Deep Learning Agents and the Emergence of Compositional Languages: Approaches, Inductive Biases and Measurement

NICHOLAS BAILEY*, City St George's, University of London, United Kingdom
CHRIS CHILD, City St George's, University of London, United Kingdom
TILLMAN WEYDE, City St George's, University of London, United Kingdom

Background: Compositional symbol-forming and symbol-relating behaviors in deep learning or neuro-symbolic systems have been repeatedly recommended as part of a solution to the shortcomings of current state-of-the-art artificial intelligence. Studying how compositional languages can emerge between *tabula rasa* deep learning agents may help us understand how to make artificial neural networks represent unstructured, continuous input data in terms of combinations of discrete symbols.
Objectives: We aim present a comprehensive overview of recent research into compositional languages emerging between deep learning agents, in a manner that will be accessible to machine learning researchers who are not already aware of emergent communication and emergent languages.
Methods: We review roughly ten years of emergent language research, particularly focusing on contributions after 2019 that pertain to measuring or eliciting compositionality in emergent languages.
Results: Systematic generalization and topographic similarity (topsim) are the most dominant measures of compositionality in recent literature. "Productivity pressure", forcing agents to use vocabularies smaller than the number of meanings they need to communicate, is clearly necessary for compositionality to emerge. Regularizing or periodically resetting receiver/listener agents is an effective way of encouraging more compositional languages, perhaps because it creates a pressure for the speaker to create languages that can be learned more efficiently. The relative benefits of various neural network architectures, particularly the Transformer architecture dominant in other areas of deep learning, remains an underexplored topic. As other authors have noted, the field relies heavily on small-scale models and simple, often symbolic environments, which may hinder the generality of current conclusions.
Conclusions: Emergent language research provides a testbed for encouraging emergent compositionality in deep learning models, which may in future contribute to the development of safer, more interpretable, and more sample-efficient neuro-symbolic foundational models. We advocate that future research converges on topsim and generalization as the standard approaches to measuring compositionality, but also works to expand topsim into a family of metrics that can detect compositionality in languages displaying forms of linguistic variation such as free word order and synonymy. We also call upon researchers to test promising techniques at larger scales, with a greater range of agent architectures, and with more complex, multi-modal referents.

JAIR Track: Surveys

JAIR Associate Editor: Myrthe Tielman

*Corresponding Author.

Authors' Contact Information: Nicholas Bailey, ORCID: 0000-0001-9211-2197, nicholas.bailey@citystgeorges.ac.uk, City St George's, University of London, London, United Kingdom; Chris Child, ORCID: 0000-0001-5425-2308, c.child@citystgeorges.ac.uk, City St George's, University of London, London, United Kingdom; Tillman Weyde, ORCID: 0000-0001-8028-9905, t.e.weyde@citystgeorges.ac.uk, City St George's, University of London, London, United Kingdom.

DOI: 10.1613/jair.1.17302

JAIR Reference Format:
Nicholas Bailey, Chris Child, and Tillman Weyde. 2026. Deep Learning Agents and the Emergence of Compositional Languages: Approaches, Inductive Biases and Measurement. *Journal of Artificial Intelligence Research* 86, Article 6 (May 2026), 43 pages. DOI: 10.1613/jair.1.17302

1 Introduction

Emergent language research involves creating multi-agent environments in which *tabula rasa* (i.e. not pretrained) deep learning agents are stimulated to invent a language in order to communicate with one another and succeed at a task. The agents are given the ability to pass symbols or sequences of symbols to each other during the task and, although these symbols have no prescribed meaning at the beginning of the experiment, the agents come to agree useful meanings as the experiment continues. These setups are called "emergent language" or "emergent communication" experiments (Lazaridou and Baroni 2020; Wagner et al. 2003). Emergent language games are comparable to grounded language learning, in which artificial agents try to learn a natural language (a language that developed naturally in human society, e.g. English) through using the language in an environment that gives experiential meaning to the words they are learning (Suglia et al. 2024). They are also comparable to approaches that involve teaching agents a predefined language and then finetuning the agents in a simulated environment to study how languages can evolve through use (Lian, Bisazza, et al. 2023; Lian, Verhoef, et al. 2024). The key distinction is that in emergent language games, the language learned by the agents has never existed before. Its meaning is grounded in the environment the agents inhabit and its structure and usage can change over time like a natural language (Brighton 2002; Kirby 2001). However, in a profound sense the language that emerges is non-human and may not follow the familiar rules of languages that emerge among humans (Kottur et al. 2017). For example, without careful experimental design an emergent language may not contain "words" in the familiar sense, or be partitioned into sentences.

This survey paper examines work in emergent languages among deep learning agents, with a particular focus on cases where these languages display compositionality, which is often colloquially defined as the property that "The meaning of a sentence is determined by the meaning of its meaningful components, plus their mode of composition" (Haugeland 1979). In other words, compositional languages have words or symbols with inherent meaning, and these symbols (meanings) can be systematically combined to produce composite meanings that would be understood by a language user who understood the meanings of the words and the rules for how they could be combined. All natural languages have this property. There is a conceptual outline of emergent language research in Section 2.1 and the challenge of defining and measuring compositionality is addressed in depth in sections 3 and 4. The time span of our survey is roughly the 10 years from the earliest uses of deep learning in emergent language research (Foerster et al. 2016) to the time of writing, though our aim is not to provide a comprehensive review of all works in this period. Instead, we focus particularly on those that make interesting contributions to the encouragement or measurement of compositionality in emergent languages, and these are mostly from 2019 onward.

Emergent language research originated as a computational method in evolutionary linguistics (Wagner et al. 2003) and indeed the field of evolutionary linguistics has contributed fertile ideas about how compositional languages might be made to emerge among artificial agents (see Section 5.4.1). Beyond that however, emergent language research is a promising source of ideas about how to help artificial neural networks (ANNs) create and manipulate symbol-like concepts and think in a more human-like way. Greff et al. (2020) dub this "the binding problem in artificial neural networks" and it is discussed in Section 2.2. Compositionality in emergent languages, in which meaningful symbols can be arranged in multiple ways to achieve useful composite meanings (though see Section 3 for further definitions) is by definition supported by binding certain atomic concepts, grounded in the environment of the agents creating the language, to certain symbols. By experimenting with the environment where these languages emerge, researchers hope to uncover the necessary or sufficient pressures that encourage

these compositional behaviors (see Section 5 for a review of what these pressures might be and how to apply them).

Although the most recent large language models (LLMs) can generate outputs that appear to exhibit symbolic reasoning and human-like common sense (Bubeck et al. 2023), they must necessarily be pre-trained on enormous amounts of pre-existing compositional language, so are not able to help us identify reasons why neural networks might take a continuous input (e.g. vision, sound) and learn to represent it in a compositional way. Additionally, today's largest AI models have some shortcomings that are inherent to language modeling as a machine learning paradigm. For example, language modeling does not put models under pressure to generate truthful text (only unsurprising text in an information-theoretic sense). In Section 2.3, we suggest that although emergent language approaches do not offer solutions to all of the problems faced by LLMs, they offer strengths that recommend them as a complementary tool and a rewarding area of research into safe and explainable AI and human-computer interaction.

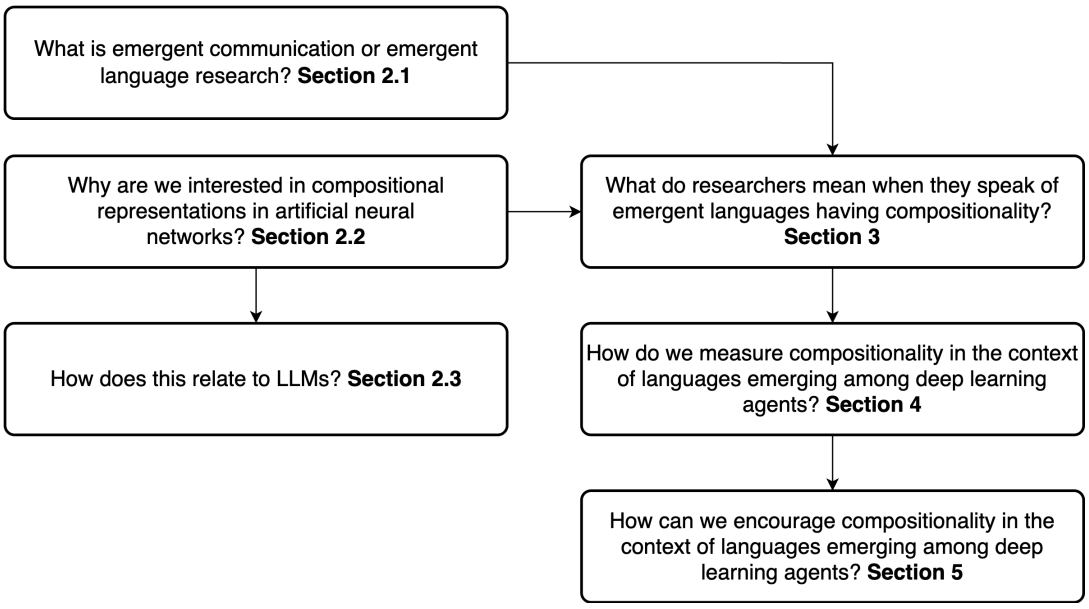

Fig. 1. A schematic of the main sections of this survey, in terms of the main question we attempt to address in each section.

The main parts of this survey, sections 3 to 5, present an overview of the ways in which compositionality in emergent languages has been defined, measured and encouraged to emerge. Section 3 outlines the breadth of definitions for compositionality used in the literature. Section 4 goes on to review the ways compositionality in emergent languages has been measured. Finally, after we have set out the different ways of defining and measuring compositionality, Section 5 reviews the inductive biases that have been shown to affect the degree of compositionality found in languages emerging between deep learning agents. Each section is arranged into subsections to highlight common themes.

We conclude this survey in Section 6, noting several seemingly effective and well-explored methods in the research to date, as well as some outstanding questions. Despite progress in understanding the conditions in which compositional languages are likely to occur, we call upon researchers to explore a wider range of methods and to test promising techniques at larger scales.

2 Context of Compositionality in Emergent Languages

As this survey is intended to be useful to researchers of deep learning more broadly, as well as those already working within the field of emergent languages, we begin our survey with some contextual information that explains the overall idea of emergent languages and the motivation for studying them. This is arranged into three sections. Section 2.1 introduces key ideas and terminology that the reader will need to be familiar with in order to understand the rest of the survey. Section 2.2 gives context around the binding problem in neuroscience and in deep learning research, which relates to forming representations of an environment in terms of nameable entities and relationships. The ability to do this is the foundation of symbolic reasoning, so having agents create languages that refer to things in their environment is an interesting route towards progress in creating reliable and explainable neuro-symbolic AI. Section 2.3 aims to address the question, "If LLMs can already master compositional languages, why are compositional emergent languages of interest?"

2.1 Emergent Language Concepts and Terminology

This section explains some terminology related to emergent language research, which will be used throughout this survey and the works it examines. Readers unfamiliar with emergent language/communication research may also wish to review other surveys of the field such as Lazaridou and Baroni (2020), Brandizzi (2023), Peters et al. (2025) and Zhu et al. (2024). The terms "emergent language research/experiment" and "emergent communication research/experiment" are used interchangeably in this survey.

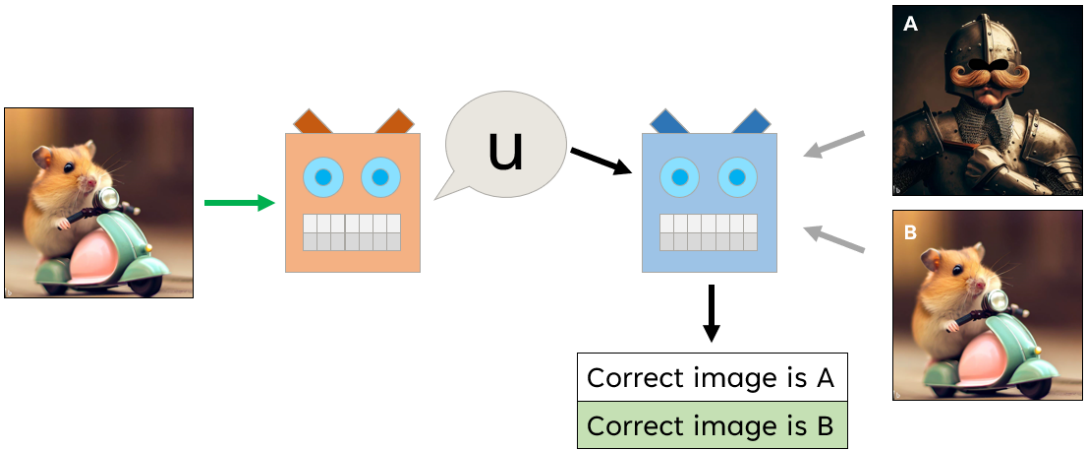

Fig. 2. A simple diagram representing a referential game, taking inspiration from the diagrams in Lazaridou and Baroni (2020). A sending agent (left) must communicate to a receiving agent (right) about a referent, using a message u. The receiving agent must correctly identify the referent from a set of candidates containing distractors.

By far the most common type of task used in emergent language research is a Lewis signaling game, originally described by Lewis (1969). In this setup, an agent must successfully communicate to a second agent by passing the latter one or more messages. A message in this case is a sequence of one or more vectors, usually one-hot, to simulate the discrete nature of words or phonemes in a natural language, where each item in the sequence may be referred to as a word, symbol, or token. There are two main Lewis games: the referential game and the reconstruction game. These may be referred to in the literature as Lewis games, or simply as "a referential game" or "a reconstruction game". Each may also be referred to by other names from time to time. For example, referential games are also referred to by works included in this survey as "a discrimination game" (Chaabouni,

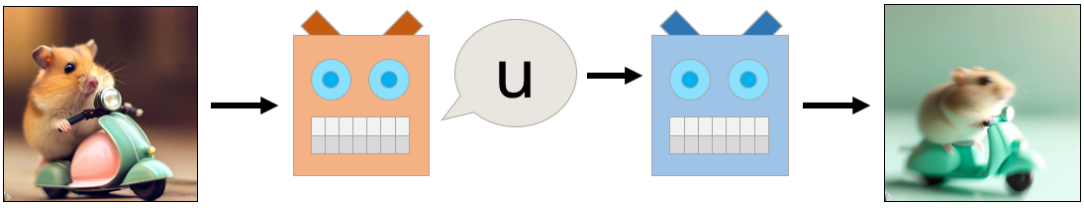

Fig. 3. A simple diagram representing a reconstruction game, taking inspiration from the diagrams in Lazaridou and Baroni (2020). A sending agent (left) must communicate to a receiving agent (right) about a referent, using a message u. The receiving agent must reconstruct the referent as accurately as possible.

Strub, et al. 2021) and "an object naming game" (Korbak, Zubek, Kuciński, et al. 2021). In both referential and reconstruction games, a "sender" (or "speaker") agent is shown a referent (a thing to which it must refer) and must pass a message (sometimes "utterance") to a "receiver" (or "listener") agent, describing the referent. Sometimes (Cogswell et al. 2019; Kottur et al. 2017, *inter alia*) agents will change roles so that each agent is at some points the receiver and at others the sender, and in these cases interactions may be a multi-step exchange. In this survey we will mostly use the terms "sender", "receiver", "message" and "symbol", as these are more general and carry fewer connotations than for example the speech-based words "speaker", "listener", "utterance" and "word". Indeed, messages sent between agents need not be sequences of symbols as in verbal conversation, e.g. Mihai and Hare (2021) have the sender agent "sketch" a message for the the receiver using a differentiable drawing procedure. The obvious exception is where we are explicitly speaking about messages exchanged in natural languages, in which case we may naturally refer to speakers, listeners and utterances. Somewhat contrary to this, the communication protocols emerging between agents in emergent language/communication experiments will tend to be referred to here as "emergent languages" rather than the common but longer "emergent communication protocols".

In a referential game (fig. 2) the receiver agent is shown the referent, often alongside other false candidates ("distractors"), and must correctly identify the true referent based on the received message. Figure 2 represents the variant of the referential game in which the sender is only shown the referent and the receiver is shown the referent and distractors. In the diagram, only one distractor is shown, but in practice any number between zero (e.g. Choi et al. 2018) and thousands (Chaabouni, Strub, et al. 2021) can be used successfully. In the case with no distractors, the receiver is not always shown the true referent and must classify whether it has been shown the true referent or a distractor (see e.g. Choi et al. 2018). In a reconstruction game (fig. 3) the receiver must generate an approximation of the referent based on the message it receives and so the sender and receiver are respectively like the encoder and decoder parts of a standard autoencoder, except that the encoding is a sequence of symbols (Devaraj et al. 2020; Resnick et al. 2020). Figures 2 and 3 respectively illustrate the referential and reconstruction games, where referents are images[1]. Images are an example of "sub-symbolic" or "entangled" input, where the factors of variation in the input are implicit and must be learned by the agents. Alternatively, emergent language experiments may use "symbolic" or "disentangled" inputs, represented by vectors where each element of the vector has a self-contained meaning. Lowe, Foerster, et al. (2019) provide an overview of some alternative referential game setups.

The set of all possible referents that agents might encounter may be referred to as their "environment", and we say that a language whose messages denote and connote concepts evident in an environment (e.g. features shared by multiple referents) is "grounded in" that environment. The intuition behind the value of compositional languages is that, even for complex (even "real world") environments, there will be a relatively small number of

[1]Images for Figures 2 and 3 were generated by DALL-E 2 (Ramesh et al. 2022) via Bing chat. Prompts: "A hamster on a Vespa", "Lightly blurred image of a hamster on a mint green Vespa, riding from left to right", "A knight with a marvelous mustache".

Table 1. Table including the main studies reviewed in this survey. Shows referent type (Symbolic, Sub-symbolic, or Mixed if both were tried), Gradient Estimation, i.e. the method for calculating the gradient for symbol choices within messages, and type of game used in the study. "Multi-step" refers to game variants where agents were permitted to exchange multiple messages (each acting alternately as the sender and the receiver).

