org/10.18653/v1/2021.emnlp-main.794

Lian, Y., Bisazza, A., & Verhoef, T. (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.

Liu, D., Lamb, A. M., Kawaguchi, K., ALIAS PARTH GOYAL, A. G., Sun, C., Mozer, M. C., & Bengio, Y. (2021). Discrete-Valued Neural Communication. *Advances in Neural Information Processing Systems, 34*, 2109–2121.

Lowe, R., Gupta, A., Foerster, J., Kiela, D., & Pineau, J. (2019). Learning to learn to communicate. *Proceedings of the 1st Adaptive & Multitask Learning Workshop*.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PloS one, 5*(1), e8559.

Lutzenberger, H., De Vos, C., Crasborn, O., & Fikkert, P. (2021). Formal variation in the kata kolok lexicon. *Glossa: a journal of general linguistics, 6*.

Meir, I., Israel, A., Sandler, W., Padden, C. A., & Aronoff, M. (2012). The influence of community on language structure: Evidence from two young sign languages. *Linguistic Variation, 12*(2), 247–291.

Michel, P., Rita, M., Mathewson, K. W., Tieleman, O., & Lazaridou, A. (2023). Revisiting Populations in multi-agent Communication. *The Eleventh International Conference on Learning Representations*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems 26*, 3111–3119.

Mitchell, T. M. (1980). The Need for Biases in Learning Generalizations.

Mudd, K., De Vos, C., & De Boer, B. (2020). An agent-based model of sign language persistence informed by real-world data. *Language Dynamics and Change*, *10*(2), 158–187.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., & Sutskever, I. (2021). Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, *2021*(12), 124003.

Ndousse, K. K., Eck, D., Levine, S., & Jaques, N. (2021). Emergent social learning via multi-agent reinforcement learning. *International conference on machine learning*, 7991–8004.

Nettle, D. (2012). Social scale and structural complexity in human languages. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1597), 1829–1836.

Newman, M. E. (2005). Power laws, pareto distributions and zipf's law. *Contemporary physics*, *46*(5), 323–351.

Nikishin, E., Schwarzer, M., D'Oro, P., Bacon, P.-L., & Courville, A. (2022). The Primacy Bias in Deep Reinforcement Learning. *Proceedings of the 39th International Conference on Machine Learning*, 16828–16847.

Nölle, J., Staib, M., Fusaroli, R., & Tylén, K. (2018). The emergence of systematicity: How environmental and communicative factors shape a novel communication system. *Cognition*, *181*, 93–104. https://doi.org/10.1016/j.cognition.2018.08.014

Ohmer, X., König, P., & Franke, M. (2020). Reinforcement of semantic representations in pragmatic agents leads to the emergence of a mutual exclusivity bias. *CogSci*.

OpenAI. (2023). GPT-4 Technical Report. *arXiv:2303.08774*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022).

Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Patel, R., & Pavlick, E. (2022). Mapping language models to grounded conceptual spaces. *Proc. of ICLR*. https://openreview.net/forum?id=gJcEM8sxHK

Perfors, A., & Navarro, D. J. (2014). Language evolution can be shaped by the structure of the world. *Cognitive science*, *38*(4), 775–793.

Portelance, E., Frank, M. C., Jurafsky, D., Sordoni, A., & Laroche, R. (2021). The emergence of the shape bias results from communicative efficiency. *Proceedings of the 25th Conference on Computational Natural Language Learning*, 607–623.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*, 140:1–140:67.

Raviv, L., de Heer Kloots, M., & Meyer, A. (2021). What makes a language easy to learn? a preregistered study on how systematic structure and community size affect language learnability. *Cognition*, *210*, 104620.

Raviv, L., Meyer, A., & Lev-Ari, S. (2019a). Compositional structure can emerge without generational transmission. *Cognition*, *182*, 151–164.

Raviv, L., Meyer, A., & Lev-Ari, S. (2019b). Larger communities create more systematic languages. *Proceedings of the Royal Society B*, *286*(1907), 20191262.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*(3), 317–328. https://doi.org/10.1016/j.cognition.2009.02.012

Resnick, C., Gupta, A., Foerster, J. N., Dai, A. M., & Cho, K. (2020). Capacity, bandwidth, and compositionality in emergent language learning. *AAMAS*, 1125–1133.

Rita, M., Chaabouni, R., & Dupoux, E. (2020). "lazimpa": Lazy and impatient neural agents learn to communicate efficiently. *CoNLL*, 335–343.

