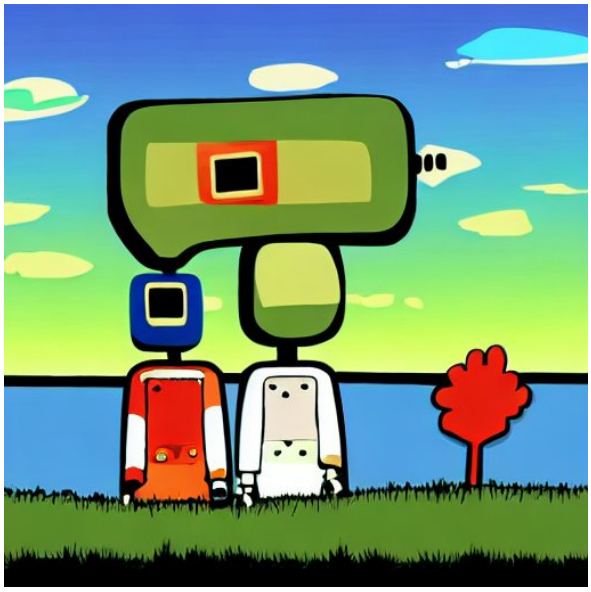
A holistic approach to artificial neural network models of language emergence and language acquisition

Case studies on pragmatic reasoning and language-perception interactions

Xenia Ohmer



Title image

The title image was created using a deep learning text-to-image model called Stable Diffusion, released by StabilityAI this year. The image was generated using the webpage https://huggingface.co/spaces/stabilityai/stable-diffusion and the prompt "Two robots walk in a landscape. One robot says something to the other robot."

Colophon

This document was typeset with the help of KOMA-Script and LaTeX using the `kaobook` class.
The source code of the `kaobook` template is available at: https://github.com/fmarotta/kaobook.

A holistic approach to artificial neural network models of language emergence and language acquisition

Case studies on pragmatic reasoning and language-perception interactions

Dissertation
zur Erlangung des Grades eines Doktors der Naturwissenschaften
eingereicht am Fachbereich Humanwissenschaften
der Universität Osnabrück

vorgelegt von

Xenia Ohmer

Osnabrück, 2022

Supervisors
Prof. Dr. Michael Franke & Prof. Dr. Peter König

Acknowledgements

I came to Osnabrück to start my Ph.D. about three and a half years ago. Back then I knew next to nothing about ongoing research in cognitive science and my only contact with linguistics had been years ago in a few introductory courses. Looking back, I cannot help but be amazed by all the things I have learned, the wonderful people I have met, and the small and big events that have made my time here so memorable.

The pandemic changed many things that I had taken for granted in scientific work. Researchers around the world missed out on the luxury of traveling to international conferences. More importantly, though, we missed out on the everyday experiences of in-person meetings, sharing an office, having lunch together, and random encounters in the hallway. Even despite these changes, my time as a Ph.D. student was a marvelous experience and I had lots of fun on the way, not least because of the support I received from family, friends, and colleagues.

First of all, I would like to express my deepest gratitude to Michael Franke and Peter König. I feel extremely grateful that two such amazing researchers and people supervised my work. Thank you for your enthusiasm, advice, support, unique perspectives, and discussions. Thank you for investing so much time in our projects and my supervision despite the many other things that require your attention. I cannot imagine two better supervisors.

I have also learned a great deal from Elia Bruni. Thank you for welcoming me into your study project and supervising the joint project with Marko Duda. I would also like to thank Marko himself, who was an extraordinary student. I enjoyed discussing all your creative research ideas and I still hope that you will end up doing a Ph.D. eventually.

Without attempting a list, I would also like to thank all the people and places that made Osnabrück so much more than just the place where I work. My biggest thank you goes to Britta Grusdt for not only being a great colleague but also a great friend. I will miss seeing you all the time! I would also like to thank Zenit for providing me with a second home and all my "climbing friends" for cheering each other on and many good conversations beyond the sport.

I would further like to thank Sören Becker, Britta Grusdt, Henrik Löfberg, and Charlotte Ohmer for reading (parts of) this thesis and providing helpful feedback.

Finally, a big thank you to my family, who supports me in everything I do. Thank you for the fun visits back and forth in the last years, going to festivals and on holidays together, playing lots of online and offline games, and sending baby photos on a regular basis.

Last but not least, I would like to thank Henrik Löfberg for moving to Germany for me, and then even to Osnabrück. Together, we will be happy anywhere.

Abstract

Humans interact with each other and with their environment. On the one hand, language *is shaped* by these interactions as we communicate about our experiences. On the other hand, language *shapes* these interactions as we use it to act in the world (by informing, requesting, promising, and so on). As a consequence, language is intertwined with many cognitive functions, such as perception, action, and social reasoning. To understand how language emerged and evolved, as well as how language is learned and used, it is important to take these connections into account. This thesis presents three case studies that investigate interfaces between language and other areas of cognition. The case studies use computational, agent-based models to study language acquisition and language emergence phenomena. We follow a long tradition of modeling interactive language use in the form of communication games. In all case studies, the agents play games that involve generating and/or interpreting references to objects, which is a simple but fundamental use of language. The agents are implemented as artificial neural networks. Artificial neural networks not only dominate machine learning research but are also popular as models of cognitive functions. We use the case studies to learn about cognitive mechanisms related to language as well as to suggest ways in which artificial neural networks may benefit from integrating such mechanisms. Case study 1 shows how reasoning about the speaker's intention and the context can help children learn the meanings of new words. Case study 2 demonstrates how considerations about the context can lead to the emergence of object references at different levels of specificity (as in "Fido", "dalmatian", "dog", "animal"). Case study 3 models aspects of the bidirectional influence between visual perception and (emergent) language. In general, I argue that the combination of communication games and artificial neural networks generates a versatile framework for studying language–cognition interfaces. A holistic approach to modeling language will increase our understanding of how humans learn and use language as well as our ability to emulate this behavior in machines.

Contents

List of Figures

List of Tables

1 General Introduction

1.1 Overview

Language is grounded in our experience of the world (Barsalou, 2008) and serves to coordinate with others (Clark, 1992; Lewis, 1969). The meaning of a word cannot solely be determined by its connections to other words. To avoid an infinite regress, the meaning of a word must ultimately be connected to reality (Harnad, 1990). Further, language is fundamentally used for interpersonal communication. While our experience of the world gives meaning to the building blocks of language, the meaning of an utterance can only be understood in relation to the agents involved, their intentions, and the context in which the utterance is made. Language is therefore intrinsically linked to other cognitive domains, including action and perception (Pecher & Zwann, 2005), social cognition (Holtgraves & Kashima, 2008), and emotion (Barrett et al., 2007). To understand how language emerges and evolves as well as how it is learned and used, it is important to take these connections into account. In this thesis, I seek to integrate insights from cognitive science, evolutionary linguistics, and artificial intelligence (AI) toward the goal of modeling language learning and language emergence phenomena in the broader context of cognition.

Cognitive science and AI research have long been working with computational models of grounded language, in which simulated or robotic agents develop their own communication system through interactions with each other and the world (Cangelosi, 2010; Lyon et al., 2007). These approaches have been used to study the link between language and cognition as well as to investigate how linguistic capabilities can be designed in artificial agents. A popular framework for simulating how agents learn or create a language are communication games (Cangelosi & Parisi, 2002; Franke & Wagner, 2014; Kirby et al., 2014; Lazaridou & Baroni, 2020; Steels, 1997). In these games, there is usually a population of agents with certain linguistic knowledge. The game requires the agents to exchange information through communication. The agents use feedback from the environment, such as rewards, to improve language production and comprehension. Communication games reduce the complexity of real linguistic interaction, allow for game-theoretic analysis, and can easily be combined with various frameworks for learning and evolution.

Artificial neural networks (ANNs) are a popular choice for modeling the agents in language learning and language emergence simulations (Cangelosi & Harnad, 2000; MacWhinney, 1998; Westermann & Twomey, 2017) because they can extract complex word-meaning associations directly from their experience with the environment and naturally simulate an incremental learning process. More recently, deep neural networks (DNNs), combined with the availability of large data sets and increasing computer power, have enabled experiments with raw pixel inputs (e.g., Lazaridou et al., 2018; Vong & Lake, 2020), simulated environments (e.g., Hermann

et al., 2017; Hill, Clark, et al., 2020), natural language (e.g., Hermann et al., 2017; Lazaridou et al., 2020), and large scale group interactions between relatively complex agents (e.g., Chaabouni et al., 2022; Harding Graesser et al., 2019). These developments unlock new possibilities for studying language learning and language emergence phenomena as well as for bridging the gap between small-scale, hand-crafted experiments and real-world applications.

Also the fields of natural language processing (NLP) and machine learning (ML) more broadly are recognizing the need to account for the grounded and functional nature of language (Bernardi et al., 2015; Bisk et al., 2020; McClelland et al., 2020). NLP research is currently dominated by DNN language models that are trained on massive amounts of text data scraped from the internet. These models achieve impressive performance on various tasks, such as translation, question answering, reading comprehension, or summarization (e.g., Devlin et al., 2019; Radford et al., 2019). However, language models are trained to predict a probability distribution over the vocabulary given some linguistic context, which means their language understanding is based on language *in isolation*. As a result, they can fail to capture grounded features of words (Lucy & Gauthier, 2017) or to perform experience-informed inferences (Peng et al., 2015), and generally display limited ability to understand situations (McClelland et al., 2020).

This thesis builds on efforts in cognitive science, AI, and evolutionary linguistics that use simulations with artificial agents to study language learning and language emergence. In particular, I present three case studies that investigate phenomena related to the physical or social context of linguistic experience. All case studies use a simple setup where artificial agents, implemented as ANNs, aim to successfully generate or interpret referential expressions. However, they go beyond traditional setups by incorporating pragmatic reasoning (case study 1), effects of context (case study 2), and interactions between language and perception (case study 3). The first case study examines the effect of pragmatic reasoning on word learning, focusing on a specific word learning bias. The second case study examines how considerations about the context can lead to the emergence of object references at different levels of specificity. The third case study is concerned with the mutual influence between language learning/emergence and the formation of perceptual representations. Together, the case studies demonstrate how specific language-related phenomena can be captured by taking a holistic stance on language, and at the same time, they suggest potential improvements to AI and ML models.

In the remainder of the introduction, I will elaborate on the aforementioned subjects. Section 1.2 explains in more detail what it means for language to be grounded and functional as well as how communication games can be used to simulate interactive language use. Section 1.3 briefly introduces ANNs and then reviews selected ANN models of language acquisition and language emergence. Section 1.4 introduces the case studies. Each case study is embedded into relevant background information and a short overview is provided. The case studies themselves will be reported in Chapters 2–4, with each chapter consisting of a lay summary followed by one of my publications. Finally, Chapter 5 provides a general discussion of this work, including an outlook and a conclusion.

1.2 Language in the world

Language is used by agents that interact with each other and with their environment. Language is *grounded* in these interactions, which means that words derive their meaning from sensorimotor, emotional, or other experiences in the world. Language is *functional* because it can be used in these interactions to bring about change in the world or in the minds of other agents. We can, for example, use language to provide information, express internal states, and make requests or suggestions. As a result, language is intertwined with other cognitive processes, such as perception and action, social reasoning, memory, and emotion. This section will provide more details on the grounded nature of language, focusing on the interface between language and perception/action, and the functional nature of language, focusing on the interface between language and socio-pragmatic reasoning. Finally, communication games will be discussed as a framework for simulating grounded and functional linguistic interaction.

1.2.1 Grounded language

In classical cognitive science, cognition was treated as symbol manipulation (Newell, 1980; Newell & Simon, 1976). Knowledge representations in the form of symbols were considered to be amodal and distinct from modal representations. In 1990, Harnad formalized the *symbol grounding problem* stating that treating cognition as symbol manipulation cannot explain how symbols derive their meaning:

> "How can the semantic interpretation of a formal symbol system be made intrinsic to the system, rather than just parasitic on the meanings in our heads? How can the meanings of the meaningless symbol tokens, manipulated solely on the basis of their (arbitrary) shapes, be grounded in anything but other meaningless symbols? The problem is analogous to trying to learn Chinese from a Chinese/Chinese dictionary alone." (Harnad, 1990, p. 1)

Similarly, treating language as an isolated system allows for descriptions of semantic similarity but cannot establish a link between words and the objects, actions, or experiences in the world that they refer to. In line with a general trend at the time, Harnad (1990) proposed that any knowledge representation must be grounded bottom-up through sensory inputs.

The research program of *embodied cognition* is based on the idea that the body or the body's interactions with the environment play an important role in many features of cognition (for a review, see Foglia & Wilson, 2013). In line with this view, there is increasing consensus that language generation and language understanding rely on sensorimotor and emotional processing. Language trivially depends on perception and action. We can talk about our experience of the world and some words, like color terms, are directly tied to these experiences. But it seems that also abstract concepts are constructed from cognitive primitives that are sensorimotor in nature.

Two of the most influential theories on how (abstract) concepts are grounded in experience are *Conceptual Metaphor Theory* and *Perceptual Symbol Systems* (Shapiro & Spaulding, 2021). Conceptual metaphor theory proposes that abstract concepts are grounded through metaphors (Gibbs et al., 1997; Lakoff & Johnson, 1980, 1999). These metaphors map from a source domain of embodied experience to an abstract target domain. For example, we can reason and talk about *anger* in terms of *heat* or *redness* ("Don't get hot under the collar", "She was scarlet with rage", "Keep cool"), a mapping that reflects changes in our physiology (increase in body temperature, red face) when we are angry (Lakoff, 1987). Accordingly, metaphors are not only a stylistic device but fundamental to our thought processes. The theory of perceptual symbol systems proposes that abstract concepts are grounded in simulations of introspective experience (Barsalou, 1999; Barsalou et al., 2003). While perceiving examples of a category, bottom-up activation patterns in sensory and motor areas are stored in long-term memory. Conceptualization of a category in absence of a stimulus reactivates these patterns. While perceptual symbol systems were more directly inspired by empirical findings, they are in principle compatible with conceptual metaphor theory.

Empirical support for the grounding of word meaning in the perception and action systems of the human brain comes from behavioral, neuroimaging, and neuropsychological studies (for reviews, see Barsalou, 2010; Glenberg et al., 2013; Kiefer & Pulvermüller, 2012; Pulvermüller, 2018). Neuroimaging studies show that word and sentence processing activates the corresponding sensory (visual, auditory, olfactory, and gustatory) and motor areas in the brain and that these activations are category-specific (Barsalou, 2008; Binder & Desai, 2011; Kemmerer, 2015; Kiefer & Pulvermüller, 2012; Pulvermüller, 2013). Among others, a functional magnetic resonance imaging (fMRI) study showed that passive reading of action words referring to face, arm, or leg actions (e.g. "lick", "pick", "kick") activates similar areas as actual movement of the face, arm, or legs (Hauk et al., 2004). Interactions between sensorimotor processing and language processing have also been observed in a series of transcranial magnetic stimulation (TMS) studies (for an overview, see Barsalou, 2010; Pulvermüller, 2018). In one of these studies, participants had to perform a lexical decision task involving words describing manual actions (normally performed with the dominant hand) and non-manual actions (Willems et al., 2011). Participants' responses were faster for manual (but not non-manual) words after stimulating the dominant hand area in the premotor cortex compared to stimulating the non-dominant hand area. Because TMS actively manipulates neural activation, these findings imply that sensorimotor activations not only correlate with linguistic processing but also have a causal impact.

The coupling between sensorimotor processing and linguistic processing also establishes a relationship between the embodiment of emotion and the understanding of emotion-related language. Different behavioral studies applied manipulations to enforce or inhibit facial expressions that are associated with certain emotions and showed a causal effect on the processing of words and sentences about emotional information (Niedenthal, 2007). For example, in one of these studies, participants received a botox injection that temporarily paralyzed the facial muscle used for frowning, which is active in the expression of negative emotions (Havas et al., 2010). The botox injection increased participants' reading times for

sentences describing angry and sad situations but not for sentences describing happy situations. These results suggest that experiencing emotions and processing emotion words rely on the same cognitive processes.

Language in turn also influences perception. Language has been shown to influence higher-level processes of perception such as recognition but also lower-level processes such as discrimination and detection (Lupyan, 2012b; Lupyan et al., 2020). A study related to visual object recognition showed that the recognition accuracy of ambiguous images[1] increases dramatically when participants are provided with labels, in the form of a forced-choice selection task or in the form of a superordinate label (e.g. "musical instrument" if the image displays a trumpet) (Samaha et al., 2018). In a study by Dils and Boroditsky (2010), participants saw real visual motion, read a story about physical motion, or read a story about abstract motion. Afterward, they had to interpret an ambiguous and unrelated image of a bird, which could be seen as facing upwards or downwards, by drawing a worm into the bird's beak. Seeing upward or downward motion but also reading about physical (but not abstract) upward or downward motion increased the probability of perceiving the bird as moving in the same direction, suggesting that language understanding can influence our interpretation of visual scenes *even when visual and linguistic content are unrelated*.

A large body of evidence for the influence of language on perceptual discrimination comes from the domain of color vision. Among others, it has been shown that differences in color terms between languages lead to differences in the ability to discriminate color stimuli (e.g., Roberson et al., 2008; Winawer et al., 2007). Differences in perception that arise due to different categorical repertoires are also known as *categorical perception* effects (Goldstone & Hendrickson, 2010). For example, Korean speakers distinguish between yellow-green ("yeondu") and green ("chorok"). Unlike English speakers, whose basic color terms do not make this distinction, Korean speakers are faster at discriminating green colors if they fall into different categories in Korean (Roberson et al., 2008). Moreover, such categorical perception effects can be eliminated when linguistic processing is suppressed by a verbal interference task (e.g., Winawer et al., 2007). But long-term experience with a certain language is not required to induce categorical perception effects. Color discrimination can also be altered short-term by teaching participants new color categories (Grandison et al., 2016; Ozgen & Davies, 2002; Zhou et al., 2010).

The effects of language on perception go beyond the perception of objects. Emotion words can play an important role in constructing perceptions and experiences of emotion (for a review, see Lindquist & Gendron, 2013). In particular, impairing participants' access to emotion words reduces their ability to recognize emotions on faces (Gendron et al., 2012; Lindquist & Gendron, 2013; Lindquist et al., 2014). This effect has, for example, been observed in emotion response studies, where participants typically have to match pictures of posed facial muscle movements (e.g. scowls, wide eyes, or smiles) with emotion words (e.g. "angry", "afraid", "happy"). Recognition accuracy drops drastically when participants perform a free labeling task

[1] The stimuli were generated by superimposing images of objects onto a patterned background, blurring the resulting images, and converting them to a black-and-white image based on a threshold.

instead of a forced-choice task (Lindquist & Gendron, 2013). Language has further been shown to influence action perception. For instance, Zarr et al. (2013) demonstrated that processing sentences that describe directional action negatively affects subsequent perception of actions in the same direction. In their experiment, participants read blocks of sentences that described either transfer away or toward the reader, using either the leg or the hand (e.g. "Ethan bicycled the mail to you"). Afterward, participants had to judge the endpoint of videotaped actions. The error increased if the action direction agreed between the video and the sentences but only if the same body part was important.[2] Taken together, these findings demonstrate a bidirectional influence between language and embodied representations (perception, emotion, action).