Citation	Referents	Gradient Estimation	Game Type
Kottur et al. (2017)	Symbolic	REINFORCE	Multi-step referential
Mordatch and Abbeel (2017)	Symbolic	STGS	Interactive environment
Bogin et al. (2018)	Mixed	Multiple	Interactive environment
Choi et al. (2018)	Sub-Symbolic	Obverter	Referential
Cogswell et al. (2019)	Symbolic	REINFORCE	Multi-step referential
S. Guo (2019)	Mixed	STGS	Referential
F. Li and Bowling (2019)	Symbolic	REINFORCE	Referential
Yi et al. (2019)	Symbolic	REINFORCE	Referential
Chaabouni, Kharitonov, Bouchacourt, et al. (2020)	Symbolic	Hybrid	Reconstruction
Dagan et al. (2020)	Symbolic	STGS	Referential
S. Guo et al. (2020)	Symbolic	STGS	Referential, reconstruction
Gupta et al. (2020)	Symbolic	STGS	Referential
Kuciński, Kołodziej, et al. (2020)	Sub-Symbolic	STGS	Reconstruction
Luna et al. (2020)	Sub-Symbolic	STGS	Referential
Resnick et al. (2020)	Symbolic	STGS	Reconstruction
Slowik et al. (2020)	Symbolic	STGS	Referential
Steinert-Threlkeld (2020)	Symbolic	REINFORCE	Referential
Chaabouni, Strub, et al. (2021)	Sub-Symbolic	Hybrid	Referential
Korbak, Zubek, Kuciński, et al. (2021)	Sub-Symbolic	STGS	Referential
Kuciński, Korbak, et al. (2021)	Sub-Symbolic	STGS	Reconstruction
Mu and Goodman (2021)	Sub-Symbolic	STGS	Referential
Perkins (2021b)	Mixed	Multiple	Referential
Perkins (2021c)	Sub-Symbolic	STGS	Referential
Garcia et al. (2022)	Sub-Symbolic	STGS	Referential
Ohmer et al. (2022)	Symbolic	STGS	Referential
Rita, Tallec, et al. (2022)	Mixed	REINFORCE	Reconstruction
E. Cheng et al. (2023)	Symbolic	REINFORCE	Referential
Conklin and K. Smith (2023)	Symbolic	REINFORCE	Reconstruction
Feng et al. (2023)	Sub-Symbolic	REINFORCE	Referential
Hazra et al. (2023)	Mixed	REINFORCE	Interactive environment
Ri et al. (2023)	Sub-Symbolic	REINFORCE	Referential
Ueda, Ishii, et al. (2023)	Symbolic	Hybrid	Reconstruction
Ueda and Taniguchi (2024)	Symbolic	Hybrid	Reconstruction
Vithanage et al. (2023)	Symbolic	STGS	Reconstruction

explicit or implicit concepts in the environment (e.g. color, shape, size) that can be combined in a relatively large number of different ways to give an approximate account of anything an agent might observe in its environment. Lewis games are not the only emergent language game and in principle any cooperative task where there is

asymmetry of information between agents and where agents are able to transmit messages to each other can be a suitable setting for language emergence. Some of the works included in this survey use simulated, interactive environments, in which agents must typically carry out some task more complex than a straight-forward Lewis game (Bogin et al. 2018; Hazra et al. 2023; Mordatch and Abbeel 2017).

Emergent language experiments that challenge deep learning agents to communicate using sequences of discrete symbols, which is almost always the case for research into compositional languages, must calculate loss gradients in relation to the symbols chosen by a sender agent. This is a technical challenge, as the natural argmax-like operation required to convert a softmax output representing a distribution over symbols to a discrete one-hot vector is not differentiable. The two most popular solutions, featured in almost all the studies reviewed in this survey, are the REINFORCE Monte Carlo policy gradient reinforcement learning (RL) algorithm (Williams 1992), or the straight-through Gumbel softmax trick (STGS), which reparameterizes a one-hot vector as the (differentiable) sum of a softmax output, a noise vector and a constant vector (Jang et al. 2016; Maddison et al. 2016). Some considerations related to choosing either REINFORCE or STGS as an approach to optimization are discussed in Section 5.2.1. Table 1 shows the main studies included in this survey, including with the type of game used in each case, the method for estimating the gradient for word choices, and whether each study used symbolic or sub-symbolic referents.

In the simplest emergent language setup, it is sufficient for the message sent by the sender agent to consist of a one-hot vector representing a single symbol. However, it is more common for emergent language experiments to feature agents that send sentence-like messages to one another, each a sequence of discrete symbols, similar to message exchange in human conversation. With design of the agents, the task and the reward structure, the grammar with which symbols are combined into messages can be encouraged to be more like the grammar of natural languages. In particular, some emerged languages may have the quality common among natural languages that "The meaning of a sentence is determined by the meaning of its meaningful components, plus their mode of composition" (Haugeland 1979). This is the so-called "Principle of compositionality", or "Frege's principle", discussed in more detail in Section 3 below. Although the grammar of emergent languages is usually much simpler than that of natural languages (Wal et al. 2020), compositional languages have been made to emerge numerous times (see Section 5). These compositional emergent languages are examples of self-supervised learning of compositional representations, i.e. representations based on the combination of discrete abstract concepts.

2.2 The Binding Problem in Brains and in Artificial Neural Networks

At a cocktail party, a neuro-typical human is able to separate the background noise into constituent conversations in order to follow a conversation of interest and ignore other conversations. The question of how people are able to do this is the "cocktail party problem" (Cherry 1953) and is a specific example of the more general problem described by Von Der Malsburg (1994) of how an animal's brain can discern whether two sensory signals are from the same source. The latter problem is typically now called the "binding problem" (Roskies 1999). This process of separating amorphous, continuous, entangled sensory inputs appears to be automatic, at least in humans, and creates a representation of the environment defined by separable entities (e.g. a bird, a door), possibly separable themselves into smaller entities (e.g. a bird's wing, its beak) and existing in various relationships with other entities. Taking in a visual scene, humans will unconsciously identify such entities as well as relationships like "is chasing", "is supporting the weight of", and others within tens of milliseconds (Hafri and Firestone 2021). The speed at which humans can parse unstructured sensory inputs into a graph-like mental model of entities and their relationships is partly due to this process beginning before the sensory data is received. There is increasing support and physiological evidence for the idea that brains do not passively receive sensory information but actively create our perceptions by combining the outputs of "bottom-up" processing of raw sensory data with anticipatory "top-down" predictions of what sensory data will be received, based on context and experience

(Carbajal and Malmierca 2018; Clark 2013; Ficco et al. 2021). In this way, humans continuously assemble a best-estimate (abductively-reasoned) mental representation of their environment based on learned expectations and recent experience as well as sensory signals (De Lange et al. 2018). Such mental representations include entities that we may or may not recognize, to each of which may be attributed certain sensory signals and between which certain relationships may exist.

According to the structure-mapping theory of analogical reasoning (Gentner 1983), experiencing the world as constituted at least partly by notionally discrete entities and their relationships allows us to recognize familiar concepts in otherwise unfamiliar situations. It proposes that graph-like mental models of the world (i.e. containing entities that relate to one another in certain ways) allow us to more readily understand new concepts if we can identify familiar relationships: an electrical battery is a little like a water reservoir, an atom is a little like a solar system, a mitochondrion is like a furnace that burns glucose to make energy for the body (Gentner 1983; Gentner and Maravilla 2017). This ability to use analogy to situate new information in a systematic understanding of the world exists in human infants (Ferry et al. 2015) and is suggested to be an important aspect of why humans are more sample-efficient learners than artificial neural networks (ANNs) (Lake and Baroni 2018). Humans are able to learn more from less data and "[generalize] far beyond their direct experiences" while ANNs "require large amounts of data [and] struggle with transfer to novel tasks" (Greff et al. 2020). For example, humans appear to learn language skills using 4–5 orders of magnitude less data than current LLM-based AI systems (Frank 2023) and this is a relevant concern as we face the prospect of running out of high-quality training data for these enormous models (Villalobos et al. 2022; F. Xue et al. 2023). Chollet (2019) even suggests that we should not conflate skill with "intelligence", instead defining human-like intelligence as the ability to efficiently acquire new skills from the minimal starting point of the evolved "knowledge priors" that humans are born with (e.g. the assumption that the world contains objects).

There is significant support for the idea that the field of deep learning should seek inductive biases that allow ANNs to model the world in a more human-like way, in terms of discrete symbol-like concepts that can be related together, systematically combined and subjected to analogical reasoning (Chollet 2019; Garnelo and Shanahan 2019; Goyal and Bengio 2020; Greff et al. 2020; LeCun 2022; Marcus 2020; Shanahan and Mitchell 2022). This is the goal of "neuro-symbolic" approaches to AI (Garcez, Bader, et al. 2022; Sarker et al. 2021), which aim to capture the best qualities of the two dominant AI paradigms of the last century, symbolic (i.e. logic-based) and connectionist (i.e. ANN-based) AI systems. Reconciling these two paradigms has long been an active area of research (Bader and Hitzler 2005; Garcez, Lamb, et al. 2009) as combining the strengths of these two approaches is desirable: we would like a system that can reason in explainable ways based on abstract concepts, like symbolic AI, while nevertheless having ANNs' ability to deal with noisy unstructured inputs (Garnelo and Shanahan 2019; Sarker et al. 2021). In order for an ANN to create the requisite "symbol-like object representations" from unstructured, entangled input data, Greff et al. (2020) consider that it is desirable that a model can disentangle each sample into the seen and unseen explanatory features of the environment that generated the data, so that these features can be represented and manipulated in a manner reminiscent of symbolic AI. The authors dub this "the binding problem in artificial neural networks". The link between disentangled representation learning, neuro-symbolic AI and the binding problem in cognitive science is perhaps most clearly expressed by Bengio et al. (2013): "An AI must fundamentally understand the world around us, and we argue that this can only be achieved if it can learn to identify and disentangle the underlying explanatory factors hidden in the observed milieu of low-level sensory data."

Compositional emergent languages between artificial agents are of interest to the study of such disentangled representation learning, as after all a compositional emergent language must have words that refer to (denote and connote) concepts evident in the environment in which the language is grounded (see Section 3). It is also of interest to neuro-symbolic AI because, as well as being effective at creating disentangled representations (Xu et al. 2022), agents in emergent communication experiments communicate these representations to one another

using discrete symbols. Therefore, examining the conditions under which emergent communication protocols (emergent languages) can be made more disentangled and compositional may contribute to important lines of research towards neuro-symbolic AI that can learn and reason in more symbolic ways, and perhaps more like a human.

2.3 Large Language Models (LLMs)

This survey is focused on research that has found ways to encourage artificial agents to invent compositional languages, motivated partly by the idea that this may contribute to solving the binding problem in artificial neural networks (see Section 2.2) and thus assist efforts to allow neural networks to reason in explainable ways based on interpretable abstract concepts. This is presented at a time when Large Language Models (LLMs) have ostensibly mastered highly compositional languages (especially natural languages), appear to be able to reason based on human-interpretable abstract concepts, and can even generate examples of analogical reasoning (Webb et al. 2023) which is understood to be deeply connected to the ability to understand the world via abstract, symbolic concepts (Mitchell 2021). In this context it is necessary to comment on whether emergent languages are still likely to be an interesting and fruitful avenue of research, given the existence of modern LLMs. In this section, we review key differences between the ways LLMs and emergent language agents learn and use language. We also highlight some of the inherent challenges faced by LLMs, some of which may be addressable while others are inherent in language modeling as an approach to generally-capable AI agents. Based on these we highlight three abiding strengths of the emergent communication paradigm. First, useful results in emergent communication research can readily be achieved with small models that can be run on consumer hardware. Second, messages exchanged by emergent language agents are under pressure to be informative, unambiguous, relevant and truthful, which lends itself to the creation of safe and interpretable AI agent behavior. Finally, because emergent language agents are usually not "shown" existing compositional languages, they represent a true testbed for studying the reasons compositional behaviors might emerge in artificial systems. We conclude the section by noting that language modeling and emergent language approaches are best characterized as different and sometimes complementary tools in developing AI.

The distinction between LLMs and traditional language modeling is that LLMs are (often unprecedentedly) enormous models, which must be trained on proportionately large data sets (Hoffmann et al. 2022). State of the art LLMs became ten times bigger each year from the release of ELMo (Edunov et al. 2019), the last state-of-the-art pretrained language model that was not based on the Transformer architecture (Vaswani et al. 2017), until the 530-billion-parameter Megatron-Turing NLG 530B (S. Smith et al. 2022). This is because scaling the size and training data of LLMs, all else being equal, reliably improves their performance (Rae et al. 2021), bearing out Richard Sutton's "bitter lesson" that leveraging greater computation in the form of machine learning or search is a reliable way to create increasingly capable AI models (Sutton 2019). Nonetheless, even models trained on data that represents most of the content available on the web have shortcomings. Sometimes issues can be improved by applying the same type of model and the same training process but using a larger version of the model and a larger amount of training data[2]. For example, the phenomenon of LLMs generating correct-looking but factually inaccurate text (Bender et al. 2021; Weidinger et al. 2021) is reduced for languages where more training resources are available (Guerreiro et al. 2023). Similarly, the weakness of LLMs in dealing with tasks involving solving multiple sub-problems (Dziri et al. 2023; O. Press et al. 2022) appears to slowly improve with model size. If models do not improve in the desired way through additional scaling alone, it may be possible to make improvements by changing the model design or training protocol, as long as the modification is applicable at sufficient scale.

[2]A topic we won't examine in detail in this paper is that it is becoming increasingly hard to imagine where we can source order-of-magnitude increases in input text (Villalobos et al. 2022; F. Xue et al. 2023) and there are risks associated with casting a wider net for training data, such as the deleterious effect of training on the outputs of earlier language models (Shumailov et al. 2024) and the increasing difficulty of excluding biased or hateful content (Birhane et al. 2023)

For example, DeepSeek-R1 (DeepSeek-AI et al. 2025) demonstrates that the performance of LLMs on reasoning problems can be improved by using large-scale reinforcement learning to fine-tune models on problems whose answers are amenable to automated verification. Whatever the approach, researching solutions to outstanding issues with LLM-based systems generally requires access to expensive, high-powered compute resources. By contrast, most interesting results in recent emergent language research have resulted from experiments using models of fewer than 6 million parameters (see our code availability statement for details of how we reviewed model sizes for the studies included in this survey). Emergent language research can thus be run on consumer-grade desktop computers and, although the ease of running experiments does not bear on the validity of the field of study, it is worthy of mention that AI and machine learning researchers without access to state of the art compute resources can yet do valid research within the field of emergent languages.

As well as using different and smaller models, emergent language agents and LLMs are generally trained to solve different problems. Like any language model, LLMs are trained to predict unknown subsequences from within a training sequence sampled from a language. This may be based on a cloze-type task (i.e. predicting masked text from the surrounding context) or take the form of predicting how a provided sequence will continue. The most successful models to date are of the latter type (Wang et al. 2022), autoregressively predicting the next word, sub-word token, or even byte (L. Xue et al. 2022) in a piece of text. This type of training explicitly aims to reduce the cross-entropy between the true continuation of a text and the distribution over likely continuations predicted by the LLM. Thus, LLMs are not under pressure to be informative in the information-theoretic sense: their continuations should be as unsurprising as possible given their training data (as represented by the model) and the provided input text (Weidinger et al. 2021). As a result, autoregressive LLMs will frequently produce factually inaccurate text, which is a valid response to the language modeling problem if the generated text is syntactically and semantically unsurprising according to the rules of the modeled language. This can take the form of simple untrue statements about the world outside the model being presented as fact, so-called "hallucinations" (J. Li et al. 2023), or a misrepresentation of how an LLM arrived at its output, so-called "unfaithfulness" (Kambhampati et al. 2025; Turpin et al. 2023). By contrast, emergent language agents, when trained to cooperate[3], are under pressure to generate sequences of symbols that follow Grice's conversational maxims (Grice 1975). That is, they learn to generate messages that are as informative, unambiguous, relevant and truthful as possible. Thus, although LLMs provide text in natural languages we may feel we understand and emergent language agents produce outputs in languages we must work harder to interpret, emergent language approaches may have better prospects in terms of safe and interpretable AI applications and faithful human-computer interactions.

The study of emergent communication among deep learning agents originates in the simulation of language emergence for the purpose of understanding the evolution of language (Wagner et al. 2003). Simulations among artificial agents are useful for investigating the necessary or sufficient pressures that lead to certain qualities in emergent languages. When the agents are parameterized by artificial neural networks, this type of investigation is also of broad interest to the deep learning community: it can help us understand how variable binding and linguistic compositionality can be emergent behaviors in such models. Emergent language agents are more helpful in this respect than LLMs, as LLMs already have a strong prior for compositionality, having already been trained on large data sets of compositional language. Thus, although LLM-based agents can rise to the challenge of an emergent language game and can usefully model the way languages change and evolve over time and generational transfer (Kouwenhoven et al. 2025), their inherent compositionality makes it difficult to use them to identify the pressures that help compositional behaviors to emerge. *Tabula rasa* agents are a true test bed for identifying these desiderata.

Pre-trained LLMs have achieved extraordinary success in a broad range of problem domains and are well able to model the complexities of natural language as it is commonly used. Although emergent languages among

[3]Communication is less likely to emerge between two agents who are competing with one another (Noukhovitch et al. 2021)

tabula rasa deep learning agents don't come close to the complexity of natural language (Ueda, Ishii, et al. 2023; Wal et al. 2020) and the problems these agents can solve are comparatively trivial, emergent language research has its own compelling benefits as a subdomain of deep learning research. Rather than competing with large-scale language modeling, emergent communication approaches are best framed as a different, additional tool in the pursuit of effective, sample-efficient and safe AI models. Accordingly, there are opportunities to combine the strengths of emergent communication and large-scale language modeling (Steinert-Threlkeld et al. 2022), or to use the attendant theories of one paradigm to analyze the other (Taniguchi et al. 2024). As Galke and Raviv (2024) note, emergent language research may well inform the creation of future LLMs, by identifying inductive biases that can encourage the latter to acquire language in a way that better resembles human language acquisition. The study of compositional emergent languages in particular may also help us understand how ANNs can bridge the gap between a noisy continuous environment and a discrete, compositional representation. If so, emergent language research may lead to insights about how to more efficiently train foundational models, including LLMs but also models that train on entangled data such as images and sound.

3 Defining Compositionality

The emergent language literature inherits its most popular definition of compositionality from semantics: "The meaning of a sentence is determined by the meaning of its meaningful components, plus their mode of composition". This is Frege's principle or the "principle of compositionality" (Haugeland 1979) and many authors of the works reviewed in Section 5 quote it directly as a definition (E. Cheng et al. 2023; Choi et al. 2018; Conklin and K. Smith 2023; Luna et al. 2020; Steinert-Threlkeld 2020) or more generally define a compositional language as one that allows complex meanings to be represented through the arrangement of simpler units of meaning (Hazra et al. 2023; Korbak, Zubek, Kuciński, et al. 2021; Kottur et al. 2017; Kuciński, Kołodziej, et al. 2020; F. Li and Bowling 2019; Mordatch and Abbeel 2017; Ri et al. 2023; Yi et al. 2019). Although the principle has been criticized as ambiguous (Szabó 2012) and does not apply to every utterance in natural language (Grice 1975), it is a broadly helpful and accurate statement about how natural languages work and a useful concept in linguistic modeling (Donatelli and Koller 2022).

The generally-compositional nature of natural language allows language-users to understand novel sentences if they understand the constitutive "meaningful components" (vocabulary) and "mode of composition" (syntax). We would expect someone who was able to understand the syntax and vocabulary elements of the phrase "John loves Mary" to also be able to parse "Mary loves John" (Fodor and Lepore 2002). According to Frege's principle, compositional languages are thus *systematic*, in that it is possible to parse the meaning of an utterance by applying some known rules, and *productive* in that knowing a few words and syntactic structures can allow us to create a relatively large number of valid utterances with unique meanings[4]. Some authors choose to define compositionality primarily in terms of systematicity (S. Guo et al. 2020; Gupta et al. 2020; Perkins 2021b), while for others productivity is positioned as more important (Cogswell et al. 2019; Perkins 2021c; Slowik et al. 2020). Different definitions of compositionality are reflected in the design of the various available compositionality metrics, such that a certain metric may only measure compositionality according to a specific definition, as discussed in Section 4.

A notable unconventional definition of linguistic compositionality is that of Korbak, Zubek, and Rączaszek-Leonardi (2020), who identify a language as compositional if it generates messages via a compositional function (see Section 4.7 for the definition of a compositional function). Informally, this means that messages in the language are dependent on the discrete concepts that are being communicated. This can include languages whose messages are made of disentangled word-like units of meaning, but also languages where words are only

[4]It has often been argued that the number of constructible grammatically-correct sentences in a natural language is infinite. See e.g. Pullum and Scholz (2010).

interpretable as part of a complete message. Concretely, it can include languages where it is possible to guess a missing symbol in a message based on the meaning of the remaining symbols, but also languages where meaning is only defined at the message level and symbols have no inherent meaning.