Rita, M., Strub, F., Grill, J.-B., Pietquin, O., & Dupoux, E. (2022). On the role of population heterogeneity in emergent communication. *ICLR*. https://openreview.net/forum?id=5Qkd7-bZfI

Rita, M., Tallec, C., Michel, P., Grill, J.-B., Pietquin, O., Dupoux, E., & Strub, F. (2022). Emergent Communication: Generalization and Overfitting in Lewis Games. *Advances in Neural Information Processing Systems*.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences, 118*(45), e2105646118. https://doi.org/10.1073/pnas.2105646118

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Selten, R., & Warglien, M. (2007). The emergence of simple languages in an experimental coordination game. *Proceedings of the National Academy of Sciences, 104*(18), 7361–7366.

Singhal, P., Goyal, T., Xu, J., & Durrett, G. (n.d.). A long way to go: Investigating length correlations in RLHF [to appear in the Conference on Language Modeling 2024]. *arXiv:2310.03716*.

Smith, K. (2022). How language learning and language use create linguistic structure. *Current Directions in Psychological Science, 31*(2), 177–186.

Smith, K., Brighton, H., & Kirby, S. (2003). Complex systems in language evolution: The cultural emergence of compositional structure. *Advances in complex systems, 6*(04), 537–558.

Smith, K., & Kirby, S. (2008). Cultural evolution: Implications for understanding the human language faculty and its evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1509), 3591–3603.

Srikant, S., Lipkin, B., Ivanova, A. A., Fedorenko, E., & O'Reilly, U.-M. (2022). Convergent representations of computer programs in human and artificial neural networks. *Advances in Neural Information Processing Systems*.

Steels, L. (2016). Agent-based models for the emergence and evolution of grammar. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1701), 20150447.

Szabó, Z. G. (2022). Compositionality. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2022). Metaphysics Research Lab, Stanford University.

Thompson, B., Raviv, L., & Kirby, S. (2020). Complexity can be maintained in small populations: A model of lexical variability in emerging sign languages.

Tieleman, O., Lazaridou, A., Mourad, S., Blundell, C., & Precup, D. (2019). Shaping representations through communication: Community size effect in artificial learning systems. *arXiv:1912.06208*.

Tomasello, M. (2008). *Origins of Human Communication*. MIT Press.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Träuble, F., Goyal, A., Rahaman, N., Mozer, M. C., Kawaguchi, K., Bengio, Y., & Schölkopf, B. (2023). Discrete Key-Value Bottleneck. *Proceedings of the 40th International Conference on Machine Learning*, 34431–34455.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems 30*, 6000–6010.

Warstadt, A., & Bowman, S. R. (2022). What Artificial Neural Networks Can Tell Us About Human Language Acquisition. *arXiv:2208.07998*.

Warstadt, A., Choshen, L., Mueller, A., Williams, A., Wilcox, E., & Zhuang, C. (2023). Call for papers – the BabyLM challenge: Sample-efficient pretraining on a developmentally plausible corpus. *arXiv:2301.11796*.

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 1–16. https://doi.org/10.1038/s41562-023-01659-w

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models. *Transactions of Machine Learning Research, 2022*. https://openreview.net/forum?id=yzkSU5zdwD

Winters, J., Kirby, S., & Smith, K. (2015). Languages adapt to their contextual niche. *Language and Cognition, 7*(3), 415–449.

Wray, A., & Grace, G. W. (2007). The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua, 117*(3), 543–578.

Zeigler, B. P., Praehofer, H., & Kim, T. G. (2000). *Theory of modeling and simulation*. Academic press.

Zhong, K., Song, Z., Jain, P., Bartlett, P. L., & Dhillon, I. S. (2017). Recovery Guarantees for One-hidden-layer Neural Networks. *Proceedings of the 34th International Conference on Machine Learning*, 4140–4149.

Zhou, H., Vani, A., Larochelle, H., & Courville, A. (2022). Fortuitous Forgetting in Connectionist Networks. *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*.

Zuidema, W. (2002). How the poverty of the stimulus solves the poverty of the stimulus. *Advances in neural information processing systems*, *15*.

Data, Code and Materials Availability Statement

This review paper does not introduce any new data, code, or materials.

Authorship and Contributorship Statement

LG conceptualized the idea, reviewed the literature and wrote the paper. LR conceptualized the idea and helped write the paper.

Acknowledgements

We thank Mitja Nikolaus and Mathieu Rita for insightful comments and discussions. We thank Eva Portelance and Michael C Frank for their valuable comments on an initial version of the manuscript.

License