In general, it has been argued that conceptual representations are dynamic and depend on prior knowledge and the current task (e.g. Cibelli et al., 2016; Lupyan, 2012b; Lupyan et al., 2020; Regier & Xu, 2017). They seem to integrate top-down linguistic processes with bottom-up sensorimotor processes. Language plays the role of abstracting knowledge from concrete experiences, and influences (often augments) cognition by contributing this abstract, categorical information (Lupyan & Lewis, 2019). For example, it is easier to learn labeled than unlabeled categories even when these labels are redundant (Lupyan et al., 2007). Similarly, categories are easier to learn if the features of the categories are easier to name (Zettersten & Lupyan, 2020). In line with that, within-category recognition memory is worse when participants label objects than when they do not because the labeling process distorts the object representations based on top-down category information. It has been suggested that the integration process of linguistic and sensorimotor information relies on standard principles of inference under uncertainty (Cibelli et al., 2016; Regier & Xu, 2017) (see Box 1). This interpretation could help reduce controversy about the relationship between language and cognition by providing an explanation of different degrees of language interference based on the degree of cognitive uncertainty.

Box 1. Influence of language on perception as inference under uncertainty

Here, we look at the hypothesis that language influences perception following the principles of inference under uncertainty. Under this framework, humans take into account categorical information as well as perceptual information when inferring a stimulus property (e.g. color). The perception of the stimulus property acts as one cue, c_1, and the information about the category (label) that is activated by that perception act as another cue, c_2. Bayes theorem states that the probability of an event A given event B can be determined by

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} . \tag{1.1}$$

Given the visual cue c_1 and the categorical cue c_2, stimulus property S can be inferred with Bayes rule as $P(S \mid c_1, c_2) \propto P(c_1, c_2 \mid S)$.

[2] The authors present different potential explanations as to why the interaction is negative and not positive, assuming that action planning and sentence processing rely on the same system. For example, one explanation could be that the action control system is fatigued. Another explanation could be that motion perception becomes relatively automatic over time, which leads to a down-regulation of activity in the action control system.

If c_1 and c_2 are normally distributed with $c_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $c_2 \sim \mathcal{N}(\mu_2, \sigma_2)$, it follows that

$$P(c_1, c_2 \mid S) \sim \mathcal{N}(\frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\mu_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}\mu_2, \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2})$$

(Regier & Xu, 2017). In other words, the mean of the distribution is a weighted average of the means of the cues, and the weight for each cue increases with the uncertainty (i.e. variance) about the other cue. As a result, activated categories will have a stronger influence on the inference process if the stimulus perception is not very reliable.

Figure 1.1 illustrates the principle of inference under uncertainty for color labels. For a yellow-green stimulus, a native speaker of Korean will activate the category "yeondu" ("yellow-green"), while a native speaker of English will activate the category "green". As a result, the reconstructed stimulus will be shifted more toward the center of the green colors for English speakers than Korean speakers.

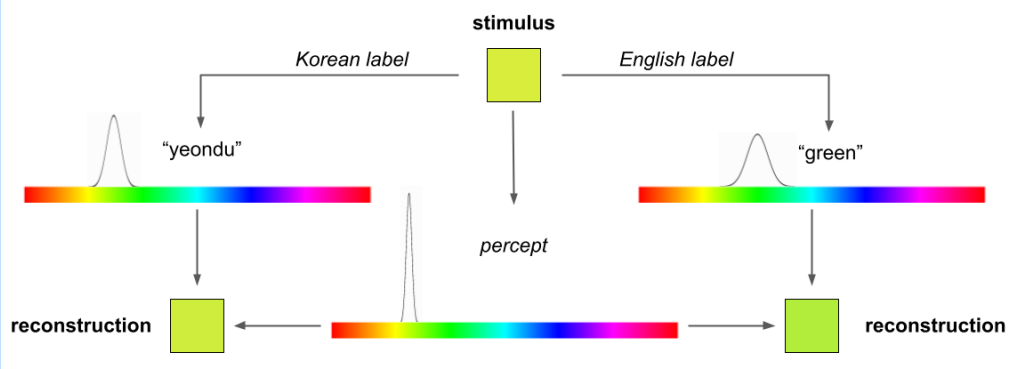


Figure 1.1: Qualitative example of how language may influence perception in a process of inference under uncertainty.

1.2.2 Functional language

We saw above that traditional cognitive science approached language as symbol manipulation that occurs in isolation from other cognitive processes. Similarly, some traditions in linguistics approached language as an abstract system that can be analyzed in isolation from its (contextual) *use* (e.g., Chomsky, 1965). However, language can also be viewed as a tool to achieve goals, whereby to use language means to perform an action (Austin, 1962; Wittgenstein, 1953). Importantly, this action involves other people and is inherently social. According to Clark (1996, p. 3), language use is "the joint action that emerges when speakers and listeners—or writers and readers—perform their individual actions in coordination, as ensembles".

The discipline in linguistics studying the flexible use of language in context is called *pragmatics*. Pragmatics emerged between 1950 and 1960 under the strong influence of John Austin and Paul Grice. Logical positivism had primarily analyzed language as a means to make factual assertions that can be evaluated in terms of their truth conditions (Bechtel, 1988). The *Speech Act Theory*

founded by Austin (1962) and later elaborated by his student Searle (1969, 1979), in contrast, moved the focus from declarative usage of language to the active results of utterances.[3]

Roughly speaking, a speech act is an act that a speaker performs by making an utterance, such as promising, warning, requesting, apologizing, and the like. Austin (1962) distinguished between different acts that are performed simultaneously when producing an utterance: the *locutionary act* (generation of a meaningful utterance), the *illocutionary* act (the act performed in saying something), and the *perlocutionary* act (the actual effect of the locutionary and illocutionary acts, whether intended or not). For example, in a specific situation, the utterance "Can you open the door" could be divided into the locutionary act of uttering a meaningful sentence, the illocutionary act of making a request, and the perlocutionary act of the listener opening the door. The example shows that speech acts can be indirect: "Can you open the door?" is not an inquiry about the listener's ability to open the door but a request.

Grice's (1975) *Theory of Conversational Implicature* made important contributions to understanding how non-literal meaning can be communicated. The theory proposes that interlocutors abide by the cooperative principle "Make your contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged" (p. 45). To abide by that principle, an utterance should be true (Maxim of Quality), as informative as required (Maxim of Quantity), relevant for the exchange (Maxim of Relation), and clear (Maxim of Manner). A listener will assume that a speaker follows these principles and will draw conversational implicatures to interpret their utterances accordingly. Accordingly, "Can you open the door?" will be interpreted as a request in order to preserve the Maxim of Relation. Thus, Grice's theory provides an explanation of how speakers convey and listeners recognize intended meaning—or in other words illocutionary force. Together, these works have laid the foundation for an action-oriented view of language and the analysis of utterances based on speaker intention and context (Holtgraves, 2008).

In communication, the speaker tries to convey an intended meaning and the listener tries to recognize that intention (Clark, 1996). The same meaning can be expressed by different utterances and the same utterance can have different meanings. Importantly, the speaker's intention is generally under-determined by the linguistic form and both agents draw on various additional sources of information to make communication succeed. These sources of information can include knowledge about the interlocutor and one's relation to them, conversational history, general circumstances, and so on. Pragmatic inference against the background of this contextual information is essentially an application of social cognition, requiring speaker and listener to coordinate their relative perspectives (Clark, 2009; Goodman & Stuhlmüller, 2013; Holtgraves, 2008; Shafto et al., 2012; Sperber & Wilson, 1995). Reasoning about the other's mental state, the speaker can choose an utterance that will best convey his or her intention to the listener, and the listener can infer that intention.

[3] Generally, utterances in pragmatics can be defined as "the intentional acts of speakers at times and places, typically involving language" (Korta & Perry, 2020).

The *Rational Speech Act* (RSA) framework provides a formal theory of language production and interpretation in context (Frank & Goodman, 2012; Goodman & Frank, 2016; Goodman & Stuhlmüller, 2013). The framework integrates the ideas of Gricean pragmatics into Bayesian models of social reasoning (e.g., Baker et al., 2009). Communication is modeled as recursive reasoning between a speaker and a listener. The listener—assuming that the speaker is rational and cooperative—applies Bayesian inference (see Equation 1.1) to arrive at the speaker's intended meaning given their utterance. So, following Bayes rule, they can calculate $P(s|u) = {}^{P(u|s)\cdot P(s)}/_{P(u)} \propto P(u|s) \cdot P(s)$, where s is the meaning, and u the utterance. The speaker, in turn, reasons about the listener to choose the utterance that will maximize the probability of being understood (while keeping utterance cost low). The recursive reasoning process between speaker and listener is grounded in the literal meaning of the utterance. Figure 1.2 illustrates the inference process of a pragmatic listener for a toy example and Box 2 gives a more formal introduction to the RSA model. The RSA framework has been successfully used to model human language use in a variety of situations and can explain complex linguistic phenomena such as vagueness, metaphor, or hyperbole (for an overview, see Goodman & Frank, 2016; Scontras et al., 2018).

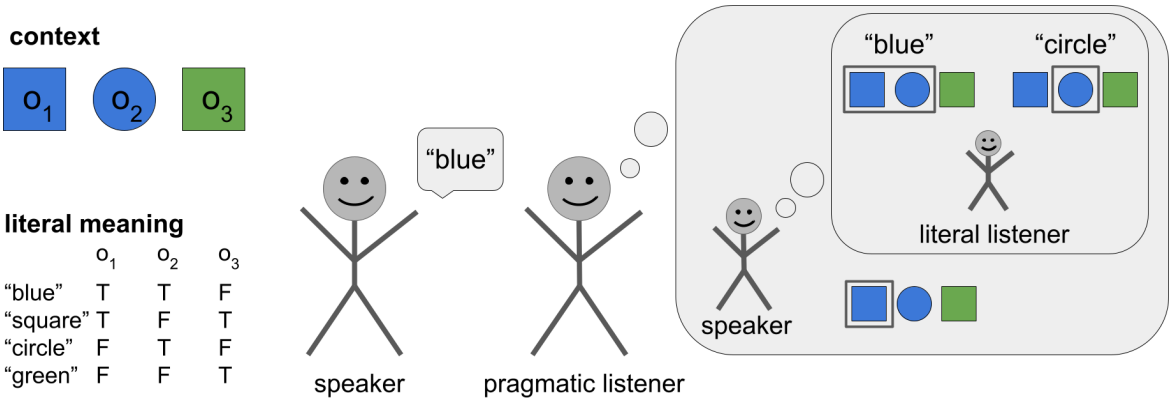


Figure 1.2: Example of a pragmatic listener's inference process according to the RSA model. There are three objects and four possible utterances. The pragmatic listener tries to identify the referent for the utterance "blue" by reasoning about a speaker who reasons about a literal listener. "blue" is true of two objects, the blue square and the blue circle. If the speaker had wanted to refer to the circle, however, they could have said "circle" to be unambiguous. Hence, the target object is probably the blue square.

Box 2. Vanilla Rational Speech Act model

The following description is adapted from Ohmer et al. (2022). In the vanilla RSA model, conditional probabilities describe how a speaker maps a state, s, onto an utterance, u, and how a listener maps an utterance onto a state, while they take into account each other's perspective.

$$P_{LL}(s \mid u) \propto [\![u]\!](s) \cdot P(s)\,, \tag{1.2}$$

$$P_{PS}(u \mid s) \propto \exp\left(\alpha \cdot [\log P_{LL}(s \mid u) - C(u)]\right), \tag{1.3}$$

$$P_{PL}(s \mid u) \propto P_{PS}(u \mid s) \cdot P(s)\,. \tag{1.4}$$

At the basis of the recursive reasoning process is a literal listener (LL: 1.2) who maps an utterance onto any state for which it is true, at the same time considering the prior probability of that state. In (1.2), $[\![u]\!](s)$ is the denotation function returning the truth value of utterance u for state s. A pragmatic speaker (PS: 1.3) chooses their utterance such that the probability of being correctly understood by a literal listener is maximized while production cost, $C(u)$, stays low. The parameter $\alpha \in \mathbb{R}^+$ regulates the speaker's optimality. For $\alpha = 0$, the speaker's choices are random, and for $\alpha \to \infty$, they will always select the utterance that yields the maximal probability of being correctly understood by the literal listener. The pragmatic listener (PL: 1.4), in turn, interprets an utterance as if coming from a pragmatic speaker, also considering the prior probability of states.

1.2.3 Communication games—simulating language in the world

Communication games provide a framework for simulating grounded, interactive language use by reducing the complexity of real-world linguistic interaction. The idea of conceptualizing language as a game can be traced back at least to Wittgenstein (1953), who developed the notion of a *language game* to illustrate the context-dependent and purposive nature of language. Let us look at an example of such a game:

> "The language is meant to serve for communication between a builder A and an assistant B. A is building with building-stones: there are blocks, pillars, slabs and beams. B has to pass the stones, in the order in which A needs them. For this purpose they use a language consisting of the words "block", "pillar", "slab", "beam". A calls them out; — B brings the stone which he has learnt to bring at such-and-such a call. Conceive this as a complete primitive language." (Wittgenstein, 1953, p. 3, Paragraph 2)

In this game, the word "block" is an order, in another game, it might have an entirely different function, such as answering a question. In that sense, language games are conventions that are part of social interaction and the meaning of a word depends on the rules of the game.

Lewis (1969) introduced the influential framework of *signaling games*, which provides a game-theoretic formalization of linguistic interaction. Inspired by the work of Schelling (1960) on coordination games, he tried to explain how conventions of meaning could emerge in a society without presupposing explicit agreement. Although there is no direct connection to Wittgenstein, it has been argued that this formalization is in line with his thoughts on language games (Correia, 2019). In a *(Lewis) signaling game* (see Figure 1.3) there are two agents, the sender and the receiver. The agents have perfect common interest. One of N world states is selected at random and observed by the sender but not the receiver. Conditioned on the state, the sender selects one of $M \geq N$ signals to send to the receiver. Conditioned on the signal, the receiver selects one of N actions. For each state, there is exactly one correct action and both agents receive a positive payoff if the correct action is selected and no payoff otherwise. *Signaling systems* are communication strategies where the sender chooses a distinct signal for each state and the receiver performs the

correct action for each signal. Signaling systems form a *Nash equilibrium*, i.e. neither sender nor receiver can increase their payoff by unilaterally changing their strategy.[4] According to Lewis, different signals can be said to *mean* different states in a given signaling system. As a result, conventions of meaning are grounded in the repeated play of signaling games, where sender and receiver coordinate their strategies.

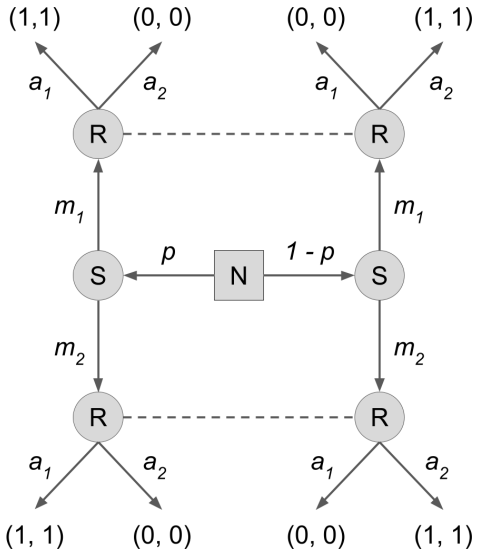


Figure 1.3: Extensive form representation of a Lewis signaling game with two messages and two states. A state is randomly selected from $N = 2$ states with probabilities $p/1 - p$ and passed to the sender (S). The sender generates one of two messages ($\{m_1, m_2\}$), which is passed to the receiver (R). The receiver does not know which state led the sender to produce the given message, indicated by the dashed lines. The receiver chooses one of two actions ($\{a_1, a_2\}$). When selecting the correct action both agents receive a payoff of 1, otherwise a payoff of 0.

The signaling game framework has been criticized and extended (for an overview, see Rescorla, 2019). Knowing that signaling conventions exist does not explain how the agents arrive at these conventions and why they stick to them. According to Lewis, agents stick to a signaling convention because they are rational and know that the other agent is rational (and will stick to the convention). In his own words, the agents must be capable of "mutual ascription of some common inductive standards and background information, rationality, mutual ascription of rationality and so on" (Lewis, 1969, pp. 56–57). To explain how the agents arrive at a certain convention Lewis uses the notion of *salience*: Agents will choose an equilibrium that stands out among other equilibria in some psychologically relevant respect. The notion of salience has been criticized for being obscure and lacking formalization, as well as presupposing that agents know that other agents will abide by the salient convention (Rescorla, 2019). Taken together, Lewis' demands on the agents' cognitive abilities seem excessive considering the evolution of language from a historical perspective or communication systems among animals.

Skyrms (1996) provided a compelling solution to both these problems by viewing signaling games in the context of evolution. When studying signaling games in evolutionary terms (e.g., Skyrms, 1996, 2010), signaling systems are not the result of a conscious decision process among

[4] Note that there are Nash equilibria that do not constitute signaling systems. E.g., such a Nash equilibrium occurs if the sender always chooses the same signal, regardless of the state, and the receiver always chooses the same action, regardless of the signal.

rational agents but simply arise from natural selection. Agents follow specific hard-wired signaling strategies, mutations can arise, and in sum, the population evolves toward strategies that increase relative fitness (i.e. payoff). The Lewis signaling game, together with its extension to evolutionary processes (Lewis-Skyrms signaling game), still serves as a foundation for a lot of research in evolutionary linguistics.

Many later approaches follow Lewis' idea to model the evolution of meaning through (variations of) signaling games (e.g., Kirby & Hurford, 2002; Nowak & Krakauer, 1999; Steels, 1997, 2001). For example, Steels and colleagues study the evolution of meaning in signaling games through self-organization (for reviews, see Steels, 1997, 2001, 2003). Aside from computer simulations, they also use robotic experiments, showing how a set of categories that are grounded in sensorimotor interaction with the environment can emerge and how this set of categories can become sufficiently shared to allow a group of agents to communicate about their environment. Prominent games in this line of work are the *naming game* and the *guessing game*, which are played between two randomly selected agents of a population (Steels, 2012). In both games, the sender has to refer to a specific object in a given context and the receiver has to recognize that object. In the naming game, there is no ambiguity about how objects differ from one another. It can, for example, also be played with symbolic object encodings lacking internal structure (e.g., Steels, 1995). Each object belongs to a distinctive category and the agents have to develop names for these categories. In the guessing game, in contrast, objects can be discriminated along different dimensions. The objects are typically represented by feature vectors and the sender chooses discriminative features for signaling (e.g., Steels & Kaplan, 1999) (see Figure 1.4). The agents follow predefined deterministic or stochastic rules of when to add or delete word-meaning associations from their lexica (Baronchelli et al., 2008). The use of feature-based representations and robotic agents makes the process of language grounding explicit in terms of sensorimotor interaction, which does not play a role in the theoretical models discussed above.