To measure the compositionality of emergent languages in terms of Frege's principle, it is necessary to ascertain whether or not the meanings of the messages passed between agents are composed from those of their meaningful symbols or meaningful groups of symbols. This is by no means straightforward, for several reasons. First, we do not necessarily know what the units of meaning in an emergent language look like. Every natural language exhibits double articulation, such that utterances are made up of meaningless parts (the individual sounds of human speech, as often represented by alphabets) which are assembled into units of meaning (morphemes or words), but it cannot be assumed that emergent languages between artificial agents will have this property. In emergent languages that have this property, the symbols in a message passed between agents will usually be meaningless individually and the smallest units of meaning will most often be represented as groups of symbols. This has only recently been explored, by Ueda, Ishii, et al. (2023), who find that, fortunately for the validity of previous results which found compositionality at the level of symbols, emergent languages do not exhibit double articulation by default. Second, even if we know the units of meaning, we may not know what an overall message between agents is supposed to communicate. Most measures of compositionality in the literature require access to the ground truth about what atomic concepts are evident in the referents about which agents are communicating and assume that agents communicating successfully must be communicating about those concepts. Naturally this presents a problem for more realistic referents such as images, for which there may be no single, objective set of concepts that allows effective compositional communication. This relates to the more general challenge of underspecification in machine learning, as discussed in Section 4.2.

4 Measuring Compositionality

This section explores the approaches used in the literature so far to measure compositionality in languages emerging between deep learning agents, which is a prelude to the later discussion of the experimental design choices that appear to help emergent languages to be more compositional. As discussed in Section 3, there is not an unambiguous, universally-accepted definition of compositionality and, even if there were, it is not always clear what artificial agents are speaking about or what the meaningful units of their messages look like. Therefore, any attempt to measure compositionality necessarily involves some design decisions. Recognizing what aspect or what sense of "compositionality" was prioritized when designing a measurement technique is important when assessing experimental design choices that helped such compositionality emerge. Because different metrics and measurement strategies are useful for spotting certain things associated with compositionality (productivity, say), many authors choose to establish the compositionality of emergent languages using more than one approach. Table 2 shows how the research surveyed in Section 5 measures compositionality. It shows the dominance of systematic generalization (Section 4.2) and topographic similarity (Section 4.3) and the fact that the majority of studies will use more than one technique. "Other" accounts for any of the other approaches to measuring compositionality, which are explored in detail in this section.

4.1 Visualization, Visual Inspection and Qualitative Analysis

Visualizing the meanings of symbols or messages in emergent languages can make for striking and convincing visuals. For example, Havrylov and Titov (2017) produce a figure comparing referents to sent messages and find that "word 5747 on the first position encodes presence of an animal in the image... [and] (5747 5747 7125 * *) corresponds to a particular type of bears." With the authors' permission, we have reproduced the figure they are referring to below in Figure 4.

Table 2. Ways of assessing compositionality in studies in this survey. Visualization allowing qualitative assessment by researchers was used in 20 studies, generalization ability in 26 studies, topographic similarity in 26 studies, and methods other than these (everything classified as "Other" is included in Section 4) in 12 studies. Visualization allows qualitative assessment while other methods are quantitative, see below.

Citation	Visualization	Generalization	TopSim	Other
Kottur et al. (2017)	✓	✓		
Mordatch and Abbeel (2017)	✓	✓		
Lazaridou, Hermann, et al. (2018)	✓	✓	✓	
Bogin et al. (2018)	✓	✓		✓
Choi et al. (2018)	✓	✓		
Cogswell et al. (2019)	✓	✓		
S. Guo (2019)	✓		✓	
F. Li and Bowling (2019)	✓		✓	
Yi et al. (2019)	✓	✓	✓	
Chaabouni, Kharitonov, Bouchacourt, et al. (2020)	✓	✓	✓	✓
Dagan et al. (2020)	✓		✓	
S. Guo et al. (2020)			✓	
Gupta et al. (2020)		✓	✓	
Kuciński, Kołodziej, et al. (2020)	✓		✓	✓
Luna et al. (2020)		✓	✓	
Resnick et al. (2020)				✓
Slowik et al. (2020)		✓		
Steinert-Threlkeld (2020)	✓	✓		
Chaabouni, Strub, et al. (2021)		✓	✓	
Korbak, Zubek, Kuciński, et al. (2021)	✓	✓	✓	✓
Kuciński, Korbak, et al. (2021)			✓	✓
Mu and Goodman (2021)	✓	✓	✓	✓
Perkins (2021b)		✓	✓	
Perkins (2021c)		✓	✓	✓
Garcia et al. (2022)		✓	✓	
Ohmer et al. (2022)	✓	✓	✓	✓
Rita, Tallec, et al. (2022)		✓	✓	
E. Cheng et al. (2023)		✓	✓	
Conklin and K. Smith (2023)		✓	✓	✓
Feng et al. (2023)	✓	✓	✓	
Hazra et al. (2023)	✓	✓	✓	✓
Ri et al. (2023)	✓	✓	✓	
Ueda, Ishii, et al. (2023)			✓	
Ueda and Taniguchi (2024)			✓	
Vithanage et al. (2023)	✓	✓		✓

Another approach, effective where the underlying factors of variation are known, is to produce a table that allows visual inspection of how certain linguistic patterns may relate to certain attributes of referents (e.g. Choi

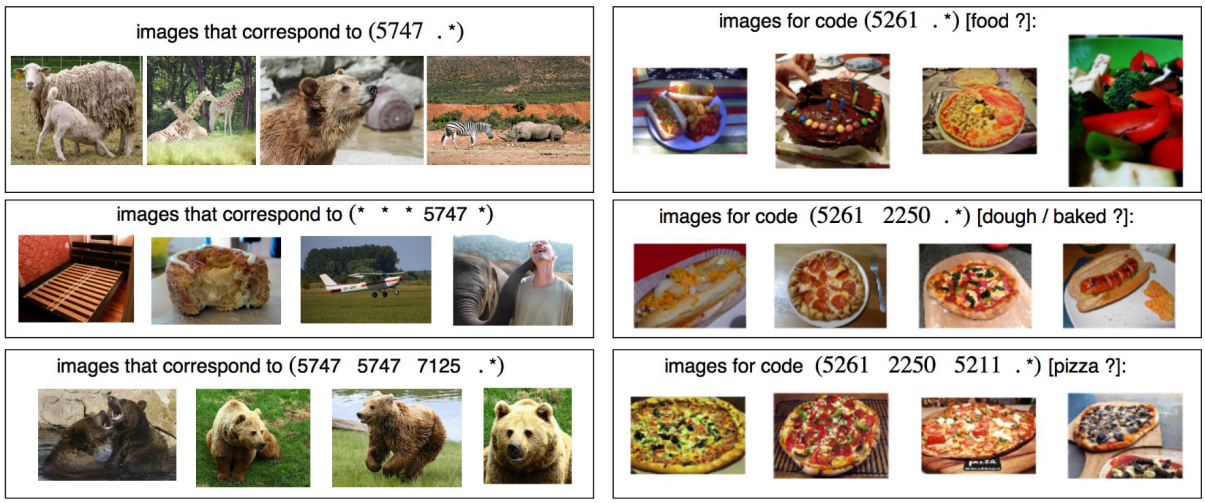

Fig. 4. Visualization showing interesting correlations between usage of symbols in an emergent language and themes/concepts in referent images, reproduced with permission from Havrylov and Titov (2017).

et al. 2018; Cogswell et al. 2019; F. Li and Bowling 2019; Yi et al. 2019). Figure 5 shows an example of this approach, found in (Yi et al. 2019) and reproduced with the authors' permission.

	blue	green	cyan	brown	red	black	yellow	white
box	aa	ea	ba	ga	da	ca	ha	fa
circle	ab	eb	bb	gb	db	cb	hb	fb
triangle	ae	**eb**	be	ge	de	ce	he	fe
square	af	ef	bf	gf	df	cf	hf	ff
star	ac	ec	bc	gc	dc	cc	**dh**	fc
diamond	ad	ed	bd	gd	dd	cd	hd	fd
pentagon	ag	eg	bg	gg	dg	cg	hg	fg
capsule	ah	eh	bh	gh	**hc**	ch	hh	fh

Fig. 5. Table reproduced with permission from Yi et al. (2019). Referents had two attributes, color and shape. The emerged language is based on messages of length 2 and generally encodes color in the first symbol and shape in the second. Exceptions are shown in bold (original emphasis).

Visually assessing how messages from an emerged language correspond to the qualities of referents is akin to field linguistics, in which a linguist observing a hitherto unexplored language may spend time with speakers of the language in the hope of spotting patterns that help decode the language. Despite its speculative nature, this approach to analysing emergent languages has remained popular (see Table 2).

4.2 Systematic Generalization

As mentioned in Section 3, systematic generalization is usually considered a defining feature of compositionality. If we accept that it is, we would expect agents to be able to use a compositional emergent language to communicate about concepts they had seen before, even if they were arranged in novel ways. Thus, many authors explicitly

test for generalization of emergent languages to novel combinations of familiar concepts (e.g. Choi et al. 2018; Kottur et al. 2017; Lazaridou, Hermann, et al. 2018). For example, Chaabouni, Strub, et al. (2021) measure the ability of agents to transfer the concepts they have learned to a relevant unseen task, given 10k training steps[5].

The relationship between compositionality and systematic generalization is not straightforward, for several reasons:

4.2.1 Compositionality May Not Be Necessary for Generalization. Compositionality appears to be sufficient for generalization of communication ability (i.e. the ability to successfully communicate about previously unseen combinations of concepts) in some cases, but it may not always be necessary. Korbak, Zubek, and Rączaszek-Leonardi (2020) find that a language that is highly "entangled", such that the meaning of a message must be memorized and cannot be derived from the meanings of the included words, performs no worse than a language displaying Fregean compositionality in terms of language acquisition speed or generalization to unseen examples. Chaabouni, Kharitonov, Bouchacourt, et al. (2020) further find that agents in emergent language experiments can sometimes generalize even when certain metrics (topographic similarity, bag of symbols disentanglement and positional disentanglement, all described below) indicate lack of compositionality in the emerged language. This may be because some form of compositionality not well-measured by those metrics is in play (see below), or because a compositional language was not required for generalization in the chosen tasks.

4.2.2 The Type of Compositionality That Emerges May Not Correlate With Chosen Compositionality Metrics. Agents may achieve systematic generalization in their ability to communicate concepts to one another without appearing compositional if the compositionality metrics used in an experiment do not measure the form of compositionality emerging in the agents' language. Conklin and K. Smith (2023) argue that some highly-compositional languages (including natural languages) include forms of linguistic variation that are deleterious to scores on popular compositionality metrics: languages with synonymy and relatively free word order are likely to score poorly on topographic similarity (a popular metric discussed in Section 4.3 below), languages exhibiting homonymy are likely to score poorly on measures assuming a one-to-one correspondence between symbols and concepts (Section 4.4). Yao et al. (2021) find that the compositional nature of natural languages (which all exhibit homonymy, synonymy, etc.) is not well measured by "simple measurements of rigid disentanglement", while Chaabouni, Strub, et al. (2021) suggest that topographic similarity in particular may become an unreliable measure of compositionality as referents in Lewis games become more complex (such as when using natural images rather than pre-processed symbolic inputs). Korbak, Zubek, and Rączaszek-Leonardi (2020) empirically show that there are identifiable types of compositionality that undermine certain popular compositionality metrics. For example, *topsim* (Section 4.3) assesses languages to be less compositional if symbol meaning somewhat depends on context (as in homonymy) and the metric *context independence* (Section 4.4.1) assesses languages to be less compositional if they make use of negation.

4.2.3 Compositional Languages May Not Generalize. Just as we cannot be sure that an emergent language generalizing to novel combinations of concepts is compositional, we cannot guarantee that a language that does not generalize in this way is not compositional. Auersperger and Pecina (2022) find that in some settings compositionality is necessary but not sufficient for generalization to held-out combinations of concepts. Such insufficiency may indicate that it is possible for agents to converge on a system of compositional concepts that is supported by the training data but cannot be extended to the type of generalization relevant to the experiment at hand (i.e. is not the system of concepts the researchers understood to be evident in the data). As Kuciński, Korbak, et al. (2021) put it, "compositionality should be defined together with features, with respect to which it

[5]The accuracy of the agents in the unseen transfer task is called "Ease and Transfer Learning" (ETL), inspired partly by the "Ease of Teaching" (F. Li and Bowling 2019) pressure often applied in emergent language experiments to elicit compositionality (see Section 5.4.1) and partly by the contrastive transfer learning approach of T. Chen et al. (2020)

holds". Moreover, in the case of sub-symbolic inputs, artificial neural networks can make use of features that humans cannot perceive (Geirhos et al. 2020; Ilyas et al. 2019) and learn compositional functions that humans cannot understand (Perkins 2021a). This relates to the broader problem of under-specification in machine learning (D'Amour et al. 2020), which is at the root of perennial concerns about deep learning, including vulnerability to adversarial examples (Ilyas et al. 2019), specification gaming (Krakovna et al. 2020), and the difficulty of aligning powerful AI with human values (Ngo 2022).

4.3 Topographic Similarity

In a paper concerning simulating the evolution of compositional languages among humans, Brighton and Kirby (2006) frame language as a mapping from some representation of a referent in a "meaning space" to some signal in a "signal space"[6]. Drawing upon Frege's definition of compositionality (see Section 3), they note that changing only a few meaningful components of a compositional message should not completely change the meaning of the message. In other words, messages that are close to one another in terms of their constituents should generally be close together in terms of meaning[7]. Early in the literature related to emergent languages among ANN agents, Lazaridou, Hermann, et al. (2018) dub this definition of compositionality *topographic similarity*[8] (*topsim*) and operationalize it as a metric that can be used to measure compositionality in emergent languages. For some set of referents and messages referring to them, the latter authors measure *topsim* as the Spearman's rank correlation coefficient (ρ) between the similarity of pairs of referents and the similarity of messages referring to those referents. Concretely, for each pair of referents and corresponding pair of messages they use the negative ρ of the Levenshtein distance between the messages and the cosine similarity between vectors enumerating features of the referents (negative ρ because they are comparing a distance to a similarity). These ways of measuring distance in the message space and similarity in the referent/semantic/meaning space have become standard, as shown in Table 3. Due to the use of the Spearman's rank correlation coefficient, *topsim* is also sometimes abbreviated to ρ. Vithanage et al. (2023) approximate *topsim* using the differentiable Pearson's correlation coefficient rather than the Spearman's rank correlation coefficient in order to create an additional loss term for their agents and optimize for this compositionality metric directly.

Although Brighton and Kirby (2006) describe topographic similarity as relating the semantic and morphological spaces of messages, i.e. the meaning and form of a message, in practice we usually will not directly know what a message in an emergent language means. Therefore, *topsim* cannot directly make use of the meaning of messages in an emergent language to test for correlation between the form of messages and their meanings, but must use a proxy for the meanings of messages: a vector representing the concepts to which the message should ideally be referring. A message's position in semantic space is thus typically represented by a vector representing what concepts are present in the message's referents. Lazaridou, Hermann, et al. (2018) use the Visual Attributes for Concepts (VisA) data set (Silberer et al. 2016), which contains both color images and human-created binary annotations of those images. This allows the agents to train on sub-symbolic image representations while *topsim* is calculated on few-hot meaning vectors. The authors find that agents trained directly on annotation vectors (i.e. disentangled symbolic input) achieve higher *topsim* than those trained on images. Possibly there are valid systems of concepts learnable from the raw images in VisA that do not perfectly align with the annotation vectors (see Section 4.2.3), making it beneficial in terms of *topsim* for agents to have the intended concepts as input.

Table 3 includes the subset of studies reviewed in Section 5 that use *topsim*. Of these 26 studies, 11 have sender agents only observing symbolic referents, i.e. one-hot or few-hot vectors. A further 11 use sub-symbolic referents, and in particular images, that are generated based on discrete parameters and thus can readily be represented as

[6]This is reminiscent of the semiotic triangle of (Ogden and Richards 1923)
[7]Possible problems with this assumption, e.g. negation, synonymy, free word order, are discussed below.
[8]Referred to in some sources as "topological similarity" (S. Guo 2019; Vithanage et al. 2023; Yi et al. 2019)

Table 3. Features of studies using *topsim* including: whether the environment was symbolic, sub-symbolic, or both were attempted; what kind of vectors were used to represent the semantic space, which may be symbolic representations such as 1-hot or few-hot vectors, or embeddings from a pre-trained vision model (PTVM) or fine-tuned vision model (FTVM); the approach to calculating distance between messages; the approach to calculating distance in semantic space. (* = We assume a symbolic representation based on the paper, although this is not explicitly stated. †= It is implied but not explicitly stated that the authors follow Lazaridou, Hermann, et al. (2018) in this respect, e.g. cases where the more general term "edit distance" is used rather than the explicit "Levenshtein distance". ‡= This study uses the complement of the Hamming distance, the count of similar entries in a fixed-length few-hot vector.)

Citation	Environment	Semantic Space	Message Distance	Semantic Distance
Lazaridou, Hermann, et al. (2018)	Mixed	Symbolic	Levenshtein	Cosine
S. Guo (2019)	Mixed	Symbolic	Levenshtein†	Hamming
F. Li and Bowling (2019)	Symbolic	Symbolic	Levenshtein	Cosine
Yi et al. (2019)	Symbolic	Symbolic	Levenshtein	Cosine
Chaabouni, Kharitonov, Bouchacourt, et al. (2020)	Symbolic	Symbolic	Levenshtein	Hamming‡
Dagan et al. (2020)	Symbolic	Symbolic	Levenshtein†	Cosine†
S. Guo et al. (2020)	Symbolic	Symbolic	Levenshtein	Hamming
Gupta et al. (2020)	Symbolic	Symbolic	Levenshtein	Cosine
Kuciński, Kołodziej, et al. (2020)	Sub-Symbolic	Symbolic*	Levenshtein†	Cosine†
Luna et al. (2020)	Sub-Symbolic	Symbolic*	Levenshtein†	Cosine†
Chaabouni, Strub, et al. (2021)	Sub-Symbolic	FTVM	Levenshtein†	Cosine
Korbak, Zubek, Kuciński, et al. (2021)	Sub-Symbolic	Symbolic	Levenshtein	Levenshtein
Kuciński, Korbak, et al. (2021)	Sub-Symbolic	Symbolic	Levenshtein	Hamming
Mu and Goodman (2021)	Mixed	Symbolic	Levenshtein	Multiple
Perkins (2021b)	Mixed	Symbolic	Hamming	Hamming
Perkins (2021c)	Sub-Symbolic	Symbolic	Not stated	Hamming
Garcia et al. (2022)	Sub-Symbolic	FTVM	Levenshtein†	Cosine
Ohmer et al. (2022)	Symbolic	Symbolic	Levenshtein†	Cosine
Rita, Tallec, et al. (2022)	Mixed	PTVM/symbolic	Levenshtein†	Multiple
E. Cheng et al. (2023)	Symbolic	Symbolic	Levenshtein	Euclidean
Conklin and K. Smith (2023)	Symbolic	Symbolic	Levenshtein†	Cosine†
Feng et al. (2023)	Sub-Symbolic	Symbolic	Levenshtein	Hamming
Hazra et al. (2023)	Mixed	Symbolic	Levenshtein†	Hamming
Ri et al. (2023)	Sub-Symbolic	Symbolic	Levenshtein†	Cosine
Ueda, Ishii, et al. (2023)	Symbolic	Symbolic	Levenshtein†	Hamming
Ueda and Taniguchi (2024)	Symbolic	Symbolic	Levenshtein	Hamming

few-hot vectors for the purpose of calculating *topsim*. As mentioned above, Lazaridou, Hermann, et al. (2018) use a data set including a few-hot annotation vector for every image, which can be used as a semantic vector. Three papers in this survey use *topsim* to measure the compositionality of languages grounded in continuous, entangled input for which the ground truth concepts are not known. These are Garcia et al. (2022), who use the embeddings from fine-tuned image classification models to represent vectors in semantic space, and Rita, Tallec, et al. (2022) and Chaabouni, Strub, et al. (2021). The latter two studies each use a ResNet-50 (He et al. 2015)

model pre-trained on ImageNet (J. Deng et al. 2009) using the BYOL self-supervised learning method (Grill et al. 2020) to pre-process referents for speaker agents. In both cases, the pre-trained model takes referent images and produces an embedding vector of dimension 2048[9]. Image embeddings are used as referents rather than the images themselves and cosine distance between referents is used as the distance in semantic space. Both Rita, Tallec, et al. (2022) and Chaabouni, Strub, et al. (2021) find that *topsim* correlates poorly with generalization performance in their settings. This may be partly because they calculate the semantic distance as the cosine distance between referents themselves, which Bouchacourt and Baroni (2018) note becomes less and less correlated with the cosine distance between agents' representations of those referents as training continues. It may also relate to the observation by Perkins (2021b) that *topsim* is only an indication of the compositionality achieved by the sender, whereas generalization may indicate that both agents are working with compositional representations. Garcia et al. (2022) use agents' representations as semantic vectors to calculate *topsim*.