Contextualized signaling games, such as the naming game and the guessing game, are in general known as *reference games* or *referential games*. A reference game is played between a sender and a receiver in a specific context (a subset of the full space of meanings). A *target* object is selected and the sender but not the receiver is aware of the target. The other objects in the context are called *distractors*. Based on a message from the sender, the receiver tries to identify the target. Reference games are also employed outside the context of meaning evolution, in particular to study pragmatic reasoning (e.g., Frank, 2016; Frank & Goodman, 2012; Qing & Franke, 2015) (see Figure 1.2). For example, experimental studies use reference games to make quantitative measurements of pragmatic inference which can then be used to inform design choices and parameters in models of pragmatic reasoning (e.g., Frank, 2016; Frank & Goodman, 2012, 2014).

This wide array of applications demonstrates that communication games, in particular formalized through game theory, provide an extremely useful framework for studying interactive language use in empirical and computational experiments.

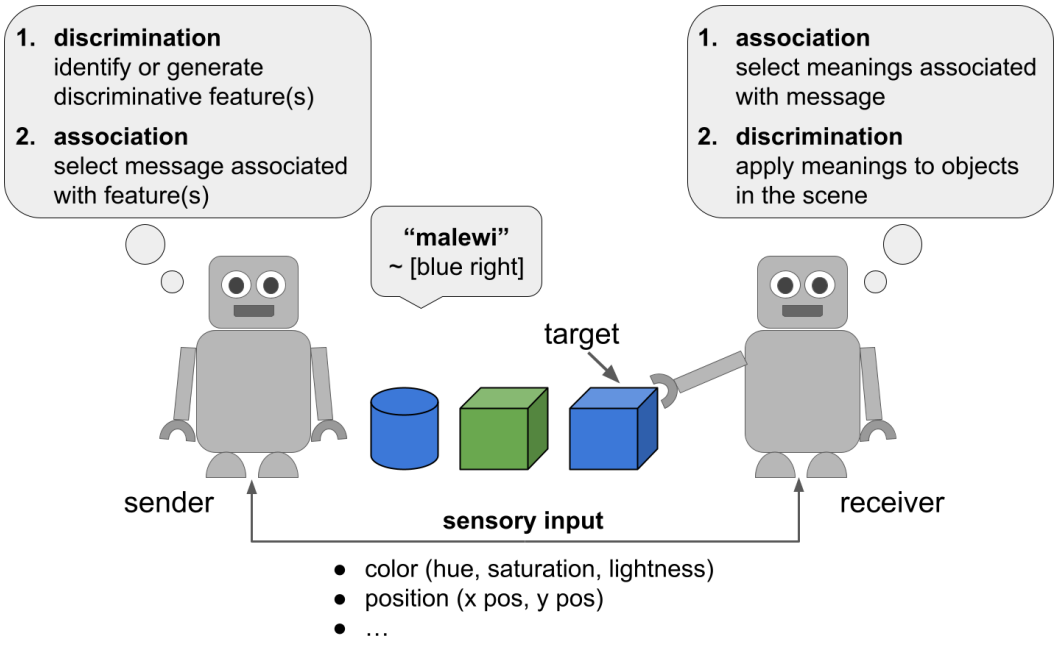


Figure 1.4: Example of a guessing game. The sender intends to communicate the target object. It can first narrow down the context by pointing (not shown here). Based on the sensory input, the sender tries to identify features that discriminate the target object from the other objects. Consulting its current lexicon, it selects the message that is most strongly associated with the discriminative feature(s). The receiver interprets the message and tests whether its meanings identify an object in the context. The receiver provides feedback by selecting an object.

1.3 Artificial neural network models of language acquisition and language emergence

This section starts by motivating the use of computational models, in particular ANN-based models, in language learning and language emergence research (for a critical discussion of ANNs as models of cognition, see Chapter 5). It then provides a brief introduction to ANNs before reviewing some of the literature on ANN models of language learning and language emergence.

1.3.1 Motivation

Why computational modeling?

From a scientific perspective, computational models serve as a method to develop and test specific theories. In particular, they require these theories to be made precise, such that hypotheses about causes and processes underlying the simulated phenomenon can be expressed by mechanistic formulations. Sometimes, empirical evidence can be used to inform the model's design or to evaluate the implementation of a certain hypothesis. But computational models also allow us to study phenomena that are difficult to access empirically. In the latter case, the models are used to find out whether a certain mechanism could in principle offer an explanation.

In the case of language acquisition, a large body of empirical evidence on word learning phenomena and changes in word learning throughout development is available. Many different theories have been proposed concerning the cognitive requirements and mechanisms underlying word learning. Computational models can be used to implement specific theories and generate (quantitative) predictions that can be compared to empirical results. Thereby, they lead to a better understanding of the potential and limits of a particular theory.

In the case of language emergence, there are no direct archaeological records about the origins and evolution of language (Hauser et al., 2014). At the same time, experiments in the lab are influenced by the participants' linguistic experience (although they may still be informative). In this context, computational models are commonly used to ask "what if" questions. They can implement agents in different environments with different cognitive capacities and study the conditions under which communication systems emerge, as well as the conditions that give rise to specific linguistic properties. Besides, they can simulate the evolution of language across multiple generations of agents, involving time scales that are difficult to reconstruct with human participants.

Humans possess the most advanced communication system we know of. Thus, from an applied perspective, mimicking characteristics of human language use may prove beneficial for artificial language systems. Computational models of human language learning or language emergence can provide implementations of such characteristics. For instance, biases that help children to learn words in ambiguous situations could also support learning in machines (e.g., Gandhi & Lake, 2020; Lake et al., 2017), and simulations that give rise to compositional language could make machines better at generalizing to novel communication scenarios (e.g., Lake et al., 2017; Lazaridou & Baroni, 2020). Looking at specific models below, we will find that the line between understanding cognition and building better machines is often blurry.

Why artificial neural network models?

ANNs, especially DNNs, are currently dominating machine learning. They have led to unprecedented improvements in various domains, including computer vision (Voulodimos et al., 2018), natural language processing (Otter et al., 2021), and control (Li, 2017). However, ANNs are not only interesting from an engineering perspective but also have a rich tradition as computational models of cognition (Rumelhart, McClelland, et al., 1986). (Readers who are not familiar with the basics of ANNs are referred to Section 1.3.2 to appreciate the following paragraphs.)

Artificial neurons are loosely inspired by biological neurons: they receive activation through input connections, apply a non-linear activation function to these inputs, and transmit the resulting signals to other neurons. Although ANNs abstract away from many features of the brain (see Section 5.2), they still capture some important mechanistic constraints on cognition (McClelland et al., 2010). In particular, both artificial and biological neural networks rely on distributed information processing. ANNs can therefore help to explain how complex cognitive

functions and behaviors can emerge from interactions among many relatively simple units (McClelland, 2010).

One key advantage of ANNs is that they can learn directly from their experience with the environment. Compared to traditional machine learning techniques (e.g. support vector machines), they require very little external guidance. Task-relevant features are extracted automatically in the learning process, rather than being determined by an expert. Relying on handcrafted rules or features is infeasible for increasingly complex tasks (Richards et al., 2019). Automated feature extraction in deep architectures, in contrast, allows ANNs to learn the substantial domain knowledge required for intelligent behavior (Kietzmann, McClure, et al., 2019).

In addition, several properties inherent to distributed processing are also central properties of cognition, such as robustness to damage and noise as well as the ability to generalize from previous experiences (McClelland et al., 2010; Robins, 1993). Unlike symbolic representations, distributed representations are structured, in that the internal properties of the representations carry information on what they are about (Clark, 1993, p. 19). As a result, distributed representations can enable generalization to novel combinations of feature values, if the network has encountered these feature values in the training data (LeCun et al., 2015). In addition, distributed representations are claimed to be relatively robust to noise in the input and to damage to the network (e.g., Buckner & Garson, 2019; Robins, 1993) as relevant information may still be recovered from the (remaining) activations. Note, however, that ANNs are less robust to noise than humans (e.g., Hendrycks & Dietterich, 2019), and may rely on different features than humans when generalizing (e.g., Geirhos et al., 2019).

In the case of language acquisition, ANN models are used to capture incremental word learning processes based on input stimuli from the environment such as co-occurrences of words and objects. Hence, they are especially suited to account for changes in cognitive abilities throughout development (Westermann & Twomey, 2017). In the case of language emergence, ANN models are used to simulate the emergence of grounded communication from interactions with other agents and the environment (Lazaridou et al., 2017). As the emergence of communication is related to the formation of concepts, and the discrimination of objects along relevant dimensions (features), ANNs are natural model candidates.

1.3.2 Brief introduction to artificial neural networks

This subsection provides a short introduction to ANNs (for more details on ANNs and DNNs and recent transformer architectures, see Bishop, 2006; Goodfellow et al., 2016; Kamath et al., 2022, respectively). Its main purpose is to familiarize the reader with the basics of ANNs in machine learning and to offer a condensed overview of the architectures and training methods relevant to this thesis (some additional concepts will be mentioned for the sake of completeness).

General overview

ANNs are computational models consisting of a collection of many simple processing units (artificial neurons), each producing a real-valued activation (Schmidhuber, 2015). These units are interconnected and typically multiple units are arranged into layers. Neural networks have at least two layers, the input and the output layer, but can have additional, so-called hidden layers in between. If the network has multiple hidden layers, it classifies as a *deep* neural network (LeCun et al., 2015; Schmidhuber, 2015). ANNs are used to approximate a function f^*. For example, a classification network approximates the function $f^*(\mathbf{x}) = y$, which maps the input $\mathbf{x}$ onto a category y. The input layer represents some external input (e.g. an image), and neurons in other layers get activated through weighted connections from previously active neurons. Most neurons perform a non-linear transformation of their input. Through the composition of many non-linear computations very complex functions can be approximated. ANNs have been around for a long time but with improving computer infrastructure (both hardware and software) it has become possible to train very large DNNs. In this "deep learning revolution" (see e.g., Hinton & LeCun, 2019), DNNs have become the most dominant machine learning method.

Most other machine learning methods are limited in their ability to process raw input data. They typically require the design of task-specific feature extractors to transform the input data into a representation on which the learning system can operate (LeCun et al., 2015). ANNs are special because they can learn from raw data without additional task-specific knowledge. In the training process, the ANN weights are adapted through an optimization procedure to minimize or maximize a certain objective function (Goodfellow et al., 2016). Returning to the example of a classifier, the data set could consist of image-category pairs, and the network could be trained to minimize the error between the true and the predicted category. During training, the network learns to extract features relevant to the task, from low-level features in early layers to higher-level features in late layers. For example, in the case of image processing networks, early layers typically function as edge detectors while later layers represent more complex shapes (LeCun et al., 2015). An important challenge is to find network weights that allow the ANN to generalize to novel inputs after training.

Architectures

Artificial neuron. The basic processing units of ANNs are artificial neurons. Artificial neurons are loosely inspired by biological neurons. They receive an input signal (postsynaptic potential at the dendrites), calculate a weighted average of these signals (different synaptic strengths), and generate an output (signal exciting via the axon) by computing a non-linear activation function of the weighted inputs and an additional bias value (the threshold value that needs to be exceeded for the neuron to fire) (Weijters & Hoppenbrouwers, 1995). Mathematically, an artificial neuron with input vector $\mathbf{x} \in \mathbb{R}^D$ is a function $f : \mathbb{R}^D \to \mathbb{R}$ that is parameterized by the weights $\mathbf{w} \in \mathbb{R}^D$, the bias $b \in \mathbb{R}$, and the non-linear activation function σ: $f(\mathbf{x}) = \sigma(\sum_{k=1}^{D} w_k x_k + b)$ (see

Figure 1.5.A). The activation function is determined by the modeler and the weights and biases are free parameters that are learned by the model. The bias parameter is usually absorbed into the weight parameter by defining an additional input variable $x_0 = 1$ and an additional weight $w_0 = b$.

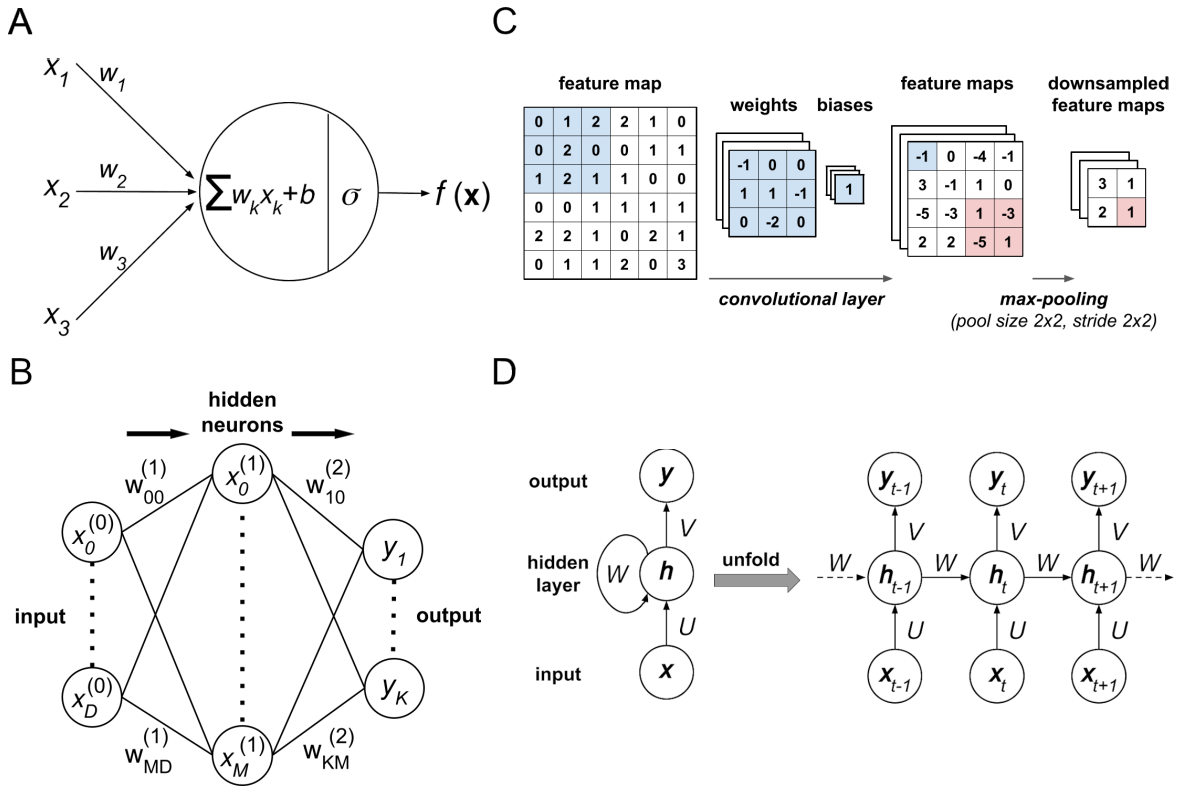


Figure 1.5: Schematic overview of different ANN architectures. **A)** Artificial neuron with input vector $\mathbf{x} \in \mathbb{R}^3$ and activation function σ. The neuron computes the function $f(\mathbf{x}) = \sigma(\sum_{k=1}^{3} w_k x_k + b)$. **B)** MLP with one hidden layer (figure inspired by Bishop, 2006, p. 228). Input dimension is D, hidden dimension M, and output dimension K. The biases are absorbed into the weight parameters by adding an input/hidden unit x_0 and a weight $w_{i0} = b_i$. **C)** Convolutional layer followed by a max-pooling layer. The input feature map is a matrix and for each of the three filters, an output feature map is generated. Convolving the blue part of the input feature map with the top kernel, and adding the bias results in the value -1. Note, if the input was a tensor, each filter would be a tensor of the same depth. In the max-pooling example, the maxima are extracted from a $2x2$ pooling window on the feature map, and the window is moved across the feature map at a stride of 2 in x- and y-direction. Calculating the maximum of the red pooling window results in the value 1. **D)** RNN in folded (left) and unfolded (right) form. The input $\mathbf{x}$ is processed sequentially. At each time step t, the hidden state $\mathbf{h}_t$ is updated based on the current input $\mathbf{x}_t$ and the previous hidden state $\mathbf{h}_{t-1}$. Biases are not included to avoid clutter.

Multi-layer perceptron. Multiple artificial neurons can be combined to form a neural network. A simple and popular type of ANN is the multilayer perceptron (MLP). Neurons are arranged into layers and there is a directed weighted connection from each neuron in layer L to each neuron in layer $L+1$. Hence, MLPs are fully-connected feed-forward networks. An MLP always has an input layer and a final output layer. There can be multiple hidden layers in between. Figure 1.5.B illustrates an MLP with one hidden layer. In the general case, each layer L calculates the function:

$$f^{(L)}(\mathbf{x}^{(L-1)}) = \sigma^{(L)}(\mathbf{W}^{(L)}\mathbf{x}^{(L-1)}) \,,$$

where the weight matrix $\mathbf{W}^{(L)} \in \mathbb{R}^{D^{(L)} \times D^{(L-1)}}$ summarizes all weights (including biases) between layer $L-1$ and layer L, and layer dimensionality is given by $D^{(L)}$. The activation function is applied to each vector element. Taken together, a network with N layers calculates

$$f(\mathbf{x}^{(0)}, \boldsymbol{\theta}) = \left(f^{(N)} \circ f^{(N-1)} \circ \cdots \circ f^{(1)}\right)(\mathbf{x}^{(0)}) \,,$$

where all weights and biases are summarized in $\boldsymbol{\theta}$ (Goodfellow et al., 2016, pp. 168–169).

Comment on activation functions. To train a neural network with the standard algorithms, all network operations must be differentiable (an explanation will follow later in this section). Typical hidden neuron activation functions include the *sigmoid*, *hyperbolic tangent* (tanh), and the *rectified linear function* ($ReL(x) = max(0,\, x)$). The activation function of the output layer depends on the task of the network. For example, the softmax activation function is used to generate a probability distribution across categories.

Convolutional neural network. Convolutional neural networks (CNNs) are a type of feed-forward ANN designed to process data with grid-like topology, such as images. They were developed and later improved by LeCun et al. (1989, 1998). CNNs are organized in layers, but unlike MLPs, at least one of their layers is *convolutional layer* and not fully-connected (see 1.5.C). Convolutional layers perform a mathematical operation called *convolution*. In mathematics, a one-dimensional convolution between two discrete functions, f and g, is defined as

$$f * g\,[i] = \sum_{m=-\infty}^{\infty} f[i-m]\, g[m] \,,$$

and analogously in two dimensions as

$$f * g\,[i,j] = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} f[i-m, j-n]\, g[m,n] \,.$$

Most machine learning libraries use a cross-correlation function because it is easier to implement but still call it convolution (Goodfellow et al., 2016, p. 329). The cross-correlation is very similar to the convolution, only the minus signs are flipped (e.g. in 1D: $f * g\,[i] = \sum_{m=-\infty}^{\infty} f[i+m] g[m]$). In a convolutional layer, this cross-correlation function is calculated between the incoming activations and so-called filters (or kernels) (see Figure 1.5.C for a 2D example). Convolution with filters is a standard operation in computer vision, and different filters have been designed to detect different features (Shapiro & Stockman, 2002). Figure 1.6 illustrates how filters can be used as edge detectors. In CNNs, the filter values are free parameters. In learning these values, the network learns useful feature extractors for the task.

Convolutional layers have several advantages over fully-connected layers (Goodfellow et al., 2016, pp. 329–335). Because the filters are usually much smaller than the input feature maps the connectivity in a convolutional layer is sparse. Weights are fewer and they are shared between

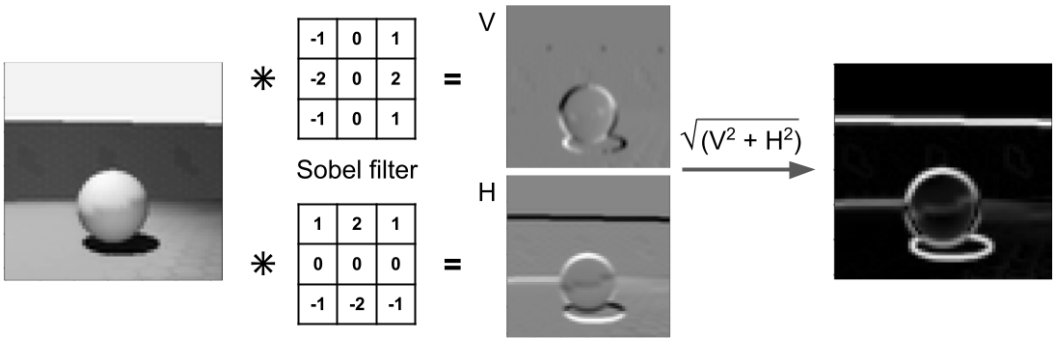


Figure 1.6: Example of edge detection with a Sobel filter (Shapiro & Stockman, 2002, p. 164). The Sobel filter approximates the image derivate in x-direction (top filter) and y-direction (bottom filter), to extract vertical (V) and horizontal (H) edges. The combined gradient magnitude is shown in the right image. The original image was taken from the *3dshapes* data set (Burgess & Kim, 2018) and transformed into a gray-scale image.

different input and output neurons because the filters are reapplied to all input locations as part of the convolution. By applying the filters to different locations they can extract the same features anywhere in the input feature map, irrespective of translations. Sparse connectivity and weight sharing significantly reduce the memory requirements and the number of operations compared to a fully-connected layer. Further, CNNs often use downsampling layers to reduce the size of the output feature maps by calculating a summary statistic. For example in *max-pooling*, collections of values in the feature map are replaced by their maximum value (see Figure 1.5.C, right side). CNN architectures vary but they usually comprise multiple convolutional layers interleaved with pooling layers, followed by multiple fully-connected layers (e.g., Krizhevsky et al., 2012). Convolutional layers and downsampling, in particular in combination with big data and highly-optimized GPU implementations of convolution (and all other neural network operations), allow for the construction of very deep networks (Gu et al., 2018).

Recurrent neural network and derivatives. Recurrent neural networks (RNNs) (based on Rumelhart, Hinton, et al., 1986) are similar to MLPs but have recurrent (cyclic) connections in their hidden layer(s). They are a family of networks developed for processing sequential data, such as language. Figure 1.5.D visualizes an RNN. The left part of the figure shows all model components. The input $\mathbf{x}$ is a sequence of D-dimensional input vectors, $\mathbf{x}_i \in \mathbb{R}^D$, $\mathbf{h}$ represents a layer of M hidden neurons with tanh activation function, and $\mathbf{y}$ is a sequence of K-dimensional output neurons, $\mathbf{y}_i \in \mathbb{R}^K$. The weight matrices represent the weights of fully-connected layers. $\mathbf{U} \in \mathbb{R}^{M \times D}$ are the weights from input to hidden layer, $\mathbf{V} \in \mathbb{R}^{K \times M}$ the weights from hidden to output layer, and $\mathbf{W} \in \mathbb{R}^{M \times M}$ are the weights of the recurrent connections. The right part of the figure visualizes how information is processed over time. This form of representation is also known as the *unfolded* or *unrolled* network. At each time step t, part $\mathbf{x}_t$ of the input sequence is processed. The hidden units $\mathbf{h}_t$, also called hidden state, are updated based on their previous activation and the current input. At each time step, an output $\mathbf{y}_t$ is generated from the current

hidden state. Taken together, the network computes the hidden state and the output as:

$$\mathbf{h}_t = \tanh\left(\mathbf{b}^{(1)} + \mathbf{W}\mathbf{h}_{t-1} + \mathbf{U}\mathbf{x}_t\right),$$
$$\mathbf{y}_t = \sigma\left(\mathbf{b}^{(2)} + \mathbf{V}\mathbf{h}_t\right),$$

where $\mathbf{b}^{(i)}$ are the bias vectors, and σ is the activation function of the output neurons (Goodfellow et al., 2016, p. 374). The hidden state functions as a (leaky) "memory" that keeps track of past states and inputs, emphasizing task-relevant information and forgetting task-irrelevant information. The Figure 1.5.D shows an RNN with a many-to-many mapping from inputs to outputs (e.g. for machine translation) but other configurations such as many-to-one (e.g. for sentiment classification) or one-to-many (e.g. for image captioning) are possible (Kaur & Mohta, 2019).

The main advantage of RNNs is that they can operate on all time steps and on all sequence lengths. Even with only one hidden layer they can be considered *deep* as the recurrent connections allow them, in principle, to create and process memories of arbitrary sequences of inputs (Schmidhuber, 2015). The weight matrices of the model are shared across time steps, which is a desirable feature for many types of sequential data. Think of the sentences "It was raining on Sunday" and "On Sunday it was raining". An MLP needs to learn to process each word at each time step, i.e. it needs to learn what "Sunday" at input position 5 means as well as what "Sunday" at input position 2 means. The RNN, in contrast, uses the same weights to process each word in different positions and can modulate the meaning based on earlier words. However, vanilla RNNs have difficulties processing long-range dependencies, roughly speaking because they assign more importance to recent information (Bengio et al., 1994; Hochreiter, 1998, 2022). To address this issue *gated* RNNs, such as the *long short-term memory* (LSTM) (Hochreiter & Schmidhuber, 1997) and the *gated recurrent unit* (GRU) (Cho et al., 2014) have been developed, which are now widely used. Gated RNNs update the hidden state more selectively by using additional trainable layers, known as gates. The gates regulate at each time step which information is written to, read from, or deleted from the hidden state.[5]

Fundamentals of neural network training

The neural network parameters (weights and biases) are learned using a method called *stochastic gradient descent* (SGD), or alternatively using *stochastic gradient ascent* (SGA). The goal is to find parameters that minimize or maximize the objective function (e.g. minimize an error function or maximize a likelihood function). As SGD and SGA follow the same principle, only SGD will be explained here (for an extensive overview of SGD methods, see Ruder, 2016). Let J be the error function that should be minimized. SGD adapts the network parameters $\boldsymbol{\theta}$ as

$$\boldsymbol{\theta} = \boldsymbol{\theta} - \epsilon \cdot \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}),$$

[5] The LSTM has a hidden state and a cell state within the hidden layer. They store different kinds of gated information.

where ϵ is a *learning rate* chosen by the modeler. SGD performs one update at a time for each data point. It calculates the gradient of the error function for that data point with respect to the network parameters, and adapts the parameters in the negative direction of the gradient. As a result, the parameters gradually move to a configuration where the error function has a local minimum. Calculating the gradients with respect to all network weights is achieved with the *back-propagation algorithm* (Rumelhart, McClelland, et al., 1986). The data set can be processed multiple times and each iteration defines a new *training epoch*. Most implementations use mini-batch SGD (but call it SGD), which means that they aggregate the error for a batch of data points (e.g. by averaging) and update the weights based on the gradients of this aggregated error. Batching has the advantage of reducing the update variance and utilizing efficient matrix calculations on GPU (Ruder, 2016).

It is important to find parameters that generalize to novel data. To test whether the network generalizes, the available data is usually divided into non-overlapping training and test sets. The training set is used to learn the model parameters. The test set is used to evaluate the trained model on novel data. Sometimes an additional validation set is split off from the data. The validation set can be used to select good hyperparameters (number of layers, numbers of neurons, activation function, learning rate, ...): Different models are trained, and the error on the validation set is used to identify the best one. It can also be used to monitor the training status. Training can be stopped just before the ANN becomes better at modeling the training data than the validation data, in other words before the network starts to *overfit* the training data at a loss of generalization. Next to this method of *early stopping*, there are various regularization techniques to avoid overfitting (see e.g., Moradi et al., 2020).

Training regimes

The main training regimes in machine learning are *supervised learning*, *unsupervised learning*, and *reinforcement learning*. Here, each regime will be explained with respect to ANN training.

Supervised learning. In supervised learning, the available data set contains ground-truth examples of input-output pairs, $(\mathbf{x}^{(i)}, \mathbf{y}^{*(i)})$, and the network is trained to produce the correct output for a given input based on these examples. The ground truth output values are called *labels*. The objective function is therefore defined as some error function between the true label $\mathbf{y}^*$ and the generated label $\mathbf{y} = f(\mathbf{x}; \boldsymbol{\theta})$. Let us again use classification as an example. To predict a distribution over categories, the output neurons have a softmax activation function: $y_i = {\exp(z_i)}/{\sum_{j=1}^{K} \exp(z_j)}$, where K are the number of categories and $\mathbf{z}$ are the values before the activation function is applied (i.e. weighted average of inputs plus bias). The network is trained to minimize the error between the true category and the predicted category. Often, it is trained to minimize the *cross-entropy loss* between the true categorical distribution, given by a one-hot vector $\mathbf{y}^*$ (a vector with a one at the true category index and zeros everywhere else), and the

predicted distribution, $\mathbf{y}$:

$$J(\mathbf{y}^*, \mathbf{y}) = -\sum_{i=1}^{K} y_i^* \cdot \ln(y_i)\,.$$

Supervised learning is the most common form of deep learning (LeCun et al., 2015) and by now there is a huge variety of large labeled data sets for learning different tasks.

Unsupervised learning. In unsupervised learning, there are no labels and the model learns some form of representation based on the input data alone. One important application of unsupervised learning is dimension reduction, where the goal is to learn a lower-dimensional representation of the input data without losing (too much) information. To solve this problem with ANNs, they are typically arranged as autoencoders, consisting of an encoder and a decoder network. The encoder takes the input $\mathbf{x}$ and maps it onto a lower-dimensional latent (hidden) representation $\mathbf{z}$. The decoder mirrors the architecture of the encoder and maps the latent representation $\mathbf{z}$ onto an output representation $\mathbf{y}$. The objective function is then defined as the reconstruction error between $\mathbf{x}$ and $\mathbf{y}$. The type of ANN for the encoder and decoder is based on the input data. For instance, K-dimensional vectors could be encoded and decoded with MLPs, and the the objective function could be the mean squared error (MSE):

$$J(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{K} (x_k - y_k)^2.$$

Representations that are learned in an unsupervised manner can also be used in downstream supervised learning or reinforcement learning tasks.

Reinforcement learning. In reinforcement learning (RL), agents learn what actions they need to take to achieve certain goals from trial-and-error interactions with an environment. RL is a major source of human knowledge (Sutton & Barto, 2018, p. 1) and therefore seems especially promising for building AI. Here, we focus on agents that are implemented as ANNs. Labeled data is not required for RL but the agent receives (potentially) sparse feedback from the environment in the form of positive or negative reward. Formally, RL can be described as a Markov decision process (for details, see Sutton & Barto, 2018, Chapter 3) but here a less formal introduction will be given. Figure 1.7 shows the main elements of RL. At each time step t, the environment generates a state s_t, and a reward r_t. Based on the current state, the agent generates an action a_t, and the cycle repeats. The agent's objective is to learn a policy that maximizes the expected *return*, R, which is defined as the discounted cumulative reward, $R = \sum_{t=1}^{T} \gamma^{t-1} \cdot r_t$, where $\gamma \in [0, 1]$ is the discount factor. So, the agent tries to maximize the cumulative reward but is more focused on rewards in the near future for smaller values of γ. T is the number of time steps in one episode. After one episode, the environment resets (think of it as starting another round of a game).

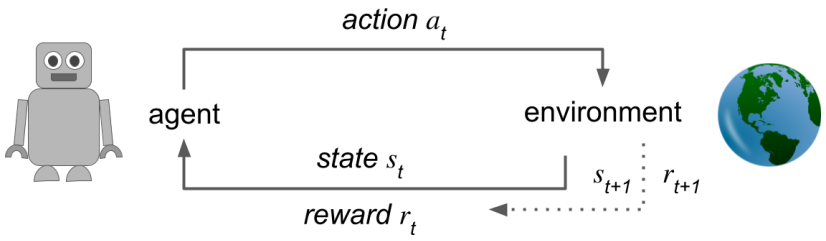


Figure 1.7: Reinforcement learning schema. At each time step t, the environment generates a state s_t and a reward r_t. Based on the current state, the agent performs an action a_t, and the cycle repeats. (Figure inspired by Sutton and Barto (2018, p. 54))

There are mainly two kinds of training algorithms for RL: value-based methods and policy gradient methods. Roughly speaking, in value-based methods, the agent tries to learn which states (or state-action pairs) maximize the return and chooses the best actions based on these estimates. The ANN maps states (or state-action pairs) to their estimated return and is trained on minimizing the error between estimated and achieved return. In policy gradient methods the agent tries to learn directly which actions in a given state maximize the return. The ANN is used to map states to actions and is trained on maximizing the return. To visualize the difference between the two methods, think of an agent on top of a hill trying to find food that is located at the bottom of the hill. In value-based methods, the agent will learn that *being at the bottom of the hill* yields a high return and will therefore move downhill. In policy gradient methods, the agent will learn that *moving downhill* yields a high return. In this thesis, the REINFORCE algorithm (Williams, 1992) (see Box 3) will be used, which is a specific policy gradient algorithm.

Box 3. The REINFORCE algorithm

The objective function is the expected return, $J(\boldsymbol{\theta}) = \mathbb{E}[R]$. The goal of the agent is to maximize this function, hence SGA is performed. According to the policy gradient theorem (Sutton & Barto, 2018, p. 334-335), the following equation holds:

$$\nabla_\theta \mathbb{E}[R] = \mathbb{E}[R \cdot \sum_{t=1}^{T} \nabla_\theta \ln \pi_\theta(a_t \mid s_t)] \,.$$

The REINFORCE algorithm operationalizes this relationship to train the agent.

The algorithm:

Randomly initialize $\boldsymbol{\theta}$.
For *each episode* $\{s_1, a_1, r_2, \ldots, s_{T-1}, a_{T-1}, r_T\} \sim \pi_\theta$
For $t = 1$ *to* $t = T - 1$
Calculate the return $R_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} r_k$.
$\theta \leftarrow \theta + \epsilon \nabla_\theta \ln \pi_\theta(s_t, a_t) R_t$

1.3.3 Artificial neural network models of language acquisition

ANNs have been used to study virtually all aspects relevant to word learning (for an overview, see Westermann & Twomey, 2017). Among others, they have been used to model the learning of speech sounds (e.g., Warlaumont et al., 2013; Westermann & Miranda, 2004; Yoshikawa et al., 2003), the segmentation of words from a continuous stream (e.g., Aslin et al., 1996; Christiansen et al., 1998; Elman, 1990), and learning word-object mappings (e.g., Li et al., 2004; Plunkett et al., 1992; Regier, 2005). The focus of this thesis lies on the formation of word-object mappings in referential contexts.

Connectionist approaches to word learning can roughly be divided into self-organizing models, which will not be discussed here (but see Appendix D for a brief overview), and multi-layer neural networks (Mayor & Plunkett, 2010), the latter of which include recent deep learning approaches. Before discussing these models, it is important to understand the phenomena they try to capture. Therefore, this section continues with a brief interlude about important word learning phenomena before providing an overview of the modeling landscape by presenting a few examples, divided into classical (shallow) ANNs and DNNs (for extensive overviews, see Frank et al., 2009; Mayor & Plunkett, 2010; Regier, 2005; Westermann & Twomey, 2017).

Interlude: Some important word learning phenomena

The formation of word-object mappings is challenging. A child will typically face a scene with many potential referents when hearing a word (Quine, 1960). In addition, it is unclear whether the word refers to a whole object, or only to some object part or feature. To complicate things further, word meanings might (currently or generally) not be observable in the scene at all. Several mechanisms seem to help children solve this disambiguation challenge. For example, children display a *mutual exclusivity (ME) bias* in mapping novel labels to novel objects, rather than familiar objects for which they already know a label (Markman & Wachtel, 1988). In addition, they generalize novel labels to other referents based on shape or functionality, referred to as a *shape bias* or *taxonomic bias* (Landau et al., 1988; Markman & Makin, 1998). Finally, they display a *whole-object bias* by behaving as if labels refer to whole objects rather than parts or features (Markman, 1990). Although a large body of empirical work supports the existence of "biased" responses in word learning, the cognitive mechanisms behind these responses are highly debated (e.g., Diesendruck & Bloom, 2003; Lewis et al., 2020; Markman, 1992; Smith & Samuelson, 2006).

While word learning is slow and errorful at the beginning several improvements can be observed around the second year of life (Regier, 2005; Westermann & Twomey, 2017). The rate at which children acquire new words accelerates, also known as *vocabulary spurt* (Behrend et al., 2001; Carey, 1978). Relatedly, children become better at *fast mapping*, i.e. the required number of word-object co-occurrences to form a link decreases. They also learn to overcome the ME response and learn second labels (Mervis et al., 1994). During that time, children are prone to

overextension and underextension errors, where they apply words more broadly (e.g. saying "dog" to various animals) or more narrowly (e.g. saying "dog" only to the family's dog) than would be allowed by their meanings (Kay & Anglin, 1982). Providing mechanistic explanations of observed phenomena, such as the word learning biases described above, is an important part of modeling word learning. Knowledge about developmental change generates additional theoretical constraints that can be used to inform and evaluate model assumptions (Lewis et al., 2020).

Classical neural networks

Many ANN models simplify the process of word learning in the following way: there is a set of objects and a set of words, and the model needs to learn the right mappings between the two. To model comprehension, the model maps words onto objects, and to model production, the model maps objects onto words.

One of the earliest ANN models of word learning used an autoencoder (Plunkett et al., 1992) to implement this idea (see Figure 1.8.A). The model received images (random dot stimuli) or labels (one-hot vectors) as input. These inputs were processed by different layers but mapped onto the same latent representation layer. From that layer, visual or linguistic output streams forked, so the model could be used to simulate comprehension (label to image) and production (image to label). The network was trained in a three-phase cycle. First, the network was trained on reconstructing an image and only the weights on the "visual side" (left side in Figure 1.8.A) were updated. Second, the network was trained on reconstructing the corresponding label and only the weights on the "linguistic side" (right side in Figure 1.8.A) were updated. Finally, the network was trained on reconstructing image and label simultaneously and all weights were updated. That way, the model learned to associate labels with images. During testing, the network then mapped images to labels and labels to images. The model implemented several aspects of early semantic development: a vocabulary spurt, over- and underextension errors, a comprehension/production asymmetry (more words are understood than used), and a prototype effect (Posner & Keele, 1968, 1970) (the model generalized to prototypical inputs although it was only trained on distorted versions). This early work showed that many important aspects of word learning can be captured by simple associative learning mechanisms.