As well as facing challenges in terms of calculating a distance in semantic space between messages, *topsim* as it is usually formulated is known to penalize certain features that may occur in compositional languages, such as synonymy, homonymy and relatively free word order (Chaabouni, Strub, et al. 2021; Conklin and K. Smith 2023; Kuciński, Kołodziej, et al. 2020). This is due to the underlying assumption that proximity of meaning is correlated to proximity of representation in compositional languages, with the latter being usually (see table 3) calculated using Levenshtein distance, which is not invariant to synonym use or word reordering. Consider "I approached the bank" (physically approaching a river bank) versus "I approached the bank" (figuratively approaching a financial institution by making a communication), and "My dog is large" versus "He is big, my hound" as examples in English of how identical messages can have dissimilar meanings and dissimilar messages can have near-identical meanings. Additionally a simple negation, a feature present in between 9-32% of English language sentences (Jiménez-Zafra et al. 2020), may also dramatically shift the meaning of a sentence without changing the constituent words or word order very much. For these reasons, Yao et al. (2021) suggest that *topsim* is likely to underestimate the compositionality of natural languages. To date, we are not aware of efforts to systematically explore the behavior of *topsim* when it is formulated with alternative distance metrics.

Ueda, Ishii, et al. (2023) and Ueda and Taniguchi (2024) make use of two distinct topographic similarity measures, *C-TopSim* (*topsim* with respect to symbols) and *W-TopSim* (*topsim* with respect to "words") and find that both significantly improve when agents base their messages on a language model prior, as described in Section 5.4.

4.4 Measures Based on Co-Occurrence of Symbols and Referent Attributes

Some approaches to measuring compositionality in emergent languages assume that, if a language is compositional, it will be easy to predict what symbols will be used in a message if we know what concepts are part of the referent, and vice versa. This is often operationalized in terms of the idea that each symbol should have a concept inherent in the environment to which it principally refers, i.e. its "principal meaning" (Kuciński, Kołodziej, et al. 2020) and each concept in an environment ought to have a symbol by which it is usually denoted. As with all other ways of measuring compositionality in the literature so far except generalization performance and (in a few exploratory cases) *topsim*, these methods assume we have access to the set of concepts within any referent.

4.4.1 Context Independence. Bogin et al. (2018) propose *context independence* as a measure of compositionality, based on the idea that in compositional languages "symbols retain their semantics in various contexts". Let the principal meaning of a symbol s be the observed concept c with which it most commonly occurs. We denote the symbol whose principle meaning is c as s_c. Then *context independence* is defined as

$$\frac{1}{n}\sum_{c} p(s_c|c) \cdot p(c|s_c)$$

[9]In Grill et al. (2020) a representation vector of the same size is taken from the final average pooling layer of a ResNet model in a similar way.

where n is the size of the set of all concepts in an environment to which messages might refer. That is, context independence is the mean of $p(s_c|c) \cdot p(c|s_c)$ over all concepts. This will be less if $p(s_c|c)$ is smaller, e.g. in a language featuring synonymy so that a concept might be denoted in several ways, or if $p(c|s_c)$ is smaller, e.g. in a language featuring homonymy so that a symbol does not always denote the same concept.

4.4.2 Bag-of-Symbols Disentanglement. Chaabouni, Kharitonov, Bouchacourt, et al. (2020) make use of *topsim*, but note that it is agnostic to what manner of compositionality it detects, as long as messages that are similar according to some similarity metric are also similar in meaning. They thus offer two additional metrics: *positional disentanglement* (described below in Section 4.5) and *bag-of-symbols disentanglement* (*bosdis*).

As the name implies, *bosdis* is a measurement of disentanglement. It measures the extent to which separate concepts exhibited by referents are encoded by dedicated symbols. It is a calculation based on a primary attribute a_1^j and a secondary attribute a_2^j whose presence or absence in a referent respectively provide the most and second-most information about the count of times n_j that a symbol j will appear in a message. The mutual information between the symbol and the most informative attribute, $\mathcal{I}(n_j; a_1^j)$, beyond the mutual information between the symbol and the second-most informative attribute, $\mathcal{I}(n_j; a_1^j)$, normalized by the marginal entropy in the distribution of times symbol j appears in messages, $\mathcal{H}(n_j)$, is

$$\frac{\mathcal{I}(n_j; a_1^j) - \mathcal{I}(n_j; a_2^j)}{\mathcal{H}(n_j)}$$

This number is maximized (to 1) for a given symbol j if the symbol j is only used to communicate the presence or absence in some referent of attribute a_1^j and if $\mathcal{I}(n_j; a_1^j)$ is equal to $\mathcal{H}(n_j)$.

Bosdis for a language is calculated as the average of this measure over all the symbols in an agent's vocabulary:

$$bosdis = 1/s_{voc} \sum_{j=1}^{s_{voc}} \frac{\mathcal{I}(n_j; a_1^j) - \mathcal{I}(n_j; a_2^j)}{\mathcal{H}(n_j)},$$

where: s_{voc} is the number of symbols in the vocabulary available to a sender agent.

4.5 Measuring the Correspondence Between Positions in a Message and Attributes of a Referent

Chaabouni, Kharitonov, Bouchacourt, et al. (2020) note that it is sometimes the case in natural languages that certain positions in a message relate to certain attributes, so that the symbol in that position will encode the value of that attribute. For example, a two-word phrase in English indicating a color and a type of fruit along the lines of "green apple" has the syntactic characteristic that the first position encodes the attribute of color and the second the attribute of fruit type, so that symbol reuse is permitted without ambiguity as in e.g. "orange orange". They thus propose *positional disentanglement*, which is a measure of the extent to which symbol positions in a language of fixed-length messages come to be associated with certain attributes. In the context of a language of binary strings (i.e. a language with two symbols in its vocabulary) Resnick et al. (2020) propose *residual entropy*, which assumes that an attribute can be jointly represented by a number of message positions (called a partition of the message) and Perkins (2021a) suggests *human compositional entropy* as a similar but less computationally intensive measure of compositionality. Vithanage et al. (2023) more straightforwardly employ Symmetrical Uncertainty, a scaled, normalized measure of mutual information between message positions and referent attributes. As usual, all of these measures assume we are measuring compositionality with respect to some known set of attributes within referents. The specific methods for calculating each of these compositionality measures is described below.

4.5.1 Symmetrical Uncertainty. Vithanage et al. (2023) make use of a scaled, normalized measure of the mutual information between what symbol is in a certain position of a message and what value a certain attribute of the referent takes:

$$SU(A,S) = 2 \cdot \frac{\mathcal{I}(A,S)}{\mathcal{H}(A) + \mathcal{H}(S)}$$

Where A is the distribution of values taken by a certain attribute of referents (e.g. an attribute might be "color" and the value might be "red") and S is the distribution over symbols in a certain position in a message.

$\mathcal{I}(A,S)$ is the mutual information between the two distributions, $\mathcal{H}(A)$ is the entropy of A and $\mathcal{H}(S)$ is the entropy of S. This is *Symmetrical*[10] *Uncertainty*, which will take a value between 0 and 1 (W. H. Press et al. 1988). The authors use this to create a visualization of how informative certain message positions are about certain attributes in referents, which they visualize as a heat map.

4.5.2 Positional Disentanglement. Positional disentanglement (*posdis*) is introduced in the same paper as *bosdis*, described in Section 4.4.2 above (Chaabouni, Kharitonov, Bouchacourt, et al. 2020). It aims to measure whether certain positions in a fixed-length message (e.g. the third symbol) are associated with particular attributes in referents. It takes a similar form to *bosdis*:

$$posdis = 1/m_{len} \sum_{j=1}^{m_{len}} \frac{\mathcal{I}(m_j; a_1^j) - \mathcal{I}(m_j; a_2^j)}{\mathcal{H}(m_j)},$$

where: m_{len} is a (fixed) message length; m_j is the symbol in the j$^{\text{th}}$ position in the messages sent between agents; a_1^j is the attribute (of referents) that has greatest mutual information with m_j; a_2^j is the attribute (of referents) that has 2$^{\text{nd}}$ greatest mutual information with m_j.

Auersperger and Pecina (2022) give encouraging indications that, for simple referential games where agent representational machinery seems to be able to support generalization to out-of-domain combinations of concepts, good results on out-of-domain data was only possible for languages that displayed high positional disentanglement and *topsim*. Korbak, Zubek, and Rączaszek-Leonardi (2020) similarly show that a combination of *bosdis*, *posdis* and *topsim* is able to successfully identify compositionality, but found shortcomings in each of the individual metrics they tested except for tree reconstruction error (discussed in Section 4.7.2).

4.5.3 Residual Entropy and Human Compositional Entropy. Resnick et al. (2020) investigate whether the information capacity of messages between agents can affect the compositionality of emergent languages, controlling message capacity by fixing the length of messages and only allowing vocabulary size of 2 so that all messages are binary sequences. In this setting, the authors consider a set p of non-intersecting partitions of message positions, such that the number of partitions is the same as the number of attributes to be communicated about in an environment. For example, the first partition might include the first and third position in a message. For any p and a language L, they define *residual entropy* as

$$re(p,L) = \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \frac{\mathcal{H}(A_i|m[p_i])}{\mathcal{H}(A_i)},$$

where:

- $\mathcal{N}$ is the number of partitions and also the number of referent attributes about which agents need to communicate
- $m[p_i]$ represents the subset of symbols in the message m included in the i^{th} partition of p
- $\mathcal{H}(A_i)$ is the marginal entropy of the distribution of values taken by the i^{th} attribute a_i of a referent

[10] "Symmetric" in some sources.

- $\mathcal{H}(A_i|m[p_i])$ is the conditional entropy of the distribution of values taken by the i^{th} attribute a_i of a referent, given that we know the tokens $m[p_i]$ in a message

In other words, the *residual entropy* for some set of partitions p and some language L is the average over all i (between 1 and $\mathcal{N}$) of the proportion of the entropy in the distribution of values taken by attribute A_i that remains when we know what symbols were in partition p_i.

Of course, if the list of partitions is not well aligned with the list of attributes about which agents are communicating, p_i might have little to do with A_i. Thus the *residual entropy* of an emerged language L is the *residual entropy*, as calculated above, for that language and some optimal partitioning p (from the set of all such partitionings $\mathcal{P}$) that minimizes the *residual entropy*:

$$re(L) = \min_{p \in \mathcal{P}} re(p, L).$$

Although Resnick et al. (2020) use *residual entropy* in a scenario with vocabulary size 2, it is applicable for languages with larger vocabularies as shown by Perkins (2021a). The latter author also improves on *residual entropy* by introducing a greedy method to approximate the optimal set of partitions, overcoming the issue that finding this set by exhaustive search is otherwise $O(|A|^l)$ in the message length l for some fixed-size set of attributes A that referents may have. The greedy partitioning uses the mutual information between position-concept pairs: a message position is included in the partition for a concept if the position has greatest mutual information with that concept out of all concepts, giving an algorithm that is instead $O(|A| \cdot l)$.

4.6 Conflict Counting

Kuciński, Kołodziej, et al. (2020) describe an experiment in which referents will always exhibit $k = 2$ concepts (color and shape) and a sender agent will be constrained to sending messages containing $l = 2$ symbols chosen from a vocabulary of $|V| >= k$ symbols. In such settings where $k = l$ they suggest conflict counting as a measure of compositionality, which is analogous to *context independence* (4.4.1) except that it also takes into account the position where a symbol occurs. They define the principal meaning of a symbol s in position j of a message as the attribute value that is most commonly evident in referents when that symbol is in that position. The *conflict count* is intuitively the count of times the symbol s is in position j but the principal meaning is not true for the referent. This correlated well with *topsim* (Kuciński, Kołodziej, et al. 2020), which is to be expected since minimizing *conflict count* would naturally maximize *topsim*, but perhaps implies that *conflict count* may have some of the same shortcomings.

4.7 Recursive Methods

Recursive methods are the most direct and reliable way to measure compositionality, in cases where it is possible to use them. One of these, *Tree Reconstruction Error*, was the only measure of compositionality able to correctly identify whether each of a selection of constructed languages was compositional in a study by Korbak, Zubek, and Rączaszek-Leonardi (2020). This could only otherwise be achieved by combining multiple metrics together. However, recursive methods not only have the familiar dependency of knowing the ground truth attributes that agents will be communicating about, but also requires knowing something about the order in which those concepts should be combined in order to accurately represent a referent. For example, if the referent is a red ball on a brown table, the correct "bracketing" of the involved concepts is (((red, ball), on), (brown, table)), to represent that the ball is on the table and not the other way around.

Before describing the two recursive methods discussed in this survey, we will review the related concepts of compositional representation functions and bracketing.

4.7.1 Compositional Representation Functions and Bracketing. Andreas (2019) follows Montague (1970) in defining a compositional representation function $f(x)$ in the following way:

- Assume:
  - There is some set of observations $x \in X$ relating to an environment
  - There is some system of salient concepts that may be evident in any x
  - There is a binary bracketing function $\langle \cdot, \cdot \rangle$ that can be applied in order to combine two concepts into a composite concept, e.g. $\langle blue, square \rangle$
  - Any atomic concept, e.g. $blue$, or bracketing of atomic concepts into a more complex concept, e.g. $\langle large, \langle blue, square \rangle \rangle$, is called a derivation[11]
  - Two derivations can always be bracketed together to form a composite derivation (i.e. the set of derivations is closed under the bracketing operation)
- Then a representation function f defined on X is compositional if there is some bracketing operation $*$ such that for any $x = \langle x_a, x_b \rangle$ we also have $f(x) = f(x_a) * f(x_b)$

Atomic derivations/concepts (i.e. those that are not made by bracketing simpler derivations) are called primitives. Informally, a compositional representation function $f(x)$ depends on the primitive concepts observable in x. Korbak, Zubek, and Rączaszek-Leonardi (2020) use this definition of a compositional function to inform their definition of a compositional language (see Section 3). Andreas (2019) assumes that there is a "derivation oracle" that can tell us the structure of any $x \in X$ in terms of the primitive concepts present and how they should be combined by repeatedly applying a bracketing function. Combining these primitives together in the order given by the derivation oracle will produce a representation of x.

4.7.2 Tree Reconstruction Error. Andreas (2019) argues that a function is compositional to the extent that an output $f(x) \rightarrow y$ can be well-approximated by applying a bracketing procedure $\hat{f}$ to a correct derivation $D(x)$ of the same input x (i.e. the set of atomic concepts evident in x and the correct order in which to combine them, as described in section 4.7 above). They define *Tree Reconstruction Error (TRE)* for a representation of a given x as

$$\delta(f(x), \hat{f}(D(x)))$$

for some distance function δ defined in the output space of $f/\hat{f}$, assuming that $\hat{f}$ is optimized to reduce this distance as much as possible. The *TRE* for a representation function is then the expected *TRE* for representations made by that function, or in other words, the irreducible error of modeling f as an explicitly compositional function.

Korbak, Zubek, and Rączaszek-Leonardi (2020) use constructed languages displaying various kinds of compositionality to test the ability of different compositionality metrics to detect different types of compositionality and find that *TRE* is the only metric that was able to detect all forms of compositionality while also correctly identifying non-compositional languages. Their work also tested systematic generalization, *topsim*, *posdis*, *bosdis*, *context independence* and *conflict count*. They measure the *TRE* of the function used to generate messages in a language, using a neural network to approximate the ground truth message via a binary bracketing process, rather than measuring the compositionality of the function that a receiver agent might use to parse a message. This leaves open the question of whether a language that is compositional in the way messages are generated is also compositional in the way messages are parsed, and how *TRE* may be applied to detect languages that are compositional in the latter sense.

4.7.3 Approximate Compositional Reconstruction (ACRe). Inspired by *Tree Reconstruction Error*, Mu and Goodman (2021) propose *Approximate Compositional Reconstruction (ACRe)* as a way to probe for compositionality in emerged languages.

[11] A term from the theory of formal grammars.

ACRe is more complex than *TRE* in the way it builds up compositional representations, using multiple trainable bracketing functions to represent the concepts of AND, OR and NOT. The authors note that this scheme is extensible and could be used to probe for complex recursive constructions as are common in natural languages.

The procedure is as follows:

(1) Sender and receiver agents are trained in an emergent language game where there is some known set of atomic concepts to learn
(2) Using messages that are known (by some oracle function) to have been produced in response to a referent exhibiting a certain primitive concept, a small language model (LM), implemented as a 2-layer Transformer (Vaswani et al. 2017), is trained to generate messages referring to that concept
(3) The process in (2) is repeated for every primitive concept
(4) Using the language models trained in 2-3 along with samples of messages that refer to intersecting concepts (e.g. blue AND square = blue square), a sequence-to-sequence model is trained to represent the AND operation: it takes as input the concatenated output of the language models representing any two primitive concepts a and b and is trained to generate messages that refer to the logical statement a AND b.
(5) In a similar way to 4, LMs are trained to represent the operations OR and NOT.
(6) The primitive concept LMs and logical operation LMs are trained to allow them to be composed together in the order prescribed by a derivation oracle (see definition above in Section 4.7.1) to predict the message a trained sender agent will use to communicate about a referent (the approximate compositional representation, ACRe).
(7) The extent to which such an ACRe can accurately model the sender's messages is a measure of the compositionality of the emerged language.

4.8 Future Directions for Compositionality Metrics

As we have shown in this section, there is a diversity of metrics available to measure different kinds of compositionality in emergent languages. As emergent language experiments aim to become more aligned to real-world communication in order that insights from deep learning research can more readily be applied to linguistic research and vice versa (Galke and Raviv 2024), it is reasonable to expect experiments to use more sophisticated (multi-modal, entangled) referents and for emergent languages to become more complex in structure. It is also less likely that data sets used in emergent language experiments will have full annotations of underlying concepts (and perhaps this is appropriate given that any system of concepts underlying referents may not be the only such system or even the preferred/easiest system, see Section 4.2.3). Although many compositionality metrics have been put forward in the literature, to our knowledge only accuracy on held-out referents (generalization) and topographic similarity have proven applicable in cases where the atomic concepts in referents are uncertain. This is not a dire problem, as *topsim* as usually formulated, using Levenshtein distance, is capable of detecting many forms of compositionality (Resnick et al. 2020). *Topsim*'s weaknesses, particularly in terms of detecting compositionality in languages with a high degree of variation (Conklin and K. Smith 2023), may yet be at least partially addressed by varying the edit distance used to calculate the metric. For example, although *topsim* is a brittle detector of compositionality in languages with relatively free word order, as discussed above, using a word-order-invariant message similarity measure such as Jaccard/Tanimoto similarity might be sufficient to detect compositionality. Thus, although the other metrics described above may each be useful in detecting certain kinds of compositionality, especially in examples where atomic concepts are knowable, we advocate a narrowing of focus towards topographic similarity and generalization in future research. This should ideally be accompanied by work to expand topographic similarity into a family of metrics using a range of morphological and semantic distance measures, in order to be able to detect compositionality in the presence of more complex languages exhibiting e.g. homonymy, synonymy and free word order. Finally, the literature suggests that following the

example of Garcia et al. (2022) and using agents' learned representations as the semantic space for calculating *topsim*, as opposed to the embedding space of a separately-trained computer vision model as in e.g. Chaabouni, Strub, et al. (2021) or Rita, Tallec, et al. (2022), will allow compositionality to be measured more successfully with more complex entangled referents.