But word learning goes beyond association. Among others, children rely on social, linguistic, and attentional information to disambiguate the meanings of words (e.g., Bloom, 2000; Hollich et al., 2000). Besides, it has been suggested that the rapid advancements in word learning during the second year of life could be explained through a transition from early associative learning to referential learning, where children are aware of the referential nature of words (Lock, 1980; McShane, 1979). This raises the question of how much of word learning can be explained by associative learning alone. The model by Plunkett et al. (1992) left several aspects unstudied, such as word learning biases, fast mapping, or the role of word structure (e.g., phonological similarities). More generally, standard ANNs—including the model by Plunkett

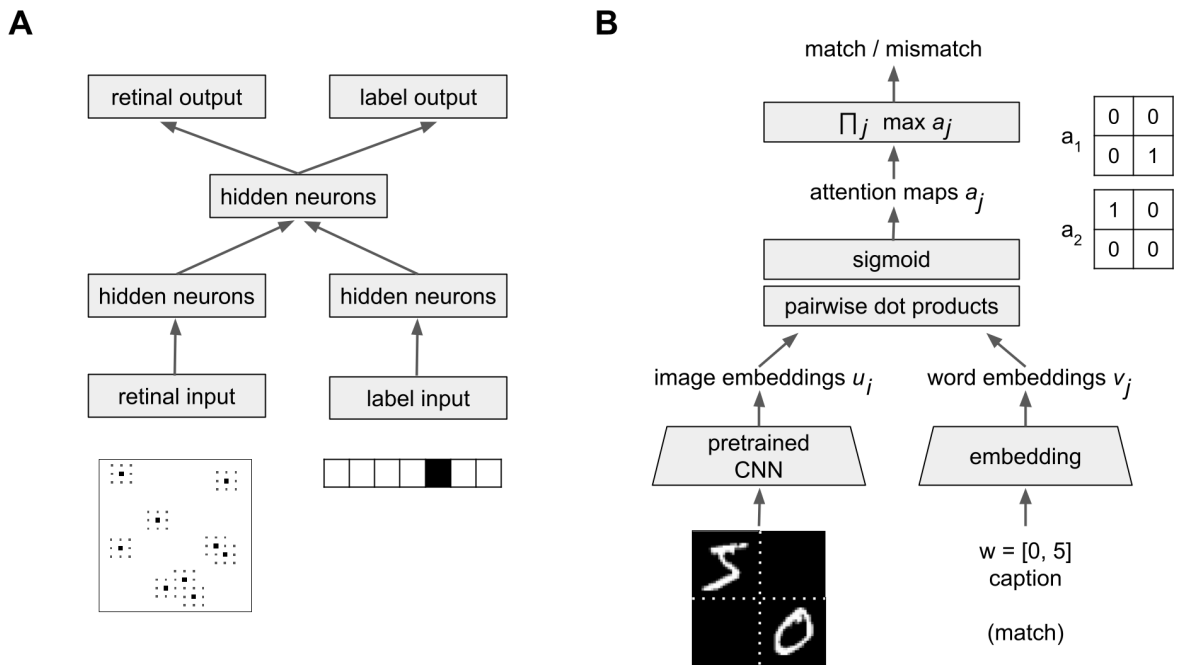


Figure 1.8: Illustration of different word learning models: A) MLP model by Plunkett et al. (1992), and B) DNN model by Vong and Lake (2020). **A)** The visual input consists of random dot images, which are additionally processed by input receptors with Gaussian receptive fields. The linguistic input consists of one-hot labels. The model is an MLP autoencoder with two pathways. Each training trial proceeds in three steps. First, the visual pathway is trained on reconstructing the image, then the linguistic pathway is trained on reconstructing the label, and finally, the entire network is trained on reconstructing both inputs. **B)** Visual inputs are sets of handwritten digits and linguistic inputs are sets of labels. The task of the model is to decide whether the labels and the images match or mismatch. Words and images are embedded and all pairwise dot products between image embeddings and word embeddings are calculated and passed through a sigmoid activation. The maximal value for each word is collected. These values are multiplied, indicating a match if the product equals 1 and a mismatch if the product equals 0.

et al. (1992)—suffer from *catastrophic forgetting*, in which learning of the first pattern is eliminated by subsequent learning of other patterns (McCloskey & Cohen, 1989). Hence, they fail to explain how children can retain the meanings of words without being continuously exposed to them.

Regier (2005) developed a new model of associative learning to overcome the problem of catastrophic forgetting and further investigate whether assuming a change in learning mechanism is necessary. The model was called *LEX* and had the same macrostructure as the model by Plunkett et al. (1992) in that either linguistic or visual information (both in the form of artificial feature vectors) could be provided as input or retrieved as output. However, the concrete implementation was an adaptation of *ALCOVE*, an ANN model of category learning relying on exemplar-based representations (Kruschke, 1991). ALCOVE is a feed-forward network consisting of an input layer, a hidden layer, and an output layer generating a category. The hidden layer does not consist of neurons but stores previously encountered input examples. During training, the weights from input to hidden layer learn to extract category-relevant features and activate the corresponding exemplars, and the weights from hidden to output layer learn to map exemplars to their category. Regier (2005) extended this model such that it processed visual and linguistic inputs at the same time. LEX stored exemplars of forms and meanings and learned attention weights for mapping new inputs to these exemplars, as well as associative weights between the two types of exemplars. The model could account for several changes and improvements in

word learning that are observed in the second year of life, such as fast mapping, the acquisition of phonologically similar words, the shape bias, as well as the learning of second labels despite an ME bias. It thus expanded the spectrum of word learning phenomena that can be explained with associative learning (in combination with attention mechanisms), making a change in the learning mechanism less likely.

More progress in resolving some of the issues with ANNs as models of word learning was made by distinguishing between processes of *learning* and *inference*. Many ANN models cannot account for phenomena such as fast mapping because gradient-based learning is incremental and typically requires many examples and iterations. Often, for example in *LEX*[6], competition mechanisms are included to allow for fast mapping despite slow learning. Against by then common modeling practices but in line with empirical findings (Horst & Samuelson, 2008), McMurray et al. (2012) argued that determining a referent for a novel word in context (online inference) and establishing word-meaning associations in the lexicon (long-term learning) are two distinct processes. They illustrated this approach with a model that relied on associative learning but performed referent selection using real-time competition dynamics. The model was able to simulate various word learning phenomena as well as interactions between situation- and developmental-time processes. The distinction between learning and inference has been picked up in later models, including ours (e.g., Gulordava et al., 2020; Ohmer et al., 2022).

In sum, associative learning can explain many word learning phenomena but not all of them. ANNs are well-suited to implement associative learning. Yet, they also display behaviors that are different from human learners (e.g., catastrophic forgetting). The models by Regier (2005) and McMurray et al. (2012) highlight how ANNs can be complemented with attention, competition, and online inference to become better models of word learning. Building models that integrate an ANN-based "associative core" with additional learning and inference mechanisms seems to be a promising approach.

Deep neural networks

Here, some deep learning models that investigate empirical word learning phenomena will be presented. Unlike the models discussed above, many of them were not developed as models of word learning per se. Some try to implement useful learning biases to build better algorithms and others simply point out interesting parallels between human and machine learning. While language models technically also learn word-meaning mappings, they are generally not concerned with human word learning phenomena and will not be discussed.

DNNs are good at processing naturalistic inputs. Vong and Lake (2020) took a first step toward building a model that learns word-object associations from raw inputs (see Figure 1.8.B). They used a cross-situational learning paradigm where multiple (artificial) words were presented

[6] In LEX, competition takes place between exemplars. The model learns to allocate attention to relevant features. Attention is used to weigh features when calculating the similarity between the input and the stored exemplars. The exemplars are then activated proportionally to the resulting similarity values.

together with an image of multiple handwritten digits. The image was processed by a CNN and the words by an embedding layer[7]. The similarity of the representations was used to evaluate whether linguistic and visual inputs matched or mismatched, and the network weights were updated with supervised learning. The model successfully learned word-referent mappings from these ambiguous inputs. The authors also quantified whether the model displayed a preference for ME. They presented the model with a novel word and an image of a novel and a familiar digit and measured which digit the model attended to. The model did not display an ME bias unless the novel word had been presented in mismatch conditions during training.

In general, the ME bias forms a challenge for ANNs. Vanilla architectures display an anti-ME bias, which means that they map novel inputs onto familiar outputs (Gandhi & Lake, 2020). People are interested in building networks with an ME bias as this would improve learning, in particular lifelong (or continuous) learning where new targets can appear throughout the training process (Gandhi & Lake, 2020).

Gulordava et al. (2020) introduced a deep learning model that was more specifically designed to tackle the ME bias challenge. Using a similar setup as Vong and Lake (2020), the model mapped visual and linguistic inputs into a joint embedding space, and association strength corresponded to the similarity of these embeddings. The authors experimented with a) symbolic encodings of co-occurring words and objects extracted from a data set of child-directed speech (called CHILDES, MacWhinney, 2000), and b) natural images together with referring expressions extracted from their captions. All possible pairs of words and objects in a scene were aligned to create a training data set of word-object pairs. The authors used three different max-margin losses, introducing competition among referents (anti-polysemy), competition among words (anti-synonymy), or competition among both. For example, the anti-synonymy loss was given by $\mathcal{L} = \sum_i \max(0, 1 - \cos(\mathbf{w}, \mathbf{o}) + \cos(\mathbf{w}_i, \mathbf{o}))$, where $\mathbf{o}$ and $\mathbf{w}$ are the embeddings of a word-object pair, $\mathbf{w}_i$ is a random negative example, and cos is the cosine similarity. In addition, learning and referent selection were separated, and referent selection could be similarity-based only or rely on Bayesian pragmatic inference. In line with observations from shallow ANNs (Yurovsky et al., 2013), competition mechanisms turned out to be crucial for generating an ME bias in both experiments. In particular competition over words seems necessary, implemented either through the loss function or through pragmatic inference.

Both the models by Vong and Lake (2020) and Gulordava et al. (2020) rely on negative examples to bring about an ME response. Negative examples make the network learn that the novel word is not associated with any of the familiar objects, leading to an ME bias when the network chooses between a novel and a familiar object at test time. It seems unlikely that children rely on the same mechanism as they also display an ME bias for entirely novel, nonsensical words (e.g., Markman & Wachtel, 1988) and negative examples may not always be available.

Another line of research studies language learning in *situated* DNN agents (Hermann et al., 2017; Hill et al., 2017; Hill, Clark, et al., 2020; Hill, Lampinen, et al., 2020). The agent typically

[7] An embedding layer turns indices (e.g. integers, one-hot vectors) into continuous vectors.

comprises a CNN for processing visual input, an RNN for processing linguistic input, an MLP for combining these inputs, and an RNN for memory and generating actions. The agent is trained on an instruction-following task with natural language text input. It navigates in a 3D simulated environment and tries to identify a referent (e.g. "red object next to the green object") or to perform a desired action ("lift a pencil"), and is trained with reinforcement learning. In an early referent identification setup, the agent showed impressive generalization abilities by generalizing to novel word combinations, applying known relations and modifiers to unfamiliar objects, and re-using knowledge about familiar concepts when learning new concepts (Hermann et al., 2017). Although these simulations are driven by a motivation to design artificial agents, several word learning phenomena have been observed across studies, such as the vocabulary spurt (Hermann et al., 2017) and—using a more complex architecture—fast mapping (Hill, Clark, et al., 2020). Another study showed that the agent develops a color bias instead of a shape bias, probably due to an inherent CNN bias toward color information (Hill, Clark, et al., 2020). More generally, it was found that the agent's ability to generalize improves with the number of word/object experiences, the diversity of the visual input, and the visual invariances afforded by the agent's perspective (Hill, Lampinen, et al., 2020), demonstrating the benefits of more "human-like" learning scenarios.

Taken together, these papers paint a mixed picture. Some properties of deep learning setups, such as complex and realistic environments may support generalization, but other properties like the inherent color bias, the inherent anti-ME bias and general data-hungriness highlight the differences from human word learners. Still, deep learning models seem best suited to study some questions, for example how word-meaning mappings can be formed based on cross-situational learning from raw visual and linguistic inputs. At the same time, human word learning abilities and mechanisms serve as useful inspirations for AI and ML.

1.3.4 Artificial neural network models of language emergence

In *language emergence* experiments, symbols have no ex-ante meaning. Rather, meaning evolves in a process of interaction with other agents and the environment. Computational models of language emergence cover different aspects of language, such as phonetics and phonology, lexicon formation, and syntax (Cangelosi, 2005; Steels, 1997). Moreover, different studies target different time scales of language evolution, from emergence within one generation, over cultural transmission, to genetic evolution (for a survey, see Wagner et al., 2003). Different time scales involve different algorithms to adapt the agents' (communicative) behavior. Usually, supervised learning or reinforcement learning are used to capture learning within one generation or cultural transmission, genetic or evolutionary algorithms are used to capture the effects of natural selection, and hybrid frameworks are used to combine learning and evolution (Wagner et al., 2003).

It seems that early ANN models of language emergence were mostly motivated by an interest in the origins and evolution of animal or human communication systems and the relationship

between the evolution of language and other cognitive capacities, while more recent DNN models are more often motivated by practical concerns. Early simulations often used a large number of simple agents and involved population and/or evolutionary dynamics. DNN simulations, in contrast, typically study language emergence between two agents of the same generation—although there is a recent trend toward populations (e.g., Chaabouni et al., 2020a; Chaabouni et al., 2022; Harding Graesser et al., 2019; Ren et al., 2020). They employ more realistic environments and often focus on how well the emergent communication system can generalize to novel situations (e.g. by virtue of compositional structure). In the following, I will provide a brief overview of ANN and recent DNN models.

Classical neural networks

I divide the literature in this field into 1) *emergence-only simulations* covering language emergence within a single generation, 2) *biological evolution simulations* covering language emergence plus evolution based on natural selection, and 3) *cultural transmission simulations* covering language emergence plus evolution through cultural transmission.

One of the earliest ANN simulations of language emergence within one generation, but also in general, was conducted by Hutchins and Hazlehurst (1995). In a proof-of-principle simulation, the authors demonstrated that a shared lexicon can emerge between agents implemented as simple autoencoders. The autoencoders had one hidden layer and the activations in that layer were interpreted as the "verbal representations" generated or received by the agents. The following training procedure was used: At each trial, a sender and a receiver agent were sampled randomly from the population, the sender generated a word and the receiver was trained with supervised learning to reconstruct the input as well as to produce the same verbal representation as the sender. One can argue that this setup does not involve proper communication as the receiver does not act on the message from the sender but rather learns to generate the same hidden representations in its reconstruction process.

Batali (1998) developed a more advanced emergence-only simulation, where the receiver really acts on the messages of the sender. He studied the emergence of structured communication with sequences of discrete symbols in a population of RNN agents. Agents used the same network to generate and process messages. During training, a sender and a receiver were selected at random.[8] At each trial, the sender produced messages for a set of meaning vectors of subject-predicate structure (e.g. a symbolic encoding of "you smile"). During production, the sender itself processed the symbols and selected those that would bring its own output closest to the original meaning vector. The receiver then processed these messages and was trained to minimize the difference between its output meaning vector and the original meaning vector. To some degree, the agents' messages mirrored the grammatical structure encoded in the

[8] To be precise, one receiver and multiple senders were selected at random. The receiver was trained together with each of these senders in random order. Then the procedure was repeated.

meaning vectors and the communication system generalized to new meaning vectors. This work established that shared and structured communication systems can emerge between ANNs.

In the context of emergence-only simulations, Luc Steels and colleagues have made important contributions by studying language emergence in groups of robotic agents playing reference games (for an overview, see Steels, 2005) (see Section 1.2.3). These models are not purely ANN-based but often involve ANN components. Working with robots, which are embodied agents, showed that artificial agents can self-organize symbolic systems that are grounded in sensorimotor interactions with the world and other agents (Bleys et al., 2009; Steels, 1998, 2001; Steels & Belpaeme, 2005). Embodiment, whether simulated or real, adds additional features that can influence emergent communication, such as different perceptions of the environment (Bleys et al., 2009) or the ability to gesture (Steels & Vogt, 1997).

In biological evolution simulations, the emergence and evolution of communication systems are studied across many generations of agents. Successful communication can constitute an evolutionary advantage and thereby increase the chances of reproduction. Studies in the field are often inspired by animal communication, focusing on the emergence of unstructured food and alarm calls. For example, Wagner (2000) used a spatial environment to study the effects of resource abundance and population density on cooperation, finding that signaling strategies evolve except when population density is high or resource abundance is low. Similarly, Reggia et al. (2001) showed that agents develop alarm calls if predators have a strong impact on survival, and food calls emerge if food is difficult to locate. Genetic algorithms are, however, arguably less suited to study how mappings between signals and meanings evolved in natural language, as these are not innate but learned (Kirby, 2002b).

To develop a model of human language evolution, Kirby and Hurford (2002) developed the *iterated learning* framework, which simulates the cultural transmission of language. In the iterated learning model, the first generation of agents (adults) produces signals for some randomly selected meanings. In a learning period, the next generation (children) is trained on these signal-meaning mappings, before becoming adults themselves. New children are introduced, and adults are removed, and the cycle is repeated until the system converges. Importantly, the language *emerges*, i.e. there is no communication system in place at the time of the first generation. Apart from vertical transmission from parents to children, the model can also include horizontal transmission in the form of within-generation communication. The iterated learning model has been particularly successful in explaining structural aspects of language such as word-order universals (Kirby, 1999), irregularity in highly frequent messages and regularity in less frequent messages (Kirby, 2001), recursive syntax (Kirby, 2002a), and compositionality (Smith et al., 2003) (for an overview, see Kirby et al., 2014).

In hybrid simulations, Cangelosi and colleagues used ANNs to simulate learning, and a genetic algorithm to simulate evolution, in a toy world of mushrooms and mushroom-foragers (see Figure 1.9.A). Among others, they studied why language production evolves even if only the receiver benefits, finding it to be a by-product of the evolving ability to discriminate different types of mushrooms (Cangelosi & Parisi, 1998) (for related work, see Mirolli and Parisi, 2005).

They also demonstrated that communication systems following a verb-object rule can emerge (Cangelosi, 1999). Moreover, they studied how the agents' internal representations are shaped by language, showing that learning through language induces additional structure which improves the agents' behavior beyond learning through sensorimotor interaction (Cangelosi, 2010; Cangelosi & Parisi, 2001). Like the work by Steels and colleagues, these simulations take important steps toward studying the emergence of language in the more general context of cognition by simulating sensorimotor interactions with an environment and considering the relationship between communication and concept formation.

Even if they are rather obvious, I would like to highlight some lessons from these studies. First, to simulate language emergence it is important that the agents not merely converge on the same signals but that they *react* to the signals of other agents. Second, to study language evolution in humans, one should focus on cultural rather than genetic evolution and the iterated learning framework provides a successful implementation. Third, taking into account perception and embodiment adds new variables that can help explain language-related cognitive phenomena, such as concept formation.

Deep neural networks

Sparked by the rapid advances in machine learning, there has been growing interest in studying language emergence in DNN agents that communicate to solve a common task (for a review, see Lazaridou & Baroni, 2020). This movement is mostly driven by AI researchers striving to build artificial agents that are capable of flexible and goal-directed language use. Deep learning allows for increasingly realistic settings. By now, agents have been trained on natural images (Dessi et al., 2021; Havrylov & Titov, 2017) and in simulated 2D and 3D environments (Das et al., 2019; Jaques et al., 2018); communication can take place in multi-turn interactions (Cao et al., 2018; Jaques et al., 2018), and agents can decide themselves when to communicate and who to communicate with (Das et al., 2019; Singh et al., 2019). However, deep learning simulations may also shed new light on the origins and evolution of human language by expanding the traditional modeling landscape.

Foerster et al. (2016) were among the first ones to demonstrate that DNNs can develop communication protocols in complex environments involving raw pixel inputs. They built two model versions, one with a discrete communication channel and one with a continuous communication channel. The choice of discrete versus continuous communication has conceptual and technical implications. In the continuous case, the learning signal can be back-propagated through the entire system, effectively collapsing the agents into a single network. In the discrete case, in contrast, back-propagation is not straightforward as the message-generating function is not differentiable. In the discrete implementation by Foerster et al. (2016), the agents treated other agents as part of the environment and were trained with reinforcement learning, leading

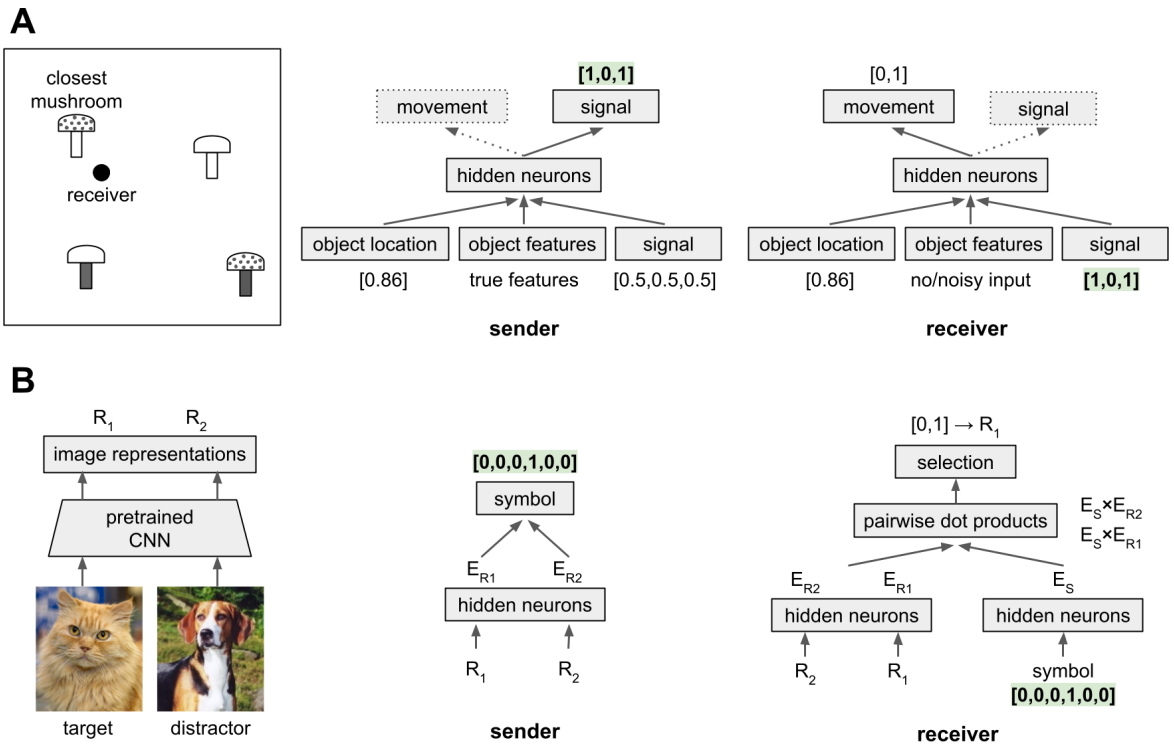


Figure 1.9: Illustration of different language emergence models: A) ANN model by Cangelosi and Parisi (1998), B) DNN model by Lazaridou et al. (2017). **A)** A sender and a receiver agent are sampled randomly from a population. The receiver navigates in a mushroom world, with mushrooms that are poisonous or edible, depending on their features. The receiver always receives information about the closest mushroom: location (normalized angles) and features (no features if too far away, noisy features if close enough). The sender has perfect vision regardless of the distance. Each agent is modeled by an ANN that receives information about the mushroom and a signal as input. Signals are binary vectors and to simulate no incoming signal for the sender the values are set to 0.5. The network outputs a horizontal and a vertical movement. If the receiver steps on a mushroom, it eats that mushroom and receives positive or negative payoffs. Only agents with relatively high cumulative payoffs generate offspring. **B)** Sender and receiver play a reference game with one distractor. Input data (including the two examples) are images from ImageNet (Deng et al., 2009), which are processed by a pretrained CNN. The agents receive these image representations, R_1 and R_2, as ordered (sender) or randomized (receiver) input. Depicted is the *agnostic sender* from the paper. It maps the representations onto internal embeddings, which it then maps onto a probability distribution over symbols, from which a symbol is sampled. The receiver generates embeddings of the image representations and the symbol. The selection probabilities are proportional to the dot products between the symbol and each image embedding.

to a true multi-agent setup more akin to human communication settings.[9]

Being able to train entire multi-agent systems with discrete communication channels marked important progress in the field. We saw examples of discrete communication in the ANN section above but only the receivers were trained in these cases (e.g., Batali, 1998; Kirby & Hurford, 2002). Reinforcement learning is one option to train systems with discrete communication. Alternatively, the *Gumbel-Softmax trick* (Jang et al., 2017; Maddison et al., 2017) can be used to generate a continuous (and therefore differentiable) approximation of the message sampling process. Discrete communication channels have the advantage that they can more easily be interfaced with natural language (Lazaridou & Baroni, 2020), for example, to analyze grammatical structure (van der Wal et al., 2020). Furthermore, discreteness introduces an additional bottleneck, which

[9] In principle, "true" multi-agent setups can also be implemented with continuous communication channels by applying the same training methods as in the discrete case instead of allowing free flow of the gradients.

might encourage the transmission of high-level, conceptual information. For these reasons, most language emergence simulations with DNN agents use discrete messages.

In a seminal paper, Lazaridou et al. (2017) established the use of reference games in deep multi-agent communication, and they are now widely used in the community (e.g., Choi et al., 2018; Dagan et al., 2021; Havrylov & Titov, 2017; Lazaridou et al., 2018; Rodríguez Luna et al., 2020). In the original setup (see Figure 1.9.B), the sender and the receiver both saw the same image of the target and the same image of the distractor. The messages consisted of single symbols. Apart from measuring performance, the authors evaluated how changes in the distractor distributions influence message content and presented a strategy for grounding the emergent messages in natural language. A follow-up investigation (Bouchacourt & Baroni, 2018) found that the agents could also successfully communicate about images of random noise, indicating that the sender—at least in part—relies on low-level information, such as (differences in) pixel values, rather than conceptual information, to point out the target. The following studies therefore came up with different countermeasures, such as showing only the target to the sender (e.g., Choi et al., 2018; Havrylov & Titov, 2017; Kottur et al., 2017; Lazaridou et al., 2018), showing different instances of the same target category to sender and receiver (e.g., Choi et al., 2018; Rodríguez Luna et al., 2020), or using data augmentation techniques to reduce low-level image similarities (Dessi et al., 2021).

These developments are related to the more general issue of understanding the content and the effect of emergent languages, which is not straightforward without predetermined symbol meanings. It seems that, without additional pressures, emergent protocols are "decidedly not interpretable or compositional" (Kottur et al., 2017, p. 1). In addition, they are anti-efficient, in that more frequent meanings are encoded in longer messages contrary to Zipf's Law of Abbreviation observed in natural language (Chaabouni et al., 2019). Moreover, in more complex setups, agents may appear to communicate—messages contain information about subsequent actions—but the messages have no impact on the environment or the receiving agents (Lowe et al., 2019). These findings underline that emergent protocols must be analyzed carefully, under consideration of counter-intuitive strategies.

Despite these differences, there are also interesting parallels between natural language and the languages emerging in deep learning simulations. For example, using populations of DNN agents with varying intra- and intergroup connectivity, Harding Graesser et al. (2019) simulated different contact linguistic phenomena, including the emergence of creole languages. As another example, DNN agents that play a reference game with colors develop color-naming systems with an accuracy/complexity tradeoff highly similar to that of human languages (Chaabouni et al., 2021). This result only holds for discrete communication. If the agents use continuous messages, the emergent protocol is more complex and less efficient. Such parallels can help to identify properties of language that do not depend on biological constraints but generally emerge from discrete communication or certain patterns of communicative interaction between agents.

1.4 Introduction to the case studies

Chapters 2–4 will present the three main publications of my Ph.D. project. The projects are heterogeneous not only in the questions they address. They use different modeling approaches, focus either on language learning or language emergence, and differ in how comfortably they fit into cognitive science versus AI research. I still like to treat them as *case studies* because they all arise from the same motivation to take a more holistic approach when modeling language-related phenomena with ANNs. Crucially, all case studies employ communication games and instantiate the idea that language must be understood through its use in the world.

All case studies rely on a reference game setup and implement agents as ANNs. Case study 1 uses this setup to simulate word learning, and case studies 2 and 3 use this setup to simulate language emergence. In the language emergence simulations, a sender and a receiver agent develop a communication system for playing a the game. The sender maps referents to words and the receiver maps words to referents. In the language learning simulations, a communication system is already in place and the game is reduced to the receiver, which must learn to identify the correct referent for a given word. In each case study, either the game or the agents are modified to study the relationship between language and other areas of cognition.

Figure 1.10 provides an overview of the case studies.

- *Case study 1* focuses on the interface between language and pragmatic reasoning. The project proposes a novel computational word learning model that combines associative learning with pragmatic reasoning as formalized by the RSA model. Pragmatic reasoning has been proposed as an explanation for the ME bias phenomenon (see Figure 1.10.CS1). The model is used to study the ME bias in pragmatic word learners but also to demonstrate how an ME bias can be integrated into ANNs.
- *Case study 2* also focuses on the interface between language and pragmatic reasoning. The standard reference game is modified such that the sender has information about the context. In particular, it has information about which object properties are relevant and should be communicated. Depending on which properties are relevant, a more or less specific level of reference is appropriate (see Figure 1.10.CS2). We study whether the agents learn to refer to objects at different levels of specificity and what strategies they use.
- *Case study 3* focuses on the interface between language and visual perception. We manipulate the agents' visual representations and study how that changes the emerging languages (Figure 1.10.CS3, left to right). In addition, we manipulate the communication protocols and study how this affects the agents' visual representations (Figure 1.10.CS3, right to left).

Together, these case studies illustrate how (deep) neural network models of language learning and language emergence can be modified to capture effects of social reasoning, context, and perception. The remainder of this section embeds each case study into relevant background information.

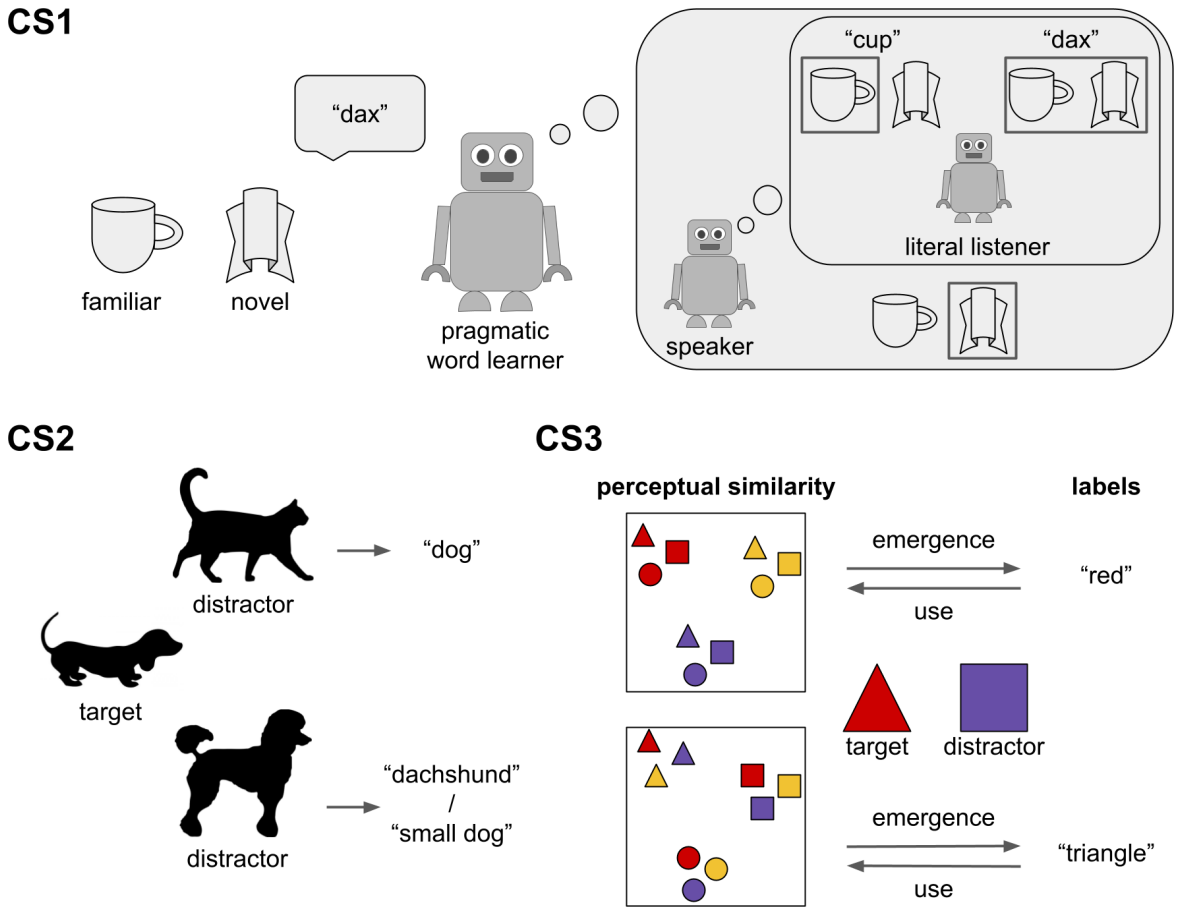


Figure 1.10: Overview of the case studies. **CS1)** If a pragmatic word learner is presented with a familiar object (cup) and a novel object (dax), and hears a novel label ("dax"), they can infer that this label likely refers to the unfamiliar object. The first case study investigates this mechanism with a computational pragmatic word learning model that integrates gradient-based learning into the RSA framework. **CS2)** Different contexts require different levels of specificity when referring to an object. For example, "dog" can be used to disambiguate the target in the upper context but not in the lower context. More specific expressions can be generated holistically ("dachshund") or compositionally ("small dog"). The second case study develops a new type of reference game, where the sender receives information about the target plus the context, to study the emergence of hierarchical reference systems in DNN agents. **CS3)** The third case study investigates the mutual influence between perception and emergent communication in DNN agents. Working with simple stimuli, we introduce different visual biases by manipulating perceived object similarities. If objects are most easily distinguished based on a certain feature like color (top) or shape (bottom), that feature should be the preferred dimension for communication. Similarly, if specific features are preferred for communication, representational space might adapt to reflect these linguistic categories.

1.4.1 Case study 1: Mutual exclusivity in pragmatic agents

One question that has been discussed extensively in the context of word learning, is whether the meanings of words are learned through hypothesis testing or association (e.g., Westermann & Twomey, 2017). In the associative learning framework, the learner forms graded associations between words and meanings. These associations are strengthened or weakened based on co-occurrence. Connectionist models naturally fall into this category. As discussed above, they demonstrate how complex word learning phenomena and generalization behaviors can arise from simple associative learning processes. Associative accounts are well suited to capture incremental learning processes, varying degrees of confidence a learner might have, and tolerance to noise (Frank et al., 2009).

In the hypothesis testing framework, upon hearing a word, the learner constructs a hypothesis space of possible meanings and evaluates these hypotheses based on the context and some principle of rational inference. *Deductive inference* approaches (e.g., Siskind, 1996) rely on word learning biases as strong prior constraints. Mechanisms like the ME bias, the whole object assumption, and the shape bias allow the learner to eliminate possible extensions of a word. Xu and Tenenbaum (2007) introduced *Bayesian inference* as an alternative to deduction. The authors argued that word learning biases are in many cases not sufficient to fix a reference class for a word. Bayesian inference provides a natural way to further constrain the hypothesis space. I will give an example from the paper. After observing multiple co-occurrences between the word "fep" and a dalmatian, according to deductive as well as associative approaches, dalmatians, dogs, and animals form equally likely referent classes.[10] By scoring hypotheses based on how well they predict the observed data, Bayesian inference will evaluate the class of dalmatians as most likely. At the same time, word learning biases can be built into the model by changing prior probabilities. Compared to deductive inference, Bayesian inference provides a more general framework for constraining the hypothesis space and—like associative models—allows for graded knowledge about word meanings, rather than keeping or eliminating hypotheses in a binary fashion.

Treating word learning as Bayesian inference instead of associative learning has the advantage that socio-pragmatic aspects of word learning can easily be integrated. While Xu and Tenenbaum (2007) only discussed this point, Frank et al. (2009) explicitly modeled the joint inference of word meanings and speaker intentions. The model was evaluated on realistic data (CHILDES dataset, MacWhinney, 2000) and was able to account for several word learning phenomena, including the ME bias, fast mapping, and the use of speaker intention as a cue to word meaning. Among others, these findings suggest that the ME response can arise from pragmatic reasoning and there is no need to postulate an innate bias. However, while Bayesian inference constitutes a general cognitive principle, that successfully explains human behavior in various domains of inductive learning, it is unclear whether young children, especially at the earliest stages of word learning, actually have the cognitive capacity to perform such inferences (for a review, see Bohn & Frank, 2019).

There is no simple answer as to which approach is "better". Some proponents of associative models argue that additional mechanisms are necessary to capture the social aspects of word learning (e.g., Regier, 2003, 2005) and proponents of Bayesian word learning admit that there might be a trajectory from early associative to more mature Bayesian learning (e.g., Xu & Tenenbaum, 2007). Considering the argument by McMurray et al. (2012) that word learning and referent selection are relatively independent processes, it could also be that associative learning and Bayesian inference are complementary.

The first case study picks up the idea that associative learning and Bayesian inference are complementary processes. In probabilistic pragmatic word learning models (like the one

[10] Markman (1989) argued that children are biased toward inferring basic-level categories (Rosch et al., 1976) to address this problem. In that case, however, it remains unclear how children learn the labels of subordinate or superordinate categories.

by Frank et al., 2009), agents reason about the speaker's intention when inferring the most likely lexicon based on a history of observed word-meaning co-occurrences. These models do not differentiate between learning and inference. Our model, in contrast, embeds pragmatic inference into a long-term associative learning process and thereby naturally incorporates different time scales for learning (incremental-associative) and inference (online-pragmatic).