5 How Experimental Design Choices Can Affect the Compositionality of Emergent Languages

Following the review in previous sections of the ways in which compositionality in emergent languages is defined and measured, this section will survey the ways in which good choices in the design of emergent communication experiments can lead to more compositional emergent languages. Previous work in emergent language research has established that compositionality may not be required for agents to perform perfectly in emergent language games or even to generalize their language to unseen data (see Section 4.2), so additional biases or constraints must be applied to emergent language games in order to increase the pressure for languages to be compositional (Chaabouni, Kharitonov, Bouchacourt, et al. 2020; Kottur et al. 2017; Kuciński, Korbak, et al. 2021).

In the subsections that follow, experimental design choices are grouped by theme. These include

- The chosen architecture for agents in an experiment
- Optimization and regularization choices, e.g. choosing to use a reinforcement learning (RL) algorithm or add additional terms to a loss function
- Game design choices, e.g. what agents have to communicate about and how they can achieve reward

There is also a separate section for game design choices that aim to subject communicating agents to pressures that may occur in real-life communication between animals, as the deep relationship between evolutionary linguistics and emergent language experiments has led to many such ideas.

Chaabouni, Strub, et al. (2021) warn that we should be wary of interventions that appear to improve compositionality on toy problems, as they may be ineffective or even inhibitive in more realistic settings. A related concern is that most emergent communication research to date has used agents that are based on relatively small ANNs and it is unclear whether the methods that led to better compositionality in small models will be as effective for large models. Based on reasonable efforts to run the code provided with the reviewed publications (see Section 8 for a link to a code repository showing how this was done), most of the speaker agents used in the research surveyed here have fewer parameters than e.g. MobileViT-S (Mehta and Rastegari 2022), a recent 5.6-million-parameter Vision Transformer model (Dosovitskiy et al. 2021) designed to be small enough to run on mobile devices. Thus, there is a need for more research into whether the many promising techniques described in this section are applicable in experiments with larger agents.

5.1 Architecture of Emergent Language Agents

5.1.1 Matched and Mismatched Architectures. The emergence of a compositional language between a sender agent and a receiver agent relies on both agents developing a compositional representation of their environment. Thus it is necessary to consider the inductive biases of both agents in order to ensure that, for example, the compositionality of messages sent by the sender is put to good use by the receiver (Garcia et al. 2022; Perkins 2021b). Perkins (2021b) notes the difference between *topsim* (Section 4.3) and generalization accuracy (Section 4.2) in this regard: *topsim* is only an indication of the compositionality achieved by the sender, whereas generalization may indicate that both agents are working with compositional representations. Garcia et al. (2022) find that using ANN architectures designed for disentangled representation learning produces more compositional languages, and using mismatched architectures offers a further benefit to compositionality but will make it more difficult for agents to create a successful language. The latter authors avoid the problem of compositional representations being lost as they move through a recurrent neural network (RNN) architecture that produces entangled representations

by using an RNN with a specially designed attention mechanism to continually inject compositional information into the hidden state.

5.1.2 Modelling Visual Perception. Lazaridou, Hermann, et al. (2018) found that agents trained directly on disentangled feature vectors created languages with higher *topsim* than those trained directly on images. However, for the reasons discussed in Section 2.2 it is desirable that agents should be able to ground an emergent language in entangled input and learn how to disentangle it in the process. Thus, several studies make use of agents with components able to process visual information.

Kuciński, Korbak, et al. (2021) find that the inductive biases inherent in a convolutional neural network (CNN) seem to support compositionality in a referential game with sub-symbolic (raw image) referents, which accords with the suggestion of Greff et al. (2020) that separating the input of ANNs into spatial or temporal parts is likely to be beneficial for learning compositional representations. Ri et al. (2023) further find that allowing sender and receiver agents to separately attend to different areas of image referents in a referential game enables significantly higher compositionality in terms of *topsim* and generalization accuracy, suggesting that "the human capacity for dynamic focus may have contributed to developing compositional language". Perkins (2021c) achieves good generalization to unseen objects in a visual data set with receiver agents that use the embedding of a received message to calculate attention over all layers of a CNN image encoder.

5.1.3 Model Capacity. Resnick et al. (2020) hypothesize that two factors influencing the emergence of compositional languages are the capacity of the agents to memorize the possible states of their environment and the bandwidth available for communication. They show that in some cases a compositional communication protocol requires more bandwidth than a non-compositional one, and so a compositional protocol cannot emerge if the Shannon information transmissible per message is restricted too much. They also hypothesize that there should be an optimal range of model size outside of which a compositional language will not emerge, as models that have too many parameters will simply memorize the input space and models with too few parameters will fail to learn at all. The expectation of sufficiently-large models memorizing their inputs aligns with the standard expectation that ANNs will memorize training examples if this is within their capacity, as this is often the simplest solution to the problem posed by their training data, i.e. the easiest "shortcut" to low loss (Geirhos et al. 2020). However, Resnick et al. (2020) only find evidence for a minimum model capacity (size) for compositionality to emerge and note that greater bandwidth and capacity both seem to benefit compositionality in their experiments.

The counterintuitive result that compositional languages can emerge between agents large enough to memorize all possible referents supports earlier work showing that compositional languages are easier to learn than languages with less systematicity (Galke, Ram, et al. 2024; F. Li and Bowling 2019). At the extreme, non-systematic languages can only be learned via memorization, so if memorization was always easier for deep learning agents there would be little pressure for compositional languages to emerge at all. Thus, the emergence of compositionality where memorization is available implies that sometimes compositionality, not memorization, is the easier route. Research elsewhere in deep learning suggests that, during training, models of sufficient capacity will go through periods of memorization followed by "interpolation" and rule learning (Nakkiran et al. 2019), and that this extends to the processes occurring in neural networks during language acquisition (A. Chen et al. 2024).

5.1.4 Sequence-Processing Architectures. Central to agent design in emergent language experiments is their facility for processing sequences. Nevertheless, and despite known differences in the expressive power of different sequence-processing architectures (Deletang et al. 2023), there has been limited exploration of what architectures are more likely to allow agents to create compositional languages and almost all studies in this survey involving agents that must produce or interpret multi-symbol messages make use of a recurrent neural network such as a GRU or LSTM. In a reconstruction game using MNIST images (L. Deng 2012) as referents, Devaraj et al. (2020) find that receiver agents reading messages left-to-right using an LSTM will ascribe more meaning to earlier symbols

and interpret later symbols as contributing progressively finer detail, implying a simple grammar involving progressive partitioning of the meaning space.

Adjacent to the emergent language literature there has been some exploration of whether certain types of sequence processing neural network are relatively more able to learn tasks that involve systematic generalization based on compositional training data (data that includes atomic concepts combined in various ways). This is of importance to emergent language research if compositional communication protocols are less likely to emerge among agents that struggle to process compositional data. Perkins (2021a) tests the ability of various neural network architectures to learn a created language in which "relocatable atomic groups of tokens" denote concepts, similar to compositionality in natural languages. He finds that dense neural networks (multilayer perceptrons, MLPs) learn this type of language inefficiently, Transformer (Vaswani et al. 2017) networks learn it successfully and recurrent neural networks learn it most efficiently of all, with GRU (Cho et al. 2014) learning more efficiently than LSTM (Hochreiter and Schmidhuber 1997). The latter work is reminiscent of, but in disagreement with, work by Hupkes et al. (2020) that suggests Transformers are more capable than LSTMs of generalizing on compositional tasks, suggesting that more work is required on this topic to fully understand the dynamics at play.

5.1.5 The Obverter Architecture. Choi et al. (2018) present a referential game between a pair of novel agents that are both able to send or receive messages. The agents alternate in these roles during training and are optimized using the obverter method, originated by Batali (1998). Each agent has a language module, which can generate an embedding from a sequence of symbols, a vision module that generates an embedding from a referent image, and a decision module that takes the language and vision embeddings as input and rates the quality of the sequence of symbols at describing the image. The obverter method entails a sender agent constructing a message by repeatedly appending the token from its vocabulary that maximizes the quality of the sequence according to its own decision module. This is repeated until the quality reaches a specified threshold or the maximum number of tokens is reached.

Although the obverter method has an obvious shortcoming in that humans do not try out every possible next word in a sentence before deciding which one works best, an agent trained in this way has the human-like quality that it produces a message that is informative according to its own assessment. This is different from the majority of emergent language games, in which a sender agent will construct messages in order to maximize the comprehension of another agent. Using this technique, the authors demonstrate that agents can develop a language that is clearly compositional based on visual inspection of the relationship between messages and concepts, and that this language allows systematic generalization to unseen combinations of concepts.

Bogin et al. (2018) further develop the obverter method to propose the context-consistent obverter (CCO). In their multi-task grid world setting, the job of a sender agent is to observe the initial state w_0 of an instance of the grid world and the goal that must be achieved g, and then send a message m to a receiver agent that sees the world state w_t at every time step but only understands the goal based on the message received from the sender. Agents are able to generate an embedding $f_m(w, m)$ from a world-message combination or an embedding $f_g(w, g)$ from a world-goal combination (if they are taking the role of the sender agent and can thus see the goal). As in the obverter technique presented by Choi et al. (2018), the sender will generate a variable-length message symbol-by-symbol and chooses the "best" next symbol with probability 1.0. The score function Ψ in Bogin et al. (2018) is

$$\Psi(w, g, m) = -d(f_m(w, m), f_g(w, g)),$$

where d is the Euclidean distance function, i.e. the perfect message is one such that the world-message embedding is the same as the world-goal embedding and the receiver agent will understand the goal from the sender's message as though it had been able to observe the goal.

The authors pre-train each agent involved in the task to complete similar tasks alone with visibility of both the environment and the goal, allowing f_g to be learned in advance of learning to communicate. Thus, when learning to communicate, f_g is fixed and f_m is trained to maximize

$$\sum_{\hat{m} \in P(m)} \alpha \cdot A \cdot \Psi(w, g, \hat{m}),$$

where A for an episode is the advantage of the utterance produced by the speaker, i.e. the reward it achieves beyond a predicted reward for the active world-goal combination. When calculating the reward for optimization purposes, α is 1 in cases where A was greater than zero, else 0.5 (to emphasize the updates for advantageous messages), Ψ is the score function above and $P(m)$ is the set of all prefixes of a message.

This approach is more stable than either policy gradient learning or the obverter method of Choi et al. (2018) and achieved better compositionality in terms of *context independence* (defined in Section 4.4.1) in both referential and grid world tasks.

5.2 Optimization Approaches

5.2.1 Supervised Learning Versus Reinforcement Learning (RL). One challenge in emergent language experiments is how to calculate gradients for the sender's choice of symbols in a message, since the natural argmax-like operation required to convert a softmax output representing a distribution over symbols to a discrete one-hot vector is not differentiable. The two most popular solutions, featured in almost all the studies reviewed in this survey, are the REINFORCE Monte Carlo policy gradient reinforcement learning (RL) algorithm (Williams 1992), or the straight-through Gumbel softmax trick (STGS), which reparameterizes a one-hot vector as the (differentiable) sum of a softmax output, a noise vector and a constant vector (Jang et al. 2016; Maddison et al. 2016). The latter allows errors from the receiver agent to be directly backpropagated through the communication channel to the policy network of the sender agent, facilitating supervised learning. It is not clear that either of these strategies is better for the emergence of compositional languages, and indeed compositionality has been observed in languages among agents using both REINFORCE (Cogswell et al. 2019; Hazra et al. 2023; Yi et al. 2019, and others) and STGS (S. Guo 2019; Kuciński, Kołodziej, et al. 2020; Mordatch and Abbeel 2017; Resnick et al. 2020, and others). Possibly reinforcement learning is somewhat preferable: E. Cheng et al. (2023) show that sender agents learning to imitate a community of pretrained senders will be more biased towards learning languages with higher *topsim* if trained using RL, which the authors attribute to the mode-seeking characteristic of RL (Bishop 2006) allowing agents to focus on imitating more easily learnable languages. Supervised learning through STGS does appear to have the potential to make compositional languages emerge more quickly (Havrylov and Titov 2017) but has the downside that the sender and receiver agent can be viewed as parts of one large neural network architecture (Lazaridou and Baroni 2020), which may be hard to justify in cases where an experiment is intended to mimic language emergence among animals.

The greater efficiency of STGS over REINFORCE relates to the general principle that reward in stochastic RL environments exhibits high variance, leading to sample inefficiency, unstable learning, and problems with reproducibility (Bjorck et al. 2021). This is particularly the case for non-stationary environments such as in multi-agent settings where the dynamics of an environment, from the perspective of any one agent, change as all agents update their policies (Lowe, Wu, et al. 2017). This non-stationarity violates the i.i.d. assumption that allows neural networks to learn effectively and, in the worst case, can precipitate catastrophic forgetting (Zhang et al. 2022). Nevertheless, the field of emergent language research is strongly reliant on the REINFORCE algorithm and has rarely experimented with more sophisticated RL tools and algorithms for the purpose of stabilizing learning. Table 1 illustrates the dominance of REINFORCE and STGS in optimizing speaker agents. Chaabouni, Strub, et al. (2021) demonstrate that if the task complexity and agent population size of emergent communication experiments

is to scale effectively, additional techniques will be required to stabilize the RL process. They find that agents trained with a policy gradient algorithm are only able to achieve 76% generalization accuracy in a referential game with over 1000 distractors, but that >98% accuracy can be achieved by stabilizing the learning process with Kullback-Leibler regularization (Geist et al. 2019), a well-established technique from the RL literature.

5.2.2 Loss Function Design and Regularization. The degree to which the ANNs underlying emergent language agents are regularized may have a significant effect on the compositionality of emerging languages. Too much or too little regularization may each be detrimental. Kuciński, Kołodziej, et al. (2020) show that adding noise to the messages sent from the sender to the receiver agent in a Lewis game can be beneficial to the compositionality of emerged languages as measured by *topsim* and *conflict count*. Introducing noise to the communication channel improved compositionality up to a point, but too much noise was detrimental, suggesting that over-regularization of a communication channel should be avoided.

Rita, Tallec, et al. (2022) note that in Lewis games there are two sources of loss: "information loss", from uninformative messages by the sender, and "co-adaptation loss", from misunderstanding by the receiver. By designing a loss function to control the weighting of each source of loss, they demonstrate that minimizing co-adaptation loss is essential for producing a language that generalizes well. However, co-adaptation loss overfits much more quickly than information loss and more compositional languages can be achieved by regularizing the receiver to address this. The authors found that iterated learning (see Section 5.4.1) was the most effective way of regularizing co-adaptation loss, but that other forms of regularization of the receiver such as dropout, layer normalization and weight decay were also beneficial.

Hazra et al. (2023) are able to increase *topsim* and *posdis* in emerging languages by adding two additional terms to the reward function. The first additional term is based on the mutual information between symbols in a message and the concepts underlying its referent. The second represents the degree to which a sender's message influences the actions of a receiver agent in a multitask grid world environment, calculated based on simulating counterfactual messages and subsequent receiver agent trajectories. Both of these additional loss terms give a sender agent access to information we might not expect it to have in a more realistic environment, i.e. the true "concepts" underlying referents and the extent to which the receiver agent is acting upon received communication, but the authors show that languages shaped by these additional reward terms achieve higher compositionality and better generalization to novel combinations of concepts.

Vithanage et al. (2023) design additional loss terms for a reconstruction game, respectively representing approximations of the complements of *topsim* and the mutual information between symbols and concepts. They find that directly optimizing for compositionality in this way makes agents more sample-efficient and more consistently able to achieve high generalization accuracy, especially where the agents vocabularies were small or training data was limited.

Ueda and Taniguchi (2024) found that using an ELBO loss based on a language model prior, i.e. penalising agents whose next-symbol choices stray from the expected distribution, led to more defined apparent word boundaries in sequences of symbols sampled from the emergent languages according to Harris's articulation scheme (Harris 1955), and better *topsim* with respect to both symbols and detected "words". This additional loss was combined with the standard reward scheme of a referential game using an adjustable weight parameter β, in a manner similar to a β-VAE (Higgins et al. 2016).

5.3 Game and Reward Design

5.3.1 Characteristics of Lewis Games. S. Guo et al. (2020) find that languages emerging from a reconstruction game are more expressive (allow better reconstruction of referents) than those emerging from a referential game, but that languages emerging from referential games tend to have higher *topsim*.

Steinert-Threlkeld (2020) shows that in a referential game where message lengths are fixed and agents must communicate about a finite number of states from a discrete state space, the optimal strategy will lead to the emergence of a language where any compositionality is "trivial" in the sense that it merely consists of taking the union of meanings of the symbols being composed. Thus, more natural and sophisticated settings for referential games (e.g. Lazaridou, Hermann, et al. 2018) may be required for more sophisticated compositionality to emerge, such as the existence of function words.

5.3.2 Game Variants Encouraging Compositionality. Mu and Goodman (2021) find that referential games in which agents must refer to a set of objects with common properties (rather than a single object), from within a superset of objects including distractors, helps agents to learn to describe properties of objects. This is called the "setref game". They also experiment with a version of the setref game where agents see objects with the same colors and shapes but do not have the same view of the objects, i.e. they may see them in different positions or in different sizes, and this is called the "concept game". The authors find that having to refer to sets of objects with shared characteristics improves compositionality as measured by *topsim*, and that referring to sets of objects when agents' viewpoints are different offers a further boost.

Ohmer et al. (2022) take inspiration from Grice's maxim of quantity Grice (1975), i.e. that a conversation participant should not give more information than is necessary, and design a referential game in which agents must learn to communicate in various levels of detail in order for the receiver agent to correctly identify a referent among a set of distractors. The sender agent is provided with a disentangled, symbolic input vector encoding attributes and a relevance vector indicating which attribute(s) will distinguish the target from the distractors. The receiver agent only sees the target and the distractors. In this setting a compositional language emerges (proven by inspection of the messages), in which sender agents learn to communicate only about relevant concepts and even to indicate to the receiver which attributes of the target object are irrelevant.

5.3.3 Multitask Learning. The benefit of multitask learning to the emergence of compositional language is possibly best understood with reference to the formalization of disentangled representation learning by Ren et al. (2023):

(1) A learner will receive observations $x \in X$ of an environment, which can be conceptualized as being generated by a deterministic generating function $f_x(\mathbf{G})$ operating on the values of some generating factors $\mathbf{G} = [G_1, G_2, ...G_m]$.
(2) For any given observation, the values of $[G_1, G_2, ...G_m]$ can be seen as having been sampled from some multivariate distribution $P(\mathbf{G})$.
(3) There will be a task to predict a label y for each x, where the label too should be thought of as having been generated by a second generating function, $f_y(\mathbf{G}, \epsilon)$ operating on $\mathbf{G}$ and some independent noise ϵ. The term "label" can be interpreted generally, e.g. identifying the most rewarding action y for an agent to take given an observation of its environment x can be cast as a labeling task.
(4) Any labeling function may only take into account a relevant subset of the factors of variation in an environment, i.e. $f_y(\mathbf{G}, \epsilon)$ may be invariant to changes in the factors of variation $\mathbf{G}$ other than in some subset $\mathbf{G_{relevant}}$.
(5) In the general paradigm of machine learning as function approximation (Vapnik 1999), the task of the learner is to learn to approximate $p(y|x)$.
(6) In the disentangled representation learning task, the learner must learn to factor $p(y|x)$ into $p_{\sim P}(\mathbf{G}|x)$, the relationship between the observed state of the environment and the values taken by the explanatory factors[12], and $p(y|\mathbf{G})$, the distribution of relevance to a specific task in that environment.

[12]The subscript $\sim P$ indicates that we are talking about probability under the distribution P, so that $p_{\sim P}(y|x)$ is to be read as "The probability of y given x, under the distribution P".

(7) We then hope that a learner can compositionally generalize to some distribution of held-out examples $Q(\mathbf{G})$ if $x = f_x(\mathbf{G})$ and $y = f_y(\mathbf{G}, \epsilon)$ remain unchanged, even if the explanatory factors $\mathbf{G}$ appear in unseen combinations (i.e. if the support of P and Q is not the same).

By this formulation, we can think of multi-task learning in emergent communication experiments as a form of meta-learning (learning to learn). While the observations $x \in X$ of the agent are related to all the factors of variation in $\mathbf{G}$ by $x = f_x(\mathbf{G})$, each of a range of tasks may relate to only a subset of those factors of variation, $\mathbf{G}_{\mathbf{relevant}}$. If an agent has learned many tasks and factored the solution $p(y|x)$ to each into the corresponding $p(\mathbf{G}|x)$ and $p(y|\mathbf{G})$ per the formulation above, we might hope that the agent has learned a representation of $p(\mathbf{G}|x)$ that is shared across tasks and accounts for more factors in $\mathbf{G}$ than are relevant to any single task. General knowledge of $p(\mathbf{G}|x)$ can be transferred to a further unseen task more easily than learning the latter from scratch if $p(y|\mathbf{G})$ is easier to learn than $p(y|x)$. Similarly, a task can in principle be transferred to a new environment where $f_x(\mathbf{G})$ is different if $\mathbf{G}$ and $f_y(\mathbf{G}, \epsilon)$ remain the same.

Languages learned in environments with only one factor of variation may form the foundation for more complex languages. Korbak, Zubek, Kuciński, et al. (2021) take inspiration from the "self-assembling games" of Barrett and Skyrms (2017) to demonstrate that relatively complex communication games involving several factors of variation can be modeled as compositions of simpler games. The authors simultaneously train a receiver agent to communicate successfully with a sender agent referring to object color and a second sender agent referring to object shapes. When such a pretrained receiver agent is paired with an untrained sender agent in a referential game requiring transmission of both color and shape information, the new sender agent learns to emulate the communication ability of the sender agents for which the receiver has already been trained. Using this technique, emergent languages were more successful in generalization to unseen concepts, and more compositional in terms of *topsim* and *context-independence*, than obverter-style (Choi et al. 2018) agents that did not benefit from pretraining. In terms of generative factors, the support of the test distribution in this example is different from each of the training distributions, but is the Cartesian product of the games played with single factors of variation.

Perkins (2021c) finds some evidence that agents trained simultaneously on multiple tasks that require communication about different combinations of generative factors tend to produce more compositional languages in terms of *topsim*.

5.4 Communication Pressures Inspired by Real-World Communication

Because multi-agent emergent language research offers a chance to explore how language might have emerged for humans and why natural languages might have common characteristics, such as compositionality, many of the explored inductive biases intended to elicit compositionality in emergent languages are based on communication pressures that have analogues in the human world (Galke and Raviv 2024). Below we have categorized these efforts as relating to transmissibility pressure (the evolutionary pressure on languages to be transmissible to new generations of senders in order to survive), productivity pressure (the pressure to be able to express a rich domain of experience using a limited vocabulary of symbols), and differing viewpoints (agents having different views of referents but needing to communicate about them effectively).

5.4.1 Transmissibility Pressure. Several works draw inspiration from the insight in evolutionary linguistics that languages are under evolutionary pressure to develop in ways that make them transmissible to new learners, i.e. children, and that compositionality is a retained trait in natural languages as it is useful in this respect (Brighton 2002; Kirby 2001). Thus, several studies investigate the hypothesis that language games necessitating periodic transfer of an emerged language to new (i.e. untrained) "generations" of agents will add additional pressure for an emergent language to be compositional. After seminal work by Kirby (2001), this pattern is usually referred to as the "iterated learning model" (Lu et al. 2020; Swarup and Gasser 2009; Yi et al. 2019). Iterated learning involves periodically introducing untrained receiver agents, or untrained sender and receiver agents, to an experiment.

New agents will spend some time interacting with more experienced agents and may ultimately become teachers for future generations of agents.

Yi et al. (2019) implement iterated learning in a similar protocol to the one originally suggested by Kirby (2001). At each iteration, new agents are initialized, spend an amount of time learning from the agents in the previous generation, and then must continue to develop the language without supervision. The authors find that resetting both the sender and receiver agents is even better for the development of compositionality than only resetting receiver agents. Additionally, they find evidence for the validity of *topsim* as a measure of compositionality, showing that it strongly correlates with communication performance on a held-out set of referents.

F. Li and Bowling (2019) investigate whether it is more beneficial to compositionality to periodically reset a small or large percentage of the receiver agents in a population. They find that the benefit to compositionality increases with the percentage of receivers that is affected by each reset and is maximized by periodically resetting all receivers at once. Further, and aligned with the intuition from evolutionary linguistics, the authors find that more compositional languages are learned more quickly by new receiver agents.

Cogswell et al. (2019) repeatedly choose pairs of agents from among a population of agents to play the Task & Talk game of Kottur et al. (2017). By periodically removing experienced agents from the population and adding untrained agents in their place, the authors add additional pressure for a language to be easy to learn and find that this improves the compositionality of the emerged languages. Like Kottur et al. (2017), Cogswell et al. (2019) obtain this result based on emergent languages where all messages are one symbol long (though see Section 5.4.2 on the compositionality of single-symbol messages in the Task & Talk game). Dagan et al. (2020) further explore the idea of resetting individuals from a population of agents, additionally investigating the effects of allowing the architecture of the agents themselves to evolve. They remove the worst-performing agents from the population and replace them with mutated versions of the best-performing agents (using a simple neural architecture search approach), finding that the resulting populations of agents outperformed populations where agents were randomly replaced in terms of both task success and compositionality as measured by *topsim*. Perkins (2021b) finds evidence that iterated learning improves *topsim* in languages emerging from referential games based on images, but also observes that holdout accuracy (a standard way to check for systematic generalization) might sometimes be inversely correlated with *topsim*. This might be a case of under-specification, such that the agents in the experiment learned a compositional (by *topsim*) system of concepts that was different to the system required for generalization accuracy (a risk highlighted by the same author in Perkins 2021a, and discussed in Section 4.2 above) or might relate to a shortcoming of *topsim*.

Iterated learning appears to improve compositionality by at least two mechanisms. It was originally suggested that the information bottleneck of teaching a language to new agents would select for language features that, like compositionality, exhibited efficient compression of information (Brighton 2002; Kirby 2001). F. Li and Bowling (2019) and Chaabouni, Kharitonov, Bouchacourt, et al. (2020) both demonstrate that, among deep learning agents, more compositional languages are indeed easier for new learners to acquire. Adjacent to the emergent communication literature, Ren et al. (2023) explore the hypothesis that compressibility pressure leads to more compositional representations by using an iterated learning approach to train generations of models to embed images for downstream tasks. At each iteration, a new agent is trained to imitate the embeddings of the agent trained in the previous iteration before being further fine-tuned on a downstream task. Further compressibility pressure is applied by constraining the model's embeddings to be simplicial embeddings (SEM), per Lavoie et al. (2022). The authors find that iterated learning without the constraint of SEM did not significantly improve compositionality beyond a ResNet-18 (He et al. 2015) baseline, but both together produced significantly more compositional representations in terms of *topsim* and generalization accuracy. Alongside positively filtering for compositional representation by imposing compressibility pressure, Rita, Tallec, et al. (2022) find that iterated learning is an effective way of ameliorating overfitting between sender and receiver agents while the sender agent is still forming an informative language (as discussed in Section 5.2.2), and that this leads to more compositional

languages. Although other forms of regularization were also effective in this regard, iterated learning proved the most effective and is perhaps most easily tied to real world evolutionary pressures.

5.4.2 Productivity Pressure. Productivity is the quality of a language that a finite collection of meaningful symbols can be reused and recombined in a relatively large number of ways to express a large number of different concepts. This quality is shared by all natural languages by virtue of their compositionality and compositional languages are always productive. On this basis, one natural pressure often exerted on emergent language agents is the pressure to use a limited vocabulary to communicate a relatively large number of concepts.

Kottur et al. (2017) demonstrate that emergent languages will only exhibit compositionality if the agents' vocabularies are carefully constrained. The authors created a referential game ("Task & Talk") where two RNN agents repeatedly exchange single-symbol messages analogous to a question and answer format, such that the questioning agent ("Q-bot") builds an understanding of what the answering agent ("A-bot") has seen over multiple turns by composing the single-symbol messages received. Using a simple environment where referents are few-hot vectors encoding a combination of 3 qualities (metaphorically described as shape, style and color), each of which may take one of four values, the authors found that it was necessary to introduce information bottlenecks in order to encourage a language to be compositional. A-bots with over-complete vocabularies that could encode all ($4 \times 4 \times 4 = 64$) unique referents would simply overfit the training data and usually fail on unseen combinations of attributes. A-bots with only enough vocabulary to communicate the atomic qualities of the referents, i.e. having a word for each color/style/shape ($4 + 4 + 4 = 12$), would use the memory available in the hidden state of their underlying RNN to learn a language where the same symbol meant different things depending on what symbols they had already used. With symbols having different meanings depending on context, memoryful A-bots were able to stretch an under-complete vocabulary into an over-complete one and somewhat overfit the training data. Finally, A-bots with under-complete vocabularies whose hidden state was reset after each conversational round, akin to having their memory erased, were able to create languages that related one attribute to one symbol and generalised to unseen combinations of attributes in the majority of cases.

Mordatch and Abbeel (2017) and Luna et al. (2020) both experiment with adding terms to the reward function to encourage agents to use a more restricted vocabulary, achieving conflicting results: the former authors find that this is beneficial to the compositionality of emergent languages, while the latter find that it is detrimental. The additional reward term of Mordatch and Abbeel (2017), appropriate to use for games with one or more agents, is

$$\sum_{i,t,k} \mathbb{1}\left[c_i^t = c_k\right] \log p(c_k)$$

That is, the sum of log probabilities of every symbol $c_k \in C$, uttered by every agent i, at every time step t, where $p(c_k)$ (the probability that the n^{th} symbol will be c_k) is calculated as

$$p(c_k) = \frac{n_k}{\alpha + (n-1)}$$

I.e., a Dirichlet process where n_k is the number of times the symbol c_k has previously been used, $n-1$ is the number of times any symbol has previously been used, and α is a small constant value representing the probability of observing and out-of-vocabulary symbol. Thus, the additional reward term is the log likelihood of the used symbols under the assumption that symbols are distributed according to a Dirichlet process, so that senders get additional reward if they favor a relatively small subset of the symbols in their vocabulary. This allows agents the flexibility to use a broad vocabulary early in training to escape local minima, but ultimately encourages them to converge on a more compositional language.

Luna et al. (2020) use a "least effort pressure" penalty term that is designed to penalize both long messages and high entropy in the distribution over symbols. This is the cross entropy loss between a sender agent's distribution over possible next tokens in a message and a one-hot vector indicating the symbol that was chosen. This is

intended to elicit highly confident symbol choices (low entropy distributions over vocabulary when choosing symbols) and also has the side effect of reducing the length of messages, since the token signaling the end of a message is expected to be among the tokens towards which the sender agent is strongly biased. In contrast to the results of Mordatch and Abbeel (2017), it was found that this pressure to reduce vocabulary size and message length had a negative impact on the agents' ability to generalize to referents with previously-unseen combinations of familiar qualities (shape and color). The difference between the results of the two papers is likely to be in the pressure applied to agents to keep messages short: Mordatch and Abbeel (2017) apply no such pressure, whereas the auxiliary loss used by Luna et al. (2020) does. The two papers thus both accord with the findings of Denamganaï and Walker (2020), who show that increasing the representational capacity of emergent language agents by increasing the maximum sentence length is beneficial to compositionality, while increasing it by increasing the size of the available vocabulary is generally detrimental. Wal et al. (2020) also suggest that longer messages may allow for more structured languages[13], which may mean that the emergence of compositionality could be suppressed by measures put in place to encourage natural-language-like brevity in messages exchanged by agents (Chaabouni, Kharitonov, Dupoux, et al. 2019; Rita, Chaabouni, et al. 2020).

5.4.3 Differing Viewpoints. Entities to which we refer in natural languages have inherent qualities that are invariant across a range of viewpoints. For example, we can instantiate many views of a round, red object without altering its apparent shape and color. Thus, the studies described in this section aim to help agents disentangle viewpoint-invariant properties of referents from properties that are simply artifacts of the viewpoint of the sending or receiving agent.

Luna et al. (2020) find that greater compositionality emerges in languages required to communicate about shape and color when the sender and receiver have different views of an object (i.e. they perceive it as being in different positions so that the language they develop must be position invariant) or when they see objects that are the same but with slightly perturbed color (in particular having slightly different brightness).

Feng et al. (2023) find that if agents in a referential game see different views of referents and additionally have different noise added to their respective images, the emergent language is more compositional based on generalization ability and *topsim*.

Mu and Goodman (2021) find that referential games in which agents must refer to a set of objects with common properties (rather than a single object), from within a superset of objects including distractors, help agents to learn to describe properties of objects (the "setref" game). They also experiment with a version of the setref game where agents see objects with the same colors and shapes, but do not have the same view of the objects, i.e. they may see them in different positions or in different sizes (the "concept" game). The authors find that having to refer to sets of objects with shared characteristics improves compositionality as measured by *topsim*, and that referring to sets of objects when agents' viewpoints are different offers a further boost.

6 Concluding Remarks

In this section we summarize the findings above, highlighting promising avenues for future research. In particular, we call upon researchers to demonstrate successful emergent communication among agents with larger policy models, find ways of measuring compositionality that work at all scales and ideally do not require knowledge of the ground truth concepts in an environment, and to study whether the (so far) discovered means of eliciting compositionality in emergent language experiments are robust to various changes in the experimental setup.

[13]The structure of messages is one of the sources of meaning in a compositional language as defined by Frege's principle.

6.1 Motivation for Studying the Emergence of Compositional Languages Between Deep Learning Agents

Language is a defining feature of the human experience and all natural languages are compositional. This means that a message in a natural language is made up of meaningful parts, which are carefully arranged in order to represent a particular meaning. Natural languages are grounded in the human world and many words in natural languages (i.e. content words, not function words) denote and connote concepts that are largely intended to be part of a shared understanding of the world between conversational participants (Grice 1975). The way that words and their attendant concepts can be almost endlessly rearranged and reused to represent complex shared experiences is reminiscent of the fact that humans naturally view the world as being made up of composable parts. We perceive the world as consisting in entities, qualities and relationships that can, and will, be remixed in learnable patterns. Understanding when one situation is partially similar to another in nameable ways appears to allow humans to learn from experience more efficiently than machine learning algorithms (Greff et al. 2020; Lake and Baroni 2018). Helping ANNs to learn world representations that are compositions of discrete, reusable concepts in order to facilitate more symbolic, more human-like reasoning has been called the "binding problem in artificial neural networks" and is deeply related to ideas about reconciling connectionist and symbolic approaches to artificial intelligence. Agents able to process sub-symbolic input but reason in interpretable and verifiable ways would represent the best of both worlds. Deep learning agents that create compositional emergent languages do so by disentangling observations of their environment, and so are an available "model organism" for how this can happen in practice.

Large language models (LLMs) have had extraordinary success in recent years, on a range of problems inside and outside natural language processing. However, they have shortcomings in terms of interpretability, fairness and factual reliability, which may not be addressable through further model scaling. As well as helping us understand how neural networks can learn compositional symbol-forming and symbol-relating behaviors, agents in emergent language experiments have a bias towards informativeness and interpretability that could complement the obvious strengths of LLMs. Thus, there have been efforts to understand whether the strengths of these two contrasting approaches can be usefully combined (Steinert-Threlkeld et al. 2022) or reconciled into a single theoretical framework (Taniguchi et al. 2024). Perhaps insights from emergent language research can inform us about how to make more efficient and effective foundational models, as suggested by Galke and Raviv (2024). If so, it is possible that the insights we need are uncoverable with only modest resources, as most interesting results in emergent language research so far have been achieved with models with numbers of parameters only in the millions, which is several orders of magnitude smaller than modern LLMs.

The motivations for emergent language research are thus broad, and we hope that this survey has succeeded in presenting it as an interesting and exciting area of deep learning research that is widely accessible and likely to support progress in other areas. The main sections in this survey identified ways of encouraging and measuring compositionality, which is a desirable quality in terms of the value of emergent language research as outlined above. In the later sections of this conclusion we make concrete recommendations regarding the most promising directions of future research in both of these areas.

6.2 The Challenge of Defining and Measuring Compositionality

Frege's principle is the standard definition of compositionality: the meaning of a compositional message is a function of the meanings of its meaningful components and their mode of composition. This definition has provided abundant inspiration for the design of metrics to measure the compositionality of emergent languages. However, to judge whether an emergent language fits this definition we must be able to compare the meanings of its messages with the meanings of their constituent symbols. In practice, neither of these comparators is easy to discern. Most approaches to measuring compositionality in emergent languages are akin to field linguistics:

with access to the concepts inherent in a referent (e.g. blue, triangle) we examine the extent to which certain symbols or combinations of symbols appear to have mutual information with a concept.

In cases where we know the concepts that are the objective, ground truth factors of variation in referents, we can express the presence or absence of these concepts in a few-hot vector and examine the extent to which the symbols and/or symbol order in a message can help us predict what concepts are present. This is the unifying idea between the metrics (discussed in Section 4) of *consistency and effectiveness* (Ohmer et al. 2022), *context independence* (Bogin et al. 2018), *bosdis*/*posdis* (Chaabouni, Kharitonov, Bouchacourt, et al. 2020), *residual entropy* (Resnick et al. 2020), *human compositional entropy* (Perkins 2021a) and *conflict count* (Kuciński, Kołodziej, et al. 2020). *TRE* (Andreas 2019) and *ACRe* (ACRe, Mu and Goodman 2021) measure the extent to which we can learn a strictly compositional way to model the representations learned by our agents, with the similar assumption that we have access to a few-hot vector of concepts. Thus, the majority of measures of compositionality advanced by the literature to date depend on having either symbolic referents or knowledge of what symbol-like concepts underpin referents. Since more realistic communication scenarios, i.e. ones that are more like how animals communicate, will need to take into account that different agents might conceptualize the world in different ways, more research is urgently needed into measures of compositionality that do not assume knowledge of an objective set of concepts for each referent.