We use the model to study the ME bias phenomenon from a pragmatic perspective. In the first part of the project, we use symbolic word and object representations and a simple associative model, where the only free parameters are associative weights between words and objects. Our results show that pragmatic reasoning can (at least qualitatively) account for the ME bias phenomenon, its developmental trajectory, and the effects of different modulating factors such as vocabulary size and the amount of exposure to familiar words. In the second part of the project, we develop a proof-of-concept implementation to demonstrate that pragmatic reasoning can also lead to an ME bias in DNNs. This project successfully combines the modeling frameworks of hypothesis testing and associative learning. Together, they provide a mechanistic explanation of the ME bias phenomenon, which could also help solve the ME bias challenge in machine learning.

1.4.2 Case study 2: Referring to objects at different levels of specificity

In natural language, multiple objects can be grouped under the same label and multiple labels can be used to describe the same objects. For example, a poodle and a dalmatian can both be called "dog", and the same dog can be called "Fido", "dalmatian", "dog", or "animal".

In most word learning and language emergence simulations, the formation *one-to-many associations* between words and objects depends on the modeler's choice of input representation. These associations form naturally if different instances of the same object are presented to the agent. Already in the early word learning model by Plunkett et al. (1992), distorted versions of different prototypes were presented and the agent learned to assign the same label to distortions of the same prototype. The same effect arises in exemplar-based models (e.g., Regier, 2005), or in DNN models trained on different images showing the same object category (e.g., Gulordava et al., 2020; Vong & Lake, 2020). In language emergence simulations, one-to-many associations arise if different object instances are presented that play the *same functional role* in the communication game. For example, in the mushroom world by Cangelosi and Parisi (1998), some mushrooms are edible and some are poisonous. The agents map different feature vectors onto the same label according to this distinction. Hence, the formation of one-to-many associations between words and objects can be simulated with ANNs when choosing the right input representation and task.

Learning multiple labels for the same object, i.e. the formation of *many-to-one* associations between words and objects is more challenging. In the context of language acquisition, the ability to learn second labels is of particular interest as it constitutes one of the word learning boosters in the second year of life. Theories that hypothesize innate word learning biases, such

as the ME bias, cannot easily explain how, over time, these biases weaken or come to be applied more selectively (e.g., Lewis et al., 2020; McMurray et al., 2012). Learning second labels can also be challenging for ANN-based models. If labels are learned sequentially, standard multi-layer networks suffer from catastrophic forgetting (see Section 1.3.3). Some models avoid this issue by making the simplifying assumption of one-to-one associations between words and objects. But, as we saw above, catastrophic forgetting can also be overcome by modified architectures. The exemplar-based model by Regier (2005) learned second labels and the model by McMurray et al. (2012) even learned labels with taxonomic structure by combining association and real-time competition dynamics.

In language emergence simulations, trivial *many-to-one* associations between words and objects in the form of synonymous expressions tend to emerge naturally (e.g., Choi et al., 2018). The question remains, how non-trivial many-to-one associations can emerge, where labels have overlapping but different meaning extensions. More generally, it is unclear how agents can learn to refer to objects at different levels of specificity, regardless of whether expressions at different levels are lexicalized (e.g. taxonomies) or not (e.g. use of modifiers). Referencing objects at different levels of specificity is a pragmatic phenomenon, and the context—next to other factors such as object typicality or cost of utterance alternatives—plays an important role in determining specificity (e.g., Graf et al., 2016). For example, human participants in a reference game develop more abstract referring expressions if the differences between targets and distractors are coarse and more specific expressions if the differences are fine-grained (Hawkins et al., 2018). In early, pre-deep-learning simulations, the game setup is typically not complex enough to study the emergence of such "hierarchical" reference systems. In modern, deep learning simulations, in turn, the sender often does not have access to the context, in order to avoid communication about low-level object differences. Even in simulations where context information is available, the emergence of reference at different levels of specificity has not yet been investigated.

To fill this gap, the second case study investigates the emergence of hierarchical reference systems in language emergence simulations with DNN agents. Context is extremely important for hierarchical reference systems to emerge. However, as discussed above, standard reference game setups with DNN agents typically outsource effects of the context from language production to interpretation. The receiver may choose different types of objects for the same message depending on the distractors but the sender will always produce the same distribution over messages for the same target object. To address this problem, we develop a modified game, the *hierarchical reference game*. In the hierarchical reference game, agents have to communicate *concepts*, defined as compositional and hierarchical abstractions over primitive feature values (e.g. "red dotted circle" $\subset$ "red circle" $\subset$ "circle"). In each game, a target object and a context are sampled randomly. The sender knows which object features are relevant in the given context and the receiver has to select a matching object (e.g. a "red circle" or a "red triangle" if the target concept is "red").

Our analyses show that the agents learn to play the game and can even generalize to novel concepts. In particular, they develop systematic hierarchical reference systems if the input space

is large enough. The agents use different strategies to achieve systematic abstraction. Among others, their protocols reflect the compositional input structure (e.g. they might say "a" for circle, "b" for red, and "ab" for red circle). Overall, the project highlights the importance of contextualization for the emergence of flexible language use.

1.4.3 Case study 3: Interactions between language and perception

An intriguing question that pertains to both language acquisition and language emergence concerns the status of labels in cognition: Are (perceptual) object representations and object labels separate from each other or do they interact? The evidence presented in Section 1.2 strongly suggests an interaction.

In the word learning literature, different theories have been proposed. According to one proposal, labels are separate from perceptual representations: first, the child builds a repertoire of concepts and then learns to describe them (Waxman & Gelman, 2009). This view corresponds to models that learn separate representations for words and objects and then learn a mapping between the two while leaving the actual representations untouched (e.g., Li et al., 2007; Mayor & Plunkett, 2010). Other proposals recognize that labels have an impact on the conceptual system. Sloutsky and colleagues proposed that labels act in the same way as perceptual features—at least in the early stages of word learning—to explain the effects of labels on infants' categorization behavior (Sloutsky & Fisher, 2004, 2012; Sloutsky et al., 2001). This proposal was instantiated in a model by Gliozzi et al. (2009): First, the feature vectors that encode visual and auditory information are concatenated and then processed together by a neural network.[11] An alternative view suggests separate but interacting processing streams for labels and objects, implemented for example in the encoder-decoder model by Plunkett et al. (1992). This view was also implemented in a model by Westermann and Mareschal (2012, 2014), which was used explicitly to study how labels affect perceived similarity. In that model, labels had the effect of warping representational space to cluster objects with the same label while maintaining internal category structure. By now, it is commonly accepted that labels do influence category formation but the exact mechanisms are still subject to debate.

The effect of language on perceptual object representations has also been studied in the context of language emergence. Experiments with robots show that agents can develop perceptually grounded categories through interactions with their environment and that they can learn to coordinate and align these categorical repertoires by playing communication games (for an overview, see Steels, 2005). Cangelosi and Harnad (2000) directly compared the visual representations learned through sensorimotor experience to visual representations learned through a combination of sensorimotor experience *and* language. Like Westermann and Mareschal (2014) in the context of language acquisition, they describe that labels warp the

[11] The input vector is projected into the same self-organizing map. See Appendix D for more information on self-organizing maps.

representational space and induce categorical structure beyond the structure emerging through sensorimotor interaction alone.

The interaction between perceptual object representations and emergent communication has received little attention in current deep learning simulations. Some designs work with symbolic input and do not model perceptual processes (e.g., Bouchacourt & Baroni, 2019; Chaabouni et al., 2020a; Chaabouni et al., 2019; Kharitonov & Baroni, 2020). Simulations that do work with pixel inputs largely rely on image features generated by pretrained CNNs and do not model the effects of communication on these representations (e.g., Havrylov & Titov, 2017; Lazaridou et al., 2017; Rodríguez Luna et al., 2020). In exceptional cases, the agents' vision modules are trained from scratch based on feedback from the communication game (e.g., Choi et al., 2018; Dessi et al., 2021; Lazaridou et al., 2018). Of those, only Dessi et al. (2021) studied the emergent visual representations, focusing on their quality as out-of-the-box features for downstream vision tasks (e.g. classification). A systematic investigation of the interaction between the formation of language and the formation of perceptual representations, similar to the pre-deep-learning models described above, is still missing.

The third case study addresses this point by investigating the interaction between emergent communication and perceptual representations in a reference game with DNN agents. While interactions between concept formation, visual representations, and (emergent) language have already been studied in simple ANNs, the use of deeper networks allows us to study these interactions in the face of more complex, structured messages and raw pixel inputs.

The agents in our simulations communicate about images of 3D geometric shapes with well-defined properties (color, scale, and shape). This simple object structure allows us to systematically manipulate the agents' visual representations in terms of perceived object similarities. We study the effects of variations in visual representations on emergent communication and vice versa the effects of variations in communication on visual representations. Our simulations mirror several empirically observed phenomena of bidirectional influence between language and perception. In particular, communication induces categorical perception, which also occurred in language emergence and word learning simulations with simpler models (Cangelosi & Harnad, 2000; Westermann & Mareschal, 2014). The mutual influence between language and perception can also lead to mutual improvement: some visual representations lead to more successful communication, and some protocols lead to more accurate visual representations. Thus, aside from accounting for co-adaptation effects between language and perception, our results point out ways to improve visual representation learning and emergent communication in artificial agents.

2 Case study 1: Mutual exclusivity in pragmatic agents

This chapter presents case study 1. The chapter starts with a lay summary, which is followed by the content of the publication:

Ohmer, X., Franke, M., and König, P. (2022). Mutual exclusivity in pragmatic agents. *Cognitive Science*, *46*(1), e13069. https://doi.org/10.1111/cogs.13069.

2.1 Lay summary

Learning the meanings of words is a central part of language acquisition. Words are typically uttered in a context with many potential referents. To learn the meaning of a new word, children have to understand whether it refers to an object in the scene, and if so whether it refers to the object as a whole or only a feature or part of it. Different mechanisms help children to perform this disambiguation. Among others, children tend to assume that objects only have one label. Under this assumption, new words should only be mapped to objects for which they have not yet learned a label. This mechanism is known as the *mutual exclusivity* (ME) bias.

There are different theories about how the ME bias arises. One of them claims that the bias arises from pragmatic reasoning processes. Pragmatic reasoning describes the process of taking into account the context and the intention of the speaker when trying to infer the meaning of what is being said. According to that explanation, the word learner reasons that the speaker could have used a familiar word to describe a familiar object. Hence, if the speaker uses a new word, this word must refer to an unfamiliar object.

In this project, we build a computational model of a word learner that applies pragmatic reasoning and use the model to investigate the ME bias. We combine the pragmatic reasoning component with an associative learning component: Every time the model identifies the correct meaning of a word, the corresponding word-meaning association is strengthened a little. While both pragmatic and associative word learning models already exist, this combination is novel. It allows us to model two relatively independent parts of word learning. The associative component simulates the incremental but sustained learning of word-meaning mappings, while the pragmatic component simulates children's ability to reason about the speaker's intention and the context to infer the meaning of a word in a given moment (although they might forget it later). Taken together, we can study whether pragmatic reasoning leads to an ME bias as well as how the ME bias develops over time according to a pragmatic explanation.

We run simulations in which our model learns several word-meaning associations. We measure whether the model has an ME bias by calculating its probability of selecting an unfamiliar object

when hearing a novel word and subtracting the probability of randomly selecting an unfamiliar object. The model is said to have an ME bias if the resulting value is larger than zero, with higher values indicating a stronger bias. The model displays an ME bias throughout the entire learning process, which confirms the idea that the bias can arise from pragmatic reasoning. The model also captures the effects of different factors that influence the ME bias strength in human word learners. For example, if the word learner knows more words, the bias becomes stronger. Similarly, if the learner is more often confronted with familiar words, the bias becomes stronger. Testing different model configurations shows that the model can only account for all these effects if it becomes increasingly certain that associations between words and objects are one-to-one, based on past observations.

The mutual exclusivity bias is also of interest to machine learning researchers. Artificial neural networks (ANNs), which are the most popular machine learning models, display a behavior that is opposite to the ME bias. ANNs are typically used to learn a mapping from certain inputs to certain outputs. For example, a model could receive images as input and output a label for the depicted objects. If ANNs receive a new input (in word learning: a word) they tend to map it to a familiar output (in word learning: an object). This anti-ME bias slows down the network's learning process, especially if it continuously encounters new inputs. In the second part of the project, We build an ANN implementation of a pragmatic word leaner and demonstrate that pragmatic reasoning can also be used as an ME bias mechanism in ANNs.

Our work shows that pragmatic reasoning can lead to an ME bias both in simple associative and in more complex ANN models. Several analyses highlight parallels between our model and human word learning, speaking in favor of a pragmatic perspective on word learning.

2.2 Abstract

One of the great challenges in word learning is that words are typically uttered in a context with many potential referents. Children's tendency to associate novel words with novel referents, which is taken to reflect a mutual exclusivity (ME) bias, forms a useful disambiguation mechanism. We study semantic learning in pragmatic agents—combining the Rational Speech Act model with gradient-based learning—and explore the conditions under which such agents show an ME bias. This approach provides a framework for investigating a pragmatic account of the ME bias in humans but also for building artificial agents that display an ME bias. A series of analyses demonstrates striking parallels between our model and human word learning regarding several aspects relevant to the ME bias phenomenon: online inference, long-term learning, and developmental effects. By testing different implementations, we find that two components, pragmatic online inference and incremental collection of evidence for one-to-one correspondences between words and referents, play an important role in modeling the developmental trajectory of the ME bias. Finally, we outline an extension of our model to a deep neural network architecture that can process more naturalistic visual and linguistic input. Until now, in contrast to children, deep neural networks have needed indirect access to (supposed to be novel) test inputs during training to display an ME bias. Our model is the first one to do so without using this manipulation.

2.3 Introduction

Word learning is central to language acquisition. The core problem of word learning is that novel words are typically encountered in situations that offer a multitude of potential referents. To learn the meaning of a new word children must understand whether the word refers to an object in the scene and if so whether it refers to the object as a whole or rather to a specific feature or part of it. Next to social, linguistic, and attentional information, inductive biases help children disambiguate the meanings of novel words (e.g., Bloom, 2000; Hollich et al., 2000; Markman, 1991). The mutual exclusivity (ME) bias accords with the property of language that word-meaning mappings tend to be bijective (Clark, 1987). In the now classical ME paradigm, Markman and Wachtel (1988) showed that when children are presented with two objects and know the label for one of them, they will tend to associate a new label with the other object. In additional experiments, they demonstrated that this inference mechanism not only helps children to learn labels for whole objects but also for object parts and features. Accordingly, the ME bias supports the identification of referents in ambiguous context and applies to a wide variety of situations, making it a key word learning mechanism.

With some due simplification, the ME-bias behavior observed in the classical ME paradigm is this. An agent is familiar with and able to recognize words $\{w_1, \ldots, w_n\}$ and objects $\{o_1, \ldots, o_n\}$. The agent has learned to associate, for simplicity, w_i with o_i for all $1 \leq i \leq n$ in a one-to-one mapping. The agent has never encountered the word w_{n+1} before, but recognizes it as different

from any word in $\{w_1, \ldots, w_n\}$. Similarly, mutatis mutandis, for novel object o_{n+1}. The agent now perceives a speaker use w_{n+1} in a referential context C, where $C \subseteq \{o_1, \ldots, o_n, o_{n+1}\}$ with $o_{n+1} \in C$. The agent shows an ME bias in the classical ME paradigm if they show an *ME response* by associating the novel object o_{n+1} with the novel word w_{n+1}. Notice that the ME bias, thus delineated, presupposes that objects and words are perceived as familiar or novel, and that the speaker's goal of using word w_{n+1} is to refer to exactly one element from the referential context C.

There has been a long-standing discussion about the mechanism underlying the ME bias (Lewis et al., 2020, for an overview, see). Two proposals dominate the literature. Under the first proposal, the ME bias is a manifestation of an innate or early emerging constraint. One important version of this proposal is the *lexical constraint account*. It posits that children are biased to consider only lexica with one-to-one mappings between words and objects (Markman & Wachtel, 1988; Markman et al., 2003). If only one-to-one mappings between words and objects are feasible, then from the assumed one-to-one associations of w_i to o_i for all $1 \leq i \leq n$, the only plausible mapping of novel word w_{n+1} is to novel meaning o_{n+1}. A second prominent approach uses a pragmatic explanation. Pragmatics studies how humans reason about each other's intentions and take into account contextual factors when producing and interpreting utterances (Clark, 1996). Social-pragmatic reasoning abilities, such as understanding others' intentions and theory-of-mind reasoning, play an important role in language acquisition (e.g., Bloom, 2000; Bohn & Frank, 2019; Clark & Amaral, 2010; Tomasello, 2001). Under the traditional *pragmatic inference account*, the ME bias is based on the assumption that the speaker follows cooperative principles of communication. According to Clark (1988), the relevant principles are the *Principle of Conventionality* (speakers use the same words to refer to the same objects) and the *Principle of Contrast* (every two types of objects contrast in meaning). Following a pragmatic explanation along these lines, the ME response, associating novel word w_{n+1} to novel meaning o_{n+1}, is supported by the pragmatic argument that *if* the speaker would have wanted to refer to a known object $o_i \neq o_{n+1}$, they would have used word w_i (by Contrast and Conventionality).

The classical ME paradigm constitutes a context-dependent inference task. The ME bias phenomenon, however, is embedded in the long-term learning process of language acquisition. In general, it has been argued that long-term word learning and online inference in situations of referential ambiguity operate on different time scales and are not straightforwardly dependent on each other (Frank et al., 2009; Gulordava et al., 2020; McMurray et al., 2012). For example, given the classical ME paradigm, children are able to identify the correct referent, arguably via online inference, but show poor retention of this novel word-meaning mapping when tested five minutes later (Horst & Samuelson, 2008). While accurate referent selection can be achieved by excluding competitors in a given context, retention requires encoding the association between the novel word and object (Axelsson et al., 2012). In addition, several studies provide insights into the developmental trajectory of the ME bias. When children of different age groups are tested within the same ME experiment, the ME bias consistently increases with age (Bion et al., 2013; Frank et al., 2016; Grassmann et al., 2015; Halberda, 2003; Lewis et al., 2020). This effect seems to be driven by two factors that increase with age: vocabulary size and linguistic exposure,

that is familiarity with "familiar" objects and labels (Grassmann et al., 2015; Lewis et al., 2020). To provide a full account of the ME bias phenomenon, it is important to consider these relations between online inference and long-term learning as well.

To address the ME bias puzzle, together with modulating developmental effects such as vocabulary size and exposure, we develop a computational model that comprises both pragmatic reasoning and long-term associative learning. In line with probabilistic pragmatic word learning models, we rely on the Rational Speech Act (RSA) framework as a computational mechanism for the ME bias. In the RSA framework, speakers and listeners recursively reason about each other's intention to enrich the literal meanings of utterances, using Bayesian inference. It has successfully modeled various pragmatic phenomena (e.g., Scontras et al., 2018), and probabilistic pragmatic word learning models either rely on the framework itself (Smith et al., 2013) or similar formalizations (Frank et al., 2009; Lewis & Frank, 2013). In all these models, agents take into account the speaker's perspective when inferring the most likely lexicon from a history of observed word-meaning pairs. Lacking situation-time dynamics, there is no differentiation between long-term learning and online inference. Our model, in contrast, embeds (pragmatic) online inference into a long-term learning process, where lexical associations are formed incrementally in a gradient-based learning process. As a result, the model can account for processes at different time scales and can be compared to behavioral data from the classical, inferential ME paradigm as well as developmental studies. This comparison allows us to evaluate which of the mechanisms hypothesized to play a role in the ME bias, and more generally word learning, agree or disagree with psychological reality.