Topsim, the most popular metric for measuring compositionality outside of generalization performance, originally relied on access to the concepts underlying referents (Lazaridou, Hermann, et al. 2018), but there have been some efforts to work around this. Three papers in this survey use *topsim* to measure the compositionality of languages grounded in continuous, entangled input for which the ground truth concepts are not known. Garcia et al. (2022) use agents' internal representations of referents as the semantic vector required to calculate *topsim* (see Section 4.3), which is a principled choice considering the finding of Bouchacourt and Baroni (2018) that agents' representations of referents will quickly diverge from the way referents are presented at input. The two other studies that use *topsim* based on calculating the semantic distance between embeddings, Rita, Tallec, et al. (2022) and Chaabouni, Strub, et al. (2021), use embeddings from pretrained computer vision models as both referents and semantic vectors. *Topsim* does not work as well in these studies, suggesting that the approach of Garcia et al. (2022) is preferable.

Given the need to measure compositionality in increasingly sophisticated emergent languages, grounded in environments increasingly like the natural world, we advocate that future research converges on *topsim* and generalization as the standard approaches to measuring compositionality, but also works to expand *topsim* into a family of metrics that can detect compositionality in languages displaying forms of linguistic variation such as free word order and synonymy. Concretely, the standard formulation of *topsim* is brittle where such variation is present due to its use of Levenshtein similarity and there is an opportunity to experiment with other similarity/distance measures. For example, Jaccard/Tanimoto similarity could be used in a word-order-invariant variant of *topsim*.

6.3 Encouraging Compositionality Through Experiment Design

Neural network architecture is a major source of useful inductive biases in deep learning, but is minimally explored in work studying compositionality in emergent languages. The obverter architecture (Choi et al. 2018) and the related context-consistent obverter (Bogin et al. 2018) stand out as obvious successes in designing agent architectures with a strong inductive bias for compositionality. However, it remains to be seen whether the benefits of obverter-type models extends to larger data sets and model sizes, and many studies achieve good results in terms of compositionality with agents based on standard architectures. There is an opportunity here for researchers to explore using agents based on more recent architectures that perform well on generating natural language, such as Transformer (Vaswani et al. 2017) or state-space models (Patro and Agneeswaran

2024), or those that have performed well in small image data sets but are under-explored in emergent language research, such as hybrid CNN-transformer vision models (e.g. Hassani et al. 2022). Relatively under-explored is the extent to which the differing expressive power of different neural network architectures (Deletang et al. 2023) might affect the ability of agents based on these architectures to create compositional languages. Outside of the emergent language literature, explorations of the ability of different architectures to learn other compositional tasks (Hupkes et al. 2020; Perkins 2021a) indicates that this would be a fruitful direction for future research.

Compositionality and disentangled representation learning are closely related concepts and it has been shown that basing agents on neural network architectures designed for the latter can help compositional languages to emerge (Garcia et al. 2022). Having sender and receiver agents with different architectures may make it difficult for expressive languages to emerge but seems to improve compositionality (Garcia et al. 2022). Perhaps more important is to ensure that compositional representations by the sender are not converted back to entangled representations by the receiver as a result of the receiver's architecture (Garcia et al. 2022; Perkins 2021b). There is a hypothesis that architectures that spatially or temporally separate an input into segments may help compositional representations to form (Greff et al. 2020) and some supporting evidence for this in that convolutional neural networks appear to help agents form compositional languages grounded in images (Kuciński, Korbak, et al. 2021).

In terms of optimization algorithms, supervised learning causes languages to emerge more quickly by dint of its better stability and does not seem to prevent compositional languages emerging. However, reinforcement learning may be preferred in settings where realism is important, e.g. in simulations for the study of evolutionary linguistics, as it does not require errors to be backpropagated from agent to agent in ways that are not possible among animals. Reinforcement learning may also be useful in that the mode-seeking, reverse Kullback-Leibler divergence objective of reinforcement learning appears to help agents choose more learnable languages in a setting where multiple languages are available. The literature to date has strongly relied on the REINFORCE algorithm where reinforcement learning is used and the straight-through Gumbel softmax trick where supervised learning is used (see table 1). Other techniques may yet be shown to be effective and there exist examples (Bogin et al. 2018; Chaabouni, Strub, et al. 2021; S. Guo 2019) where techniques from the wider reinforcement learning literature have allowed more stable and reliable language emergence.

The benefit of regularization techniques, especially in regularizing the receiver, is clear. Rita, Tallec, et al. (2022) give a compelling account of why this might be the case. They point out that, in the multi-agent emergent communication setting, loss may come from poor representation of referents by the sender or poor alignment between agents. It appears that alignment between agents is typically more easily learned than good disentangled representations, so making "co-adaptation" more difficult is effective in improving the compositionality of emergent languages. Among all the forms of receiver regularization available, the most effective and well-explored is iterated learning, in which receivers are periodically replaced with untrained receivers. This is a technique from the evolutionary linguistics literature, based on the idea that the pressure for languages to be transmissible to new learners encourages compositionality (Kirby 2001). Other supplemental loss or reward terms, whether to exert pressures on agents similar to those that might exist for human speakers (Luna et al. 2020; Mordatch and Abbeel 2017) or to directly reward compositionality (Hazra et al. 2023; Vithanage et al. 2023) have also proven effective. Over-regularizing receiver agents appears to be detrimental (Kuciński, Kołodziej, et al. 2020).

Alongside transmissibility pressure, other effects and pressures inspired by real world communication also seem effective. It is helpful for compositionality if agents communicating about a referent have different views of it (Feng et al. 2023; Luna et al. 2020; Mu and Goodman 2021), and it is also helpful if the agents are independently able to focus their attention (Ri et al. 2023). This suggests that agent architectures using attention, such as the popular transformer (Vaswani et al. 2017), might warrant further exploration. Multitask emergent language games, where each task features a different underlying distribution among the learnable factors of variation in referents, seem conducive to compositionality (Perkins 2021c). Perhaps the most important pressure after

transmissibility pressure in the real world is productivity pressure, the pressure to communicate a wide variety of meanings using a relatively small set of symbols (Denamganaï and Walker 2020; Kottur et al. 2017; Mordatch and Abbeel 2017). Increasing agents' number of parameters (model size) and available message length is beneficial to compositionality (Resnick et al. 2020). However, this can be undermined by affording agents a too-large vocabulary (Kottur et al. 2017; Mordatch and Abbeel 2017).

At the time of writing, the state of the art on many machine learning benchmarks is represented by models that achieve high accuracy by scaling up model size and training procedures to unprecedented levels. In this context, one notable thing about emergent language research is that it has mostly used agents based on relatively small models operating in simple environments. The observation of Chaabouni, Strub, et al. (2021) that "the emergent communication field mostly relies on small-scale games where only one speaker and one listener communicate about disentangled stimuli, which can hinder the generality of its conclusions" remains accurate. They warn that interventions that appear to improve compositionality in simple settings may be ineffective in more complex settings. Nevertheless, more complex settings in terms of referent complexity or number of agents remain largely unexplored. In terms of model sizes, an exploration of the code provided with the works surveyed above leads us to estimate that most of the speaker agents used in the research surveyed here have fewer parameters than e.g. MobileViT (Mehta and Rastegari 2022), a recent Vision Transformer (Dosovitskiy et al. 2021) model designed to be small enough to run on mobile devices (see Section 8 for a link to the code used). The emergent communication literature to date has produced many intriguing results and there is now a need to investigate whether these are robust to various changes in the experimental setup, including various forms of scaling and increasing complexity. We thus call upon researchers to demonstrate successful emergent communication among agents with larger policy models. This could be achievable by making use of model architectures, particularly the Transformer (Vaswani et al. 2017) that have been demonstrated to scale well but can also be adapted to learn effectively from small datasets (Hassani et al. 2022). Other architectures that perform well on language-related tasks at scale, for example state space models (Patro and Agneeswaran 2024), should also be considered. Previous experiments at encouraging compositionality in emergent language (section 5) have tended to be on a small scale. If experimental design choices that encourage the emergence of compositional behavior are also applicable at a larger scale, this is a promising indication that insights from emergent language research can be used to support modern approaches to AI involving extremely large models and data sets.

7 Author Contribution Statement

Nicholas Bailey: conceptualization, methodology, project administration, data curation, investigation, software, formal analysis, writing – original draft, writing – review & editing. **Chris Child:** conceptualization, methodology, supervision, writing – review & editing. **Tillman Weyde:** conceptualization, methodology, supervision, writing – review & editing.

8 Code Availability Statement

The code used to get an overall sense of the size of speaker agents in recent emergent language research is available at https://github.com/nicholasbailey87/2024_07_08_speaker_agent_sizes at the time of writing. The authors will not actively maintain the code in future and cannot guarantee its future availability or functionality.

9 Data Availability Statement

Approximate sizes for speaker agents used in works reviewed in this survey can be generated by running the provided code (see Section 8). All other data brought together during the course of this survey is included in the tables throughout the publication.

10 Competing Interests Statement

The authors have no financial or proprietary interests in any material discussed in this article.

11 Funding Statement

No funds, grants, or other support was received.

12 Ethical Approval Statement

As secondary research into a low-risk topic, not involving human or animal subjects and not using identifiable personal data, this survey did not require ethical review.

References

J. Andreas. 2019. "Measuring compositionality in representation learning." *arXiv preprint arXiv:1902.07181*.

M. Auersperger and P. Pecina. 2022. "Defending Compositionality in Emergent Languages." *arXiv preprint arXiv:2206.04751*.

S. Bader and P. Hitzler. 2005. *Dimensions of Neural-symbolic Integration - A Structured Survey*. (2005). arXiv: cs/0511042 (cs.AI).

J. A. Barrett and B. Skyrms. 2017. "Self-assembling games." *The British Journal for the Philosophy of Science*.

J. Batali. 1998. "Computational Simulations of the Emergence of Grammar." In: *Approaches to the Evolution of Language: Social and Cognitive bases*. Ed. by J. Hurford, C. Knight, and M. Studdert-Kennedy. Cambridge University Press, Cambridge, 405–426.

E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '21). Association for Computing Machinery, Virtual Event, Canada, 610–623. ISBN: 9781450383097. doi:10.1145/3442188.3445922.

Y. Bengio, A. Courville, and P. Vincent. 2013. "Representation learning: A review and new perspectives." *IEEE transactions on pattern analysis and machine intelligence*, 35, 8, 1798–1828.

A. Birhane, S. Han, V. Boddeti, S. Luccioni, et al.. 2023. "Into the LAION's Den: Investigating Hate in Multimodal Datasets." In: *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

C. M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg. ISBN: 0387310738.

J. Bjorck, C. P. Gomes, and K. Q. Weinberger. 2021. "Is High Variance Unavoidable in RL? A Case Study in Continuous Control." In: *International Conference on Learning Representations*.

B. Bogin, M. Geva, and J. Berant. 2018. "Emergence of communication in an interactive world with consistent speakers." *arXiv preprint arXiv:1809.00549*.

D. Bouchacourt and M. Baroni. 2018. *How agents see things: On visual representations in an emergent language game*. (2018). arXiv: 1808.10696 (cs.CL).

N. Brandizzi. 2023. "Toward More Human-Like AI Communication: A Review of Emergent Communication Research." *IEEE Access*, 11, 142317–142340.

H. Brighton. 2002. "Compositional syntax from cultural transmission." *Artificial life*, 8, 1, 25–54.

H. Brighton and S. Kirby. 2006. "Understanding linguistic evolution by visualizing the emergence of topographic mappings." *Artificial life*, 12, 2, 229–242.

S. Bubeck et al.. 2023. "Sparks of artificial general intelligence: Early experiments with gpt-4." *arXiv preprint arXiv:2303.12712*.

G. V. Carbajal and M. S. Malmierca. 2018. "The neuronal basis of predictive coding along the auditory pathway: from the subcortical roots to cortical deviance detection." *Trends in Hearing*, 22, 2331216518784822.

R. Chaabouni, E. Kharitonov, D. Bouchacourt, E. Dupoux, and M. Baroni. 2020. "Compositionality and Generalization In Emergent Languages." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4427–4442.

R. Chaabouni, E. Kharitonov, E. Dupoux, and M. Baroni. 2019. "Anti-efficient encoding in emergent communication." *Advances in Neural Information Processing Systems*, 32.

R. Chaabouni, F. Strub, et al.. May 2021. "Emergent communication at scale." In: *International Conference on Learning Representations*. (May 2021).

A. Chen, R. Shwartz-Ziv, K. Cho, M. L. Leavitt, and N. Saphra. 2024. "Sudden Drops in the Loss: Syntax Acquisition, Phase Transitions, and Simplicity Bias in MLMs." In: *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=MO5PiKHELW.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. 2020. "A simple framework for contrastive learning of visual representations." In: *International conference on machine learning*. PMLR, 1597–1607.

E. Cheng, M. Rita, and T. Poibeau. 2023. *On the Correspondence between Compositionality and Imitation in Emergent Neural Communication*. (2023). arXiv: 2305.12941 (cs.CL).

E. C. Cherry. 1953. "Some experiments on the recognition of speech, with one and with two ears." *The Journal of the acoustical society of America*, 25, 5, 975–979.

K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Oct. 2014. "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by A. Moschitti, B. Pang, and W. Daelemans. Association for Computational Linguistics, Doha, Qatar, (Oct. 2014), 1724–1734. doi:10.3115/v1/D14-1179.

E. Choi, A. Lazaridou, and N. De Freitas. 2018. "Compositional obverter communication learning from raw visual input." *arXiv preprint arXiv:1804.02341*.

F. Chollet. 2019. "On the Measure of Intelligence." *CoRR*, abs/1911.01547. http://arxiv.org/abs/1911.01547 arXiv: 1911.01547.

A. Clark. 2013. "Whatever next? Predictive brains, situated agents, and the future of cognitive science." *Behavioral and brain sciences*, 36, 3, 181–204.

M. Cogswell, J. Lu, S. Lee, D. Parikh, and D. Batra. 2019. "Emergence of compositional language with deep generational transmission." *arXiv preprint arXiv:1904.09067*.

H. Conklin and K. Smith. 2023. "Compositionality with Variation Reliably Emerges in Neural Networks." In: *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=-Yzz6vlX7V-.

A. D'Amour et al.. 2020. *Underspecification Presents Challenges for Credibility in Modern Machine Learning*. (2020). doi:10.48550/ARXIV.2011.03395.

G. Dagan, D. Hupkes, and E. Bruni. 2020. "Co-evolution of language and agents in referential games." *arXiv preprint arXiv:2001.03361*.

F. P. De Lange, M. Heilbron, and P. Kok. 2018. "How do expectations shape perception?" *Trends in cognitive sciences*, 22, 9, 764–779.

DeepSeek-AI et al.. 2025. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. (2025). https://arxiv.org/abs/2501.12948 arXiv: 2501.12948 (cs.CL).

G. Deletang et al.. 2023. "Neural Networks and the Chomsky Hierarchy." In: *The Eleventh International Conference on Learning Representations*.

K. Denamganaï and J. A. Walker. 2020. *On (Emergent) Systematic Generalisation and Compositionality in Visual Referential Games with Straight-Through Gumbel-Softmax Estimator*. (2020). arXiv: 2012.10776 (cs.CL).

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. "Imagenet: A large-scale hierarchical image database." In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

L. Deng. 2012. "The mnist database of handwritten digit images for machine learning research." *IEEE Signal Processing Magazine*, 29, 6, 141–142.

C. Devaraj, A. Chowdhury, A. Jain, J. R. Kubricht, P. Tu, and A. Santamaria-Pang. 2020. "From symbols to signals: symbolic variational autoencoders." In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3317–3321.

L. Donatelli and A. Koller. 2022. "Compositionality in Computational Linguistics." *Annual Review of Linguistics*, 9.

A. Dosovitskiy et al.. 2021. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. (2021). https://arxiv.org/abs/2010.11929 arXiv: 2010.11929 (cs.CV).

N. Dziri et al.. 2023. *Faith and Fate: Limits of Transformers on Compositionality*. (2023). arXiv: 2305.18654 (cs.CL).

S. Edunov, A. Baevski, and M. Auli. 2019. "Pre-trained language model representations for language generation." *arXiv preprint arXiv:1903.09722*.

Y. Feng, B. An, and Z. Lu. 2023. *Learning Multi-Object Positional Relationships via Emergent Communication*. (2023). arXiv: 2302.08084 (cs.LG).

A. L. Ferry, S. J. Hespos, and D. Gentner. 2015. "Prelinguistic relational concepts: Investigating analogical processing in infants." *Child development*, 86, 5, 1386–1405.

L. Ficco, L. Mancuso, J. Manuello, A. Teneggi, D. Liloia, S. Duca, T. Costa, G. Z. Kovacs, and F. Cauda. 2021. "Disentangling predictive processing in the brain: a meta-analytic study in favour of a predictive network." *Scientific Reports*, 11, 1, 16258.

J. A. Fodor and E. Lepore. 2002. *The compositionality papers*. Oxford University Press.

J. Foerster, I. A. Assael, N. De Freitas, and S. Whiteson. 2016. "Learning to communicate with deep multi-agent reinforcement learning." *Advances in neural information processing systems*, 29.

M. C. Frank. 2023. "Bridging the data gap between children and large language models." *Trends in Cognitive Sciences*. doi:https://doi.org/10.1016/j.tics.2023.08.007.

L. Galke, Y. Ram, and L. Raviv. 2024. "Deep neural networks and humans both benefit from compositional language structure." *Nature Communications*, 15, 1, 10816.

L. Galke and L. Raviv. 2024. "Learning and communication pressures in neural networks: Lessons from emergent communication." *arXiv preprint arXiv:2403.14427*.

A. d'Avila Garcez, S. Bader, H. Bowman, L. C. Lamb, L. de Penning, B. Illuminoo, H. Poon, and C. G. Zaverucha. 2022. "Neural-symbolic learning and reasoning: A survey and interpretation." *Neuro-Symbolic Artificial Intelligence: The State of the Art*, 342, 1, 327.

A. d'Avila Garcez, L. C. Lamb, and D. M. Gabbay. 2009. *Neural-symbolic learning systems*. Springer.

W. Garcia, H. Clouse, and K. Butler. July 2022. "Disentangling Categorization in Multi-agent Emergent Communication." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. (July 2022), 4523–4540.

M. Garnelo and M. Shanahan. 2019. "Reconciling deep learning with symbolic artificial intelligence: representing objects and relations." *Current Opinion in Behavioral Sciences*, 29, 17–23.

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

M. Geist, B. Scherrer, and O. Pietquin. 2019. "A theory of regularized markov decision processes." In: *International Conference on Machine Learning*. PMLR, 2160–2169.

D. Gentner. 1983. "Structure-mapping: A theoretical framework for analogy." *Cognitive science*, 7, 2, 155–170.

D. Gentner and F. Maravilla. 2017. "Analogical reasoning." In: *International handbook of thinking and reasoning*. Routledge, 186–203.

A. Goyal and Y. Bengio. 2020. "Inductive Biases for Deep Learning of Higher-Level Cognition." *CoRR*, abs/2011.15091. https://arxiv.org/abs/2011.15091 arXiv: 2011.15091.

K. Greff, S. Van Steenkiste, and J. Schmidhuber. 2020. "On the binding problem in artificial neural networks." *arXiv preprint arXiv:2012.05208*.

H. P. Grice. 1975. "Logic and conversation." In: *Speech acts*. Brill, 41–58.

J.-B. Grill et al.. 2020. "Bootstrap your own latent-a new approach to self-supervised learning." *Advances in neural information processing systems*, 33, 21271–21284.

N. M. Guerreiro, D. Alves, J. Waldendorf, B. Haddow, A. Birch, P. Colombo, and A. F. Martins. 2023. "Hallucinations in large multilingual translation models." *arXiv preprint arXiv:2303.16104*.

S. Guo. Nov. 2019. *Emergence of Numeric Concepts in Multi-Agent Autonomous Communication*. Master's thesis. Available at https://arxiv.org/pdf/1911.01098. Edinburgh, Scotland, (Nov. 2019).

S. Guo, R. Yi, A. Słowik, and K. Mathewson. 2020. "Inductive bias and language expressivity in emergent communication." *arXiv preprint arXiv:2012.02875*.

A. Gupta, A. Słowik, W. L. Hamilton, M. Jamnik, S. B. Holden, and C. Pal. May 2020. "Analyzing structural priors in multi-agent communication." In: *Workshop on Adaptive and Learning Agents at AAMAS*. (May 2020).

A. Hafri and C. Firestone. 2021. "The perception of relations." *Trends in Cognitive Sciences*, 25, 6, 475–492.

Z. S. Harris. 1955. "From Phoneme to Morpheme." *Language*, 31, 2, 190–222. Retrieved June 27, 2024 from http://www.jstor.org/stable/411036.

A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi. 2022. *Escaping the Big Data Paradigm with Compact Transformers*. (2022). https://arxiv.org/abs/2104.05704 arXiv: 2104.05704 (cs.CV).

J. Haugeland. 1979. "Understanding Natural Language." *The Journal of Philosophy*, 76, 11, 619–632. Retrieved Feb. 25, 2023 from http://www.jstor.org/stable/2025695.

S. Havrylov and I. Titov. 2017. "Emergence of Language with Multi-agent Games: Learning to Communicate with Sequences of Symbols." *CoRR*, abs/1705.11192. http://arxiv.org/abs/1705.11192.

R. Hazra, S. Dixit, and S. Sen. 2023. "Intrinsically Motivated Compositional Language Emergence." *arXiv preprint arXiv:2012.05011*.

K. He, X. Zhang, S. Ren, and J. Sun. 2015. *Deep Residual Learning for Image Recognition*. (2015). arXiv: 1512.03385 (cs.CV).

I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. 2016. "beta-vae: Learning basic visual concepts with a constrained variational framework." In: *International conference on learning representations*.

S. Hochreiter and J. Schmidhuber. 1997. "Long short-term memory." *Neural computation*, 9, 8, 1735–1780.

J. Hoffmann et al.. 2022. *Training Compute-Optimal Large Language Models*. (2022). arXiv: 2203.15556 (cs.CL).

D. Hupkes, V. Dankers, M. Mul, and E. Bruni. 2020. "Compositionality decomposed: How do neural networks generalise?" *Journal of Artificial Intelligence Research*, 67, 757–795.

A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. 2019. "Adversarial Examples Are Not Bugs, They Are Features." In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc.

E. Jang, S. Gu, and B. Poole. 2016. "Categorical reparameterization with gumbel-softmax." *arXiv preprint arXiv:1611.01144*.

S. M. Jiménez-Zafra, R. Morante, M. Teresa Martín-Valdivia, and L. A. Ureña-López. Mar. 2020. "Corpora Annotated with Negation: An Overview." *Computational Linguistics*, 46, 1, (Mar. 2020), 1–52. eprint: https://direct.mit.edu/coli/article-pdf/46/1/1/1847769/coli_a_00371.pdf. doi:10.1162/coli_a_00371.

S. Kambhampati et al.. 2025. *Stop Anthropomorphizing Intermediate Tokens as Reasoning/Thinking Traces!* (2025). https://arxiv.org/abs/2504.09762 arXiv: 2504.09762 (cs.AI).

S. Kirby. 2001. "Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity." *IEEE Transactions on Evolutionary Computation*, 5, 2, 102–110.

T. Korbak, J. Zubek, Ł. Kuciński, P. Miłoś, and J. Rączaszek-Leonardi. 2021. "Interaction history as a source of compositionality in emergent communication." *Interaction Studies*, 22, 2, 212–243.

T. Korbak, J. Zubek, and J. Rączaszek-Leonardi. 2020. "Measuring non-trivial compositionality in emergent communication." *arXiv preprint arXiv:2010.15058*.

S. Kottur, J. M. Moura, S. Lee, and D. Batra. 2017. "Natural language does not emerge 'naturally' in multi-agent dialog." *arXiv preprint arXiv:1706.08502*.

T. Kouwenhoven, M. Peeperkorn, and T. Verhoef. 2025. "Searching for Structure: Investigating Emergent Communication with Large Language Models." In: *Proceedings of the 31st International Conference on Computational Linguistics*, 9977–9991.

V. Krakovna, J. Uesato, V. Mikulik, M. Rahtz, T. Everitt, R. Kumar, Z. Kenton, J. Leike, and S. Legg. 2020. "Specification gaming: the flip side of AI ingenuity." *DeepMind Blog*.

Ł. Kuciński, P. Kołodziej, and P. Miłoś. July 2020. "Emergence of compositional language in communication through noisy channel." In: *Language in Reinforcement Learning Workshop at ICML 2020*. (July 2020).

Ł. Kuciński, T. Korbak, P. Kołodziej, and P. Miłoś. 2021. "Catalytic role of noise and necessity of inductive biases in the emergence of compositional communication." *Advances in Neural Information Processing Systems*, 34, 23075–23088.

B. Lake and M. Baroni. 2018. "Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks." In: *International conference on machine learning*. PMLR, 2873–2882.

S. Lavoie, C. Tsirigotis, M. Schwarzer, A. Vani, M. Noukhovitch, K. Kawaguchi, and A. Courville. 2022. "Simplicial Embeddings in Self-Supervised Learning and Downstream Classification." In: *The Eleventh International Conference on Learning Representations*.

A. Lazaridou and M. Baroni. 2020. "Emergent multi-agent communication in the deep learning era." *arXiv preprint arXiv:2006.02419*.

A. Lazaridou, K. M. Hermann, K. Tuyls, and S. Clark. 2018. "Emergence of linguistic communication from referential games with symbolic and pixel input." *arXiv preprint arXiv:1804.03984*.

Y. LeCun. 2022. "A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27." *Open Review*, 62.

D. Lewis. 1969. *Convention: a philosophical study*. (1969).

F. Li and M. Bowling. 2019. "Ease-of-teaching and language structure from emergent communication." *Advances in Neural Information Processing Systems*, 32.

J. Li, X. Cheng, X. Zhao, J.-Y. Nie, and J.-R. Wen. 2023. "HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models." In: *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Y. Lian, A. Bisazza, and T. Verhoef. 2023. "Communication Drives the Emergence of Language Universals in Neural Agents: Evidence from the Word-order/Case-marking Trade-off." *Transactions of the Association for Computational Linguistics*, 11, 1033–1047.

Y. Lian, T. Verhoef, and A. Bisazza. 2024. "NeLLCom-X: A Comprehensive Neural-Agent Framework to Simulate Language Learning and Group Communication." In: *Proceedings of the 28th Conference on Computational Natural Language Learning*, 243–258.

R. Lowe, J. Foerster, Y.-L. Boureau, J. Pineau, and Y. Dauphin. 2019. "On the pitfalls of measuring emergent communication." *arXiv preprint arXiv:1903.05168*.

R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch. 2017. "Multi-agent actor-critic for mixed cooperative-competitive environments." *Advances in neural information processing systems*, 30.

Y. Lu, S. Singhal, F. Strub, A. Courville, and O. Pietquin. July 2020. "Countering language drift with seeded iterated learning." In: *International Conference on Machine Learning*. PMLR. (July 2020), 6437–6447.

D. R. Luna, E. M. Ponti, D. Hupkes, and E. Bruni. 2020. "Internal and external pressures on language emergence: least effort, object constancy and frequency." *arXiv preprint arXiv:2004.03868*.

C. J. Maddison, A. Mnih, and Y. W. Teh. 2016. "The concrete distribution: A continuous relaxation of discrete random variables." *arXiv preprint arXiv:1611.00712*.

G. Marcus. 2020. "The next decade in AI: four steps towards robust artificial intelligence." *arXiv preprint arXiv:2002.06177*.

S. Mehta and M. Rastegari. 2022. "MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer." In: *International Conference on Learning Representations*. https://openreview.net/forum?id=vh-0sUt8HlG.

D. Mihai and J. Hare. 2021. "Learning to draw: Emergent communication through sketching." *Advances in Neural Information Processing Systems*, 34, 7153–7166.

M. Mitchell. 2021. "Abstraction and analogy-making in artificial intelligence." *Annals of the New York Academy of Sciences*, 1505, 1, 79–101.

R. Montague. 1970. "Universal grammar." *Theoria*, 36, 3, 373–398.

I. Mordatch and P. Abbeel. 2017. "Emergence of Grounded Compositional Language in Multi-Agent Populations." *CoRR*, abs/1703.04908. http://arxiv.org/abs/1703.04908.

J. Mu and N. Goodman. 2021. "Emergent Communication of Generalizations." *Advances in Neural Information Processing Systems*, 34, 17994–18007.

P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever. 2019. *Deep Double Descent: Where Bigger Models and More Data Hurt*. (2019). arXiv: 1912.02292 (cs.LG).

R. Ngo. 2022. "The alignment problem from a deep learning perspective." *arXiv preprint arXiv:2209.00626*.

M. Noukhovitch, T. LaCroix, A. Lazaridou, and A. Courville. 2021. "Emergent Communication under Competition." In: *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems* (AAMAS '21). International Foundation for Autonomous Agents and Multiagent Systems, Virtual Event, United Kingdom, 974–982. ISBN: 9781450383073.

C. K. Ogden and I. A. Richards. 1923. "The meaning of meaning: A study of the influence of thought and of the science of symbolism."

X. Ohmer, M. Duda, and E. Bruni. 2022. "Emergence of hierarchical reference systems in multi-agent communication." *arXiv preprint arXiv:2203.13176*.

B. N. Patro and V. S. Agneeswaran. 2024. *Mamba-360: Survey of State Space Models as Transformer Alternative for Long Sequence Modelling: Methods, Applications, and Challenges*. (2024). https://arxiv.org/abs/2404.16112 arXiv: 2404.16112 (cs.LG).

H. Perkins. 2021a. "A Framework for Measuring Compositional Inductive Bias." *arXiv preprint arXiv:2103.04180*.

H. Perkins. 2021b. "Compositionality Through Language Transmission, using Artificial Neural Networks." *arXiv preprint arXiv:2101.11739*.

H. Perkins. 2021c. "TexRel: a Green Family of Datasets for Emergent Communications on Relations." *arXiv preprint arXiv:2105.12804*.

J. Peters, C. Waubert de Puiseau, H. Tercan, A. Gopikrishnan, G. A. Lucas de Carvalho, C. Bitter, and T. Meisen. 2025. "Emergent language: a survey and taxonomy." *Autonomous Agents and Multi-Agent Systems*, 39, 1, 1–73.

O. Press, M. Zhang, S. Min, L. Schmidt, N. A. Smith, and M. Lewis. 2022. "Measuring and Narrowing the Compositionality Gap in Language Models." *arXiv preprint arXiv:2210.03350*.

W. H. Press, W. T. Vetterling, S. A. Teukolsky, and B. P. Flannery. 1988. *Numerical Recipes in C: The Art of Scientific Computing*. (Second ed.). Cambridge University Press, London, England.

Ed. by H. van der Hulst. "6. Recursion and the infinitude claim." *Recursion and Human Language*. De Gruyter Mouton, Berlin, New York, 111–138. ISBN: 9783110219258. doi:doi:10.1515/9783110219258.111.

J. W. Rae et al.. 2021. "Scaling Language Models: Methods, Analysis & Insights from Training Gopher." *CoRR*, abs/2112.11446. https://arxiv.org/abs/2112.11446 arXiv: 2112.11446.

A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. 2022. *Hierarchical Text-Conditional Image Generation with CLIP Latents*. (2022). arXiv: 2204.06125 (cs.CV).

Y. Ren, S. Lavoie, M. Galkin, D. J. Sutherland, and A. Courville. 2023. "Improving Compositional Generalization using Iterated Learning and Simplicial Embeddings." In: *Thirty-seventh Conference on Neural Information Processing Systems*.

C. Resnick, A. Gupta, J. Foerster, A. M. Dai, and K. Cho. 2020. "Capacity, Bandwidth, and Compositionality in Emergent Language Learning." In: *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems* (AAMAS '20). International Foundation for Autonomous Agents and Multiagent Systems, Auckland, New Zealand, 1125–1133. ISBN: 9781450375184.

R. Ri, R. Ueda, and J. Naradowsky. 2023. *Emergent Communication with Attention*. (2023). arXiv: 2305.10920 (cs.CL).

M. Rita, R. Chaabouni, and E. Dupoux. 2020. ""LazImpa": Lazy and Impatient neural agents learn to communicate efficiently." In: *Proceedings of the 24th Conference on Computational Natural Language Learning*, 335–343.

M. Rita, C. Tallec, P. Michel, J.-B. Grill, O. Pietquin, E. Dupoux, and F. Strub. 2022. "Emergent Communication: Generalization and Overfitting in Lewis Games." *arXiv preprint arXiv:2209.15342*.

A. L. Roskies. 1999. "The binding problem." *Neuron*, 24, 1, 7–9.

M. K. Sarker, L. Zhou, A. Eberhart, and P. Hitzler. 2021. *Neuro-Symbolic Artificial Intelligence: Current Trends*. (2021). arXiv: 2105.05330 (cs.AI).

M. Shanahan and M. Mitchell. 2022. "Abstraction for deep reinforcement learning." *arXiv preprint arXiv:2202.05839*.

I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal. 2024. "AI models collapse when trained on recursively generated data." *Nature*, 631, 8022, 755–759. ISBN: 1476-4687. doi:10.1038/s41586-024-07566-y.

C. Silberer, V. Ferrari, and M. Lapata. 2016. "Visually grounded meaning representations." *IEEE transactions on pattern analysis and machine intelligence*, 39, 11, 2284–2297.

A. Slowik, A. Gupta, W. L. Hamilton, M. Jamnik, and S. B. Holden. 2020. "Towards Graph Representation Learning in Emergent Communication." *CoRR*, abs/2001.09063. https://arxiv.org/abs/2001.09063.

S. Smith et al.. 2022. "Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model." *arXiv preprint arXiv:2201.11990*.

S. Steinert-Threlkeld. 2020. "Toward the emergence of nontrivial compositionality." *Philosophy of Science*, 87, 5, 897–909.

S. Steinert-Threlkeld, X. Zhou, Z. Liu, and C. Downey. 2022. "Emergent communication fine-tuning (ec-ft) for pretrained language models." In: *Emergent Communication Workshop at ICLR 2022*.

A. Suglia, I. Konstas, and O. Lemon. 2024. "Visually grounded language learning: a review of language games, datasets, tasks, and models." *Journal of Artificial Intelligence Research*, 79, 173–239.

R. Sutton. Mar. 2019. "The bitter lesson." *Incomplete Ideas*, (Mar. 2019). http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

S. Swarup and L. Gasser. 2009. "The iterated classification game: A new model of the cultural transmission of language." *Adaptive Behavior*, 17, 3, 213–235.

Z. G. Szabó. Feb. 2012. "64 The case for compositionality." In: *The Oxford Handbook of Compositionality*. Oxford University Press, (Feb. 2012). ISBN: 9780199541072. eprint: https://academic.oup.com/book/0/chapter/350861452/chapter-ag-pdf/44422005/book_41264_section_350861452.ag.pdf. doi:10.1093/oxfordhb/9780199541072.013.0003.

T. Taniguchi, R. Ueda, T. Nakamura, M. Suzuki, and A. Taniguchi. 2024. *Generative Emergent Communication: Large Language Model is a Collective World Model*. (2024). https://arxiv.org/abs/2501.00226 arXiv: 2501.00226 (cs.AI).

M. Turpin, J. Michael, E. Perez, and S. R. Bowman. 2023. "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting." In: *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=bzs4uPLXvi.

R. Ueda, T. Ishii, and Y. Miyao. 2023. "On the Word Boundaries of Emergent Languages Based on Harris's Articulation Scheme." In: *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=b4t9_XASt6G.

R. Ueda and T. Taniguchi. 2024. "Lewis's Signaling Game as beta-VAE For Natural Word Lengths and Segments." In: *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=HC0msxE3sf.

V. Vapnik. 1999. *The nature of statistical learning theory*. Springer science & business media.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. "Attention Is All You Need." *CoRR*, abs/1706.03762. http://arxiv.org/abs/1706.03762 arXiv: 1706.03762.

P. Villalobos, J. Sevilla, L. Heim, T. Besiroglu, M. Hobbhahn, and A. Ho. 2022. *Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning*. (2022). arXiv: 2211.04325 (cs.LG).

K. Vithanage, R. Wijesinghe, A. Xavier, D. Tissera, S. Jayasena, and S. Fernando. Dec. 2023. "Accelerating language emergence by functional pressures." *PLOS ONE*, 18, 12, (Dec. 2023), 1–28. doi:10.1371/journal.pone.0295748.

C. Von Der Malsburg. 1994. "The correlation theory of brain function." In: *Models of neural networks: Temporal aspects of coding and information processing in biological systems*. Springer, 95–119.

K. Wagner, J. A. Reggia, J. Uriagereka, and G. S. Wilkinson. 2003. "Progress in the simulation of emergent communication and language." *Adaptive Behavior*, 11, 1, 37–69.

O. van der Wal, S. de Boer, E. Bruni, and D. Hupkes. 2020. "The grammar of emergent languages." *arXiv preprint arXiv:2010.02069*.

T. Wang, A. Roberts, D. Hesslow, T. Le Scao, H. W. Chung, I. Beltagy, J. Launay, and C. Raffel. 2022. "What language model architecture and pretraining objective works best for zero-shot generalization?" In: *International Conference on Machine Learning*. PMLR, 22964–22984.

T. Webb, K. J. Holyoak, and H. Lu. 2023. "Emergent analogical reasoning in large language models." *Nature Human Behaviour*, 1–16.

L. Weidinger et al.. 2021. "Ethical and social risks of harm from language models." *arXiv preprint arXiv:2112.04359*.

R. J. Williams. 1992. "Simple statistical gradient-following algorithms for connectionist reinforcement learning." *Machine learning*, 8, 229–256.

Z. Xu, M. Niethammer, and C. A. Raffel. 2022. "Compositional generalization in unsupervised compositional representation learning: A study on disentanglement and emergent language." *Advances in Neural Information Processing Systems*, 35, 25074–25087.

F. Xue, Y. Fu, W. Zhou, Z. Zheng, and Y. You. 2023. *To Repeat or Not To Repeat: Insights from Scaling LLM under Token-Crisis*. (2023). arXiv: 2305.13230 (cs.LG).

L. Xue, A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale, A. Roberts, and C. Raffel. 2022. "ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models." *Transactions of the Association for Computational Linguistics*, 10, 291–306.

S. Yao, M. Yu, Y. Zhang, K. R. Narasimhan, J. B. Tenenbaum, and C. Gan. 2021. "Linking Emergent and Natural Languages via Corpus Transfer." In: *International Conference on Learning Representations*.

R. Yi, S. Guo, M. Labeau, S. B. Cohen, and S. Kirby. 2019. "Compositional languages emerge in a neural iterated learning model." In: *International Conference on Learning Representations*.

T. Zhang, X. Wang, B. Liang, and B. Yuan. 2022. "Catastrophic interference in reinforcement learning: A solution based on context division and knowledge distillation." *IEEE Transactions on Neural Networks and Learning Systems*.

C. Zhu, M. Dastani, and S. Wang. 2024. "A survey of multi-agent deep reinforcement learning with communication." *Autonomous Agents and Multi-Agent Systems*, 38, 1, 4.

Received 30 October 2024; accepted 12 December 2025