We introduce different pragmatic agent models with explicit lexical representations and use them for an in-depth investigation of the ME bias phenomenon from a pragmatic perspective. One implementation we evaluate is the same as in our earlier work (Ohmer et al., 2020), where agents have a fixed lexicon size, corresponding to the number of words and objects in the data set. In addition, we test a novel implementation with a dynamically growing lexicon. There is an important conceptual difference between the two implementations. The pragmatic inference account describes how pragmatic reasoning during online inference leads to an ME bias. This *inferential ME bias* occurs in both implementations. However, pragmatic reasoning has different long-term learning effects on the two lexicon types. In the fixed lexicon, but not the dynamic lexicon, an additional *lexical ME bias*, as proposed by the lexical constraint account, arises. (Section 2.4.3 provides a detailed explanation of how the two bias components relate to the two lexicon types.) Using these two models, we study the ME bias, its developmental trajectory, and its role in long-term learning; and simultaneously evaluate the influence of inferential and lexical bias components.

The ME bias is also of interest to the machine learning community. Recently, Gandhi and Lake (2020) showed that neural networks lack an ME bias and even have the reverse tendency of selecting familiar labels for novel objects. They further demonstrated that this anti-ME bias slows down learning in various types of networks. As a result, they proposed the *ME bias challenge* for neural networks, which is not only a technical challenge applying to general classification or

translation models but concerns any artificial agent design based on standard neural network architectures. Compared to traditional probabilistic pragmatic models, the gradient-based learning mechanism used by our approach is compatible with neural network training, and we explore it as a potential solution to the challenge.

Recently, deep neural word learning models have been introduced, which try to address the ME bias challenge (Gulordava et al., 2020; Vong & Lake, 2020). These models operate with two networks: a visual module processing raw images and a language module processing words. Whether a word maps to an object is determined by the similarity of their embeddings. In the classical ME paradigm, the child's ability to recognize familiar words and objects, as well as to recognize the novel word and object as such, is given. While the two networks can learn to recognize familiar words and objects, due to the anti-ME bias they are not guaranteed to recognize the novel word, w_{n+1}, as different from any familiar word or the novel object, o_{n+1}, as different from any familiar object. So far, deep neural word learning models use *negative sampling* to solve this problem. In the classical ME paradigm, the bias is tested with words and objects the child is entirely unfamiliar with. In contrast, these models rely on presenting the test items as negative examples, so mismatching combinations of words and objects with explicit negative feedback during training. Negative examples make the network learn that the novel word (object) is not associated with any of the familiar objects (words). However, negative sampling is not an empirically justified assumption about human word learning because it is far from obvious that negative samples are available at a sufficient rate. Neither does it solve the ME bias challenge of building neural networks that map novel inputs onto novel, hitherto unseen outputs. Consequently, it is important to find mechanisms for generating ME-like behavior in neural network architectures that do not require negative sampling.

The remainder of the paper is structured as follows. Section 2.4 introduces the computational pragmatic models with explicit lexical representations and investigates their behavior in a classical ME paradigm, developmental effects, as well as the relation between online inference and long-term learning. Section 2.5 explores a proof-of-concept extension of our approach to a deep neural network architecture. We follow existing joint-embedding space architectures and test whether pragmatic reasoning can also achieve ME in deep neural networks (without negative sampling). Section 2.6 critically assesses the approach taken here and the results obtained from it before Section 2.7 concludes.

2.4 Mutual exclusivity in pragmatic agents with explicit lexical representations

We set out to explain not only the ME bias behavior as it is observed in the classical ME paradigm but also interactions between ME bias and long-term learning, in particular the effects of vocabulary size and linguistic exposure on bias strength, and whether the ME bias supports learning. In the following experiments, we explore these questions from a computational

pragmatic perspective. We develop two different computational pragmatic models, one where the lexicon size is fixed and one where it grows dynamically, both of which embed pragmatic inference into an associative long-term learning process. As this is a novel approach, we initially analyze the mechanisms by which these implementations lead to an ME bias on a theoretical level. We find that they make different assumptions about how pragmatic reasoning causes an ME effect. The computational experiments therefore not only test whether our pragmatic model can account for empirical findings but also which of these assumptions are more plausible.

In our first analysis, we test whether both implementations lead to an ME bias in a long-term learning context. We then proceed to investigate the development of the ME bias over time. As vocabulary size and linguistic exposure have been identified as the main drivers behind the increasing bias strength in early development, we conduct two separate analyses to examine the influence of these two variables on our model. In an additional analysis we test whether, even though making the correct inference is not sufficient for long-term learning, it could still serve as a supporting factor. In a final analysis, we use differences in prediction between the two implementations to identify the pressures of pragmatic reasoning on learning and inference that can best explain empirical findings. To account for the difference between the two processes of long-term learning and online inference, they are monitored separately throughout the analyses.

2.4.1 Pragmatic agent model

The agents in our model feature two main components: a) explicit lexical representations and b) rules of pragmatic behavior telling them how to use these representations to produce and interpret messages. We consider agents with a fixed lexicon size (Ohmer et al., 2020) and also explore an implementation with a dynamically growing lexicon.

Lexical representations

The lexicon B_A of agent A is a matrix providing a mapping between words and objects. Each matrix entry $B(o_i, w_j) \in \mathbb{R}^+$ is an unnormalized value of how appropriate (in a semantic sense) word j is for object i. The matrix entries are the only trainable parameters of the model.

Fixed lexicon. Working with a fixed lexicon size is in line with other word learning models (e.g., Lewis & Frank, 2013; McMurray et al., 2012; Regier, 2005). The agents' lexicon is a matrix of size $N \times N$, where N is the total number of word-object pairs in the training data. Because agents are pragmatic, their reasoning process during learning encompasses all words and objects in the lexicon, even those that have not yet appeared in the learning process, which is similar to the negative sampling strategy employed in neural word learning models. In this implementation, however, the use of negative examples is not modeled explicitly but arises naturally from pragmatic reasoning. While the idea that agents reason about unfamiliar words

and objects is questionable when trying to draw a connection to human word learning, one can defend this approach by arguing that agents are aware that there are unknown states and messages in the world. A fixed lexicon size implements the idea that agents extend their reasoning to future novel inputs for which they reserve lexicon space.

Dynamic lexicon. Alternatively, the dynamic lexicon only encompasses familiar items and is extended for novel inputs. This type of implementation has also been used before (e.g., Kachergis et al., 2012). The agents start out with a minimal lexicon of size 2×2. Every time the agents encounter a novel object or word, the lexicon is extended by one row or column, respectively. The newly created lexicon entries are initialized with the mean of the old entries. As the mean of the lexicon entries changes throughout the word learning process, learning has an indirect effect on associations between novel words and objects. But, unlike with a fixed lexicon, they are not updated in the training process itself. Using the average of the old lexicon as initialization for the new slots is a natural choice because the lexicon entries are not upper-bounded. A constant initialization value would have different effects depending on the hyperparameter setting (which influences the range of values the lexicon entries take on) and the stage of the training process as lexicon entries tend to keep changing over time.

Rules of pragmatic behavior

For modeling pragmatic behavior, the vanilla RSA model is used. In the RSA model, conditional probabilities describe how a speaker maps a state, s, onto an utterance, u, and how a listener maps an utterance onto a state, while they take into account each other's perspective.

$$P_{LL}(s \mid u) \propto [\![u]\!](s) \times P(s)\,, \tag{1a}$$

$$P_{PS}(u \mid s) \propto \exp\left(\alpha \times [\log P_{LL}(s \mid u) - C(u)]\right), \tag{2a}$$

$$P_{PL}(s \mid u) \propto P_{PS}(u \mid s) \times P(s)\,. \tag{3a}$$

At the basis of the recursive reasoning process is a literal listener (LL: 1a) who maps an utterance onto any state for which it is true, at the same time considering the prior probability of that state. In (1a), $[\![u]\!]$(s) is the denotation function returning the truth value of utterance u for state s. A pragmatic speaker (PS: 2a) chooses her utterance such that the probability of being correctly understood by a literal listener is maximized while production cost, $C(u)$, stays low. The parameter $\alpha \in \mathbb{R}^+$ regulates the speaker's optimality. For $\alpha = 0$, the speaker's choices are random, and for $\alpha \to \infty$, she will always select the utterance that yields the maximal probability of being correctly understood by the literal listener. The pragmatic listener (PL: 3a), in turn, interprets an utterance as if coming from a pragmatic speaker, also considering the prior probability of states. We model our agents as pragmatic listeners.

We adapt the vanilla RSA model in several ways. The vanilla model assumes that agents have access to the lexicon which is a truth table of utterances across states. In our case, the agent learns

the (non-negative, real-valued) entries of its lexicon, which it also uses for its internal reasoning process. We assume a flat prior over states, zero costs for every utterance, and set $\alpha = 5$.[1] As our agents only reason about words and objects we change the notation accordingly:

$$P_{LL}(o \mid w, B_{LL}) \propto B_{LL}(o, w)\,, \tag{1b}$$

$$P_{PS}(w \mid o, B_{PS}) \propto P_{LL}(o \mid w, B_{PS})^{\alpha}\,, \tag{2b}$$

$$P_{PL}(o \mid w, B_{PL}) \propto P_{PS}(w \mid o, B_{PL})\,. \tag{3b}$$

The agents are myopic with respect to the possibility of different lexica, that is, they only consider their own current lexicon and do not reason about which lexicon their interlocutor might likely have as in some pragmatic-inferential accounts (Frank & Goodman, 2014; Frank et al., 2009; Lewis & Frank, 2013). We do not use literal listener or pragmatic speaker models directly; they only appear as part of the pragmatic listener's inference process.

2.4.2 Reinforcement learning

We train our agents with reinforcement learning.[2] Following the pragmatic reasoning process in (3b), the agents map an input word onto a probability distribution over objects, $P_{PL}(o \mid w, B_{PL})$, which defines their policy. The agents' selection is sampled from this policy. If they select the right object, they obtain a positive reward, $R = 1$; otherwise they obtain zero reward, $R = 0$. The loss function is defined as the negative expected reward,

$$\mathcal{L}(\theta) = -\mathbb{E}\big[R\big]\,, \tag{4}$$

and the parameters to be optimized correspond to the agents' lexicon entries, $\theta = B_{PL}$. The parameters are updated using REINFORCE (Williams, 1992), which belongs to the family of policy gradient algorithms. In our case, gradients are calculated as

$$\nabla_{\theta}\mathcal{L} = -\mathbb{E}\big[R\,\nabla_{\theta} \ln P_{PL}(o \mid w, \theta)\big]\,. \tag{5}$$

Gradient-based updates lead to incremental changes in the lexical associations between words and objects and simulate the long-term learning process.

[1] Because we use the model in a learning context α also influences the trade-off between exploration and exploitation. While $\alpha = 5$ provides a good balance, most qualitative results are robust across other values ($\alpha \in \{2.5, 10\}$). Additional explanations and analyses are provided in our OSF project.

[2] Under a supervised learning regime, which provides a stronger feedback signal, the qualitative results stay the same. The main quantitative differences are that a) training is faster, b) performance is higher, and c) performance increases more substantially for agents with a fixed lexicon where all possible word-referent mappings are updated at every training step.

2.4.3 Mutual exclusivity in pragmatic reasoning

The consideration of alternative words and meanings by the pragmatic listener, as in (3b), leads to an ME bias during online inference. This inferential ME bias applies equally to implementations with a fixed and a dynamic lexicon. Fig. 2.1.A illustrates this effect, with the help of an example, for the dynamic lexicon implementation. Here, the agent encounters a new word and a new object in the context of familiar objects, which is why the lexicon is extended. The entries in the new column and row are all identical to the average over the existing matrix. Nevertheless, the rows corresponding to familiar objects have entries with above-average values because associations between familiar words and objects have already been formed. That is, a pragmatic speaker would probably have chosen one of the familiar words if referring to one of the familiar objects (see the speaker mapping P_{PS} in the RSA readout shown in Fig 2.1.A). Taking the reasoning process of the pragmatic speaker into account, the pragmatic listener can infer that the new word probably refers to the new object (see the listener mapping P_{PL} in the RSA readout in Fig. 2.1.A) although the last row in the lexicon has no information on the choice of utterance for a new object. Thus, even without a learned lexical association for the input word, the agent has a high probability of selecting the correct object. In other words, pragmatic reasoning causes an ME bias that is independent of lexical constraints.

In our model, pragmatic inference is embedded into a long-term learning process. During each inference, the pragmatic agent consults all word-object associations to infer a policy. Accordingly, the learning signal created by a single inference affects all lexicon entries, as shown for a fixed lexicon implementation in Fig. 2.1.B. In particular, a one-to-one mapping between the current word-object pair is reinforced (green updates), and at the same time, associations between all other words and objects in the lexicon are strengthened (orange updates). As the lexicon has a different structure in the two implementations, the learning process has a different effect. In the dynamic lexicon, only associations between familiar words and objects are updated. In the fixed lexicon, in contrast, the learning process affects all associations, which induces assumptions about associations between unfamiliar words and objects. These assumptions implement a lexical ME bias: Associations between novel words and novel objects become increasingly more likely than associations between novel words and familiar objects or novel objects and familiar words. In conclusion, pragmatic reasoning causes an inferential ME bias regardless of lexicon type, and an additional lexical ME bias for the fixed but not the dynamic lexicon.

The learning process does not induce a lexical ME bias if agents have a dynamic lexicon. Still, it has an indirect effect on the associations between novel words and objects via the initialization mechanism. During learning, associations between words and their referents are strengthened. At the same time, associations between words and other objects are weakened. As true associations are sparse, a majority of the lexicon entries converge to zero. As a consequence, the mean of all lexicon entries—with which new slots are initialized—moves further away from the true association weights, such that under pragmatic principles, the novel word becomes an increasingly less likely choice for any of the familiar samples. Thus, the initialization mechanism

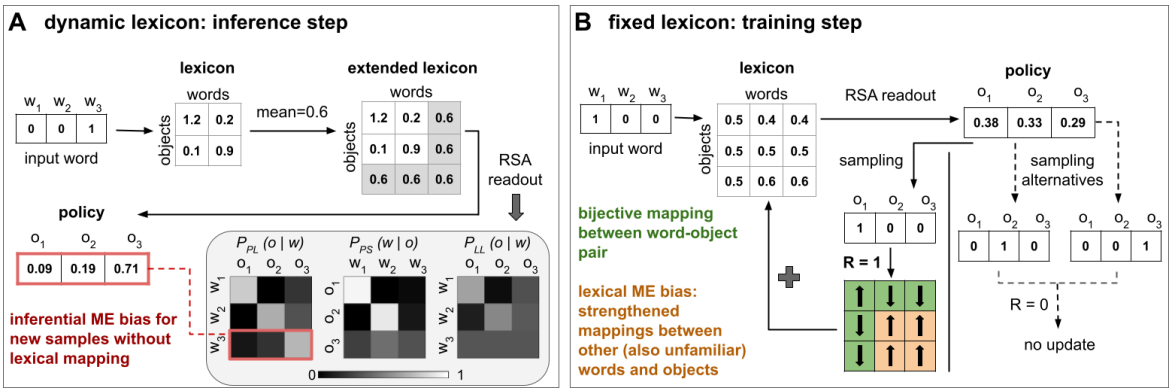


Figure 2.1: A Example of the (pragmatic listener) agent's inference step with a dynamic lexicon. When the agent encounters a new word-object pair, its lexicon is extended by one row and one column. The "RSA readout" shows the recursive reasoning steps that lead to the agent's policy, $P_{PL}(o \mid w)$. The agent reasons about the policy of a pragmatic speaker, $P_{PS}(w \mid o)$, which in turn reasons about the policy of a literal listener, $P_{LL}(o \mid w)$. Note that the matrices here visualize policies and not lexica. Following this inference process, the selection probability for the unfamiliar object is highest when receiving an unfamiliar word. Pragmatic reasoning leads to an inferential ME bias. **B** Example of a training step with a fixed lexicon, adapted from Ohmer et al. (2020). If the listener selects the correct object, it obtains a positive reward, $R = 1$, else $R = 0$. The weight update following a correct selection reinforces one-to-one mappings between the current word-object pair (green) and strengthens associations between all other words and objects in the lexicon (orange). With a fixed lexicon, the update includes unfamiliar words and objects, leading to a lexical ME bias.

in the dynamic lexicon gradually builds up evidence for one-to-one-mappings between words and objects.

In both implementations, the agent reasons pragmatically during learning and inference but the consequences on the lexicon are different. With a fixed lexicon, pragmatic reasoning during learning and inference divides the ME bias into lexical and inferential components. Other fixed lexicon implementations, combining literal listener and pragmatic listener components, are conceivable. A literal learner, performing pragmatic inference will have an exclusively inferential ME bias. A pragmatic learner, performing literal inference will have an exclusively lexical ME bias. And indeed, pragmatic inference can lead to an early emerging ME bias both by its effect on the lexicon during learning and by its effect on the online reasoning process (Ohmer et al., 2020). A fixed lexicon with literal learning and pragmatic inference provides an interesting alternative to the dynamic lexicon implementation, in that it implements an inferential ME bias component without the indirect effects of the initialization mechanism. First and foremost, we are interested in the predictions of fully pragmatic agents with a fixed or dynamic lexicon. In a separate analysis, however, we use different literal-pragmatic combinations for the fixed lexicon to disentangle the contribution of lexical and inferential pressures on the ME bias.

2.4.4 Methods

Agents and training were implemented with Tensorflow 2.0 (Abadi et al., 2015).

Training

Main setup. We train agents on a word learning task to simulate the long-term learning process. We use a simple learning scenario in which the frequencies of words follow Zipf's law (Zipf, 1949)[3]. Objects and words are represented by one-hot vectors and have a predetermined one-to-one mapping such that object i is associated with word i, for $1 \leq i \leq 100$. We define a *sample* as a word-object pair that belongs together. The training set contains all words and objects that the agent can encounter and grows monotonically with the number of epochs. At the beginning, agents are exposed to a single sample, and every k epochs a new sample is added until a total number of $N = 100$ is reached. Given that word-object pairs are always added together, the dynamic lexicon is always extended by one row and column simultaneously. Every epoch consists of a fixed number of trials. At each trial, a word from the current training set is randomly selected and presented to the agent. The agent maps the input word to a policy over all objects in its current lexicon. The lexicon entries are updated using reinforcement learning, as described above. The exposure interval k determines after how many epochs new samples are added to the training set, and thereby regulates how well the agent has learned familiar words when it encounters a novel word. If not mentioned otherwise we report results for $k = 15$, which provides near-optimal learning conditions in that agents have enough time to learn familiar words almost perfectly. We consider other exposure intervals ($k \in \{1, 3, 6, 9, 12, 15\}$) to investigate the relation between linguistic exposure and ME bias and how differences in ME bias relate to differences in learning success. To obtain robust results, we run 100 simulations for every value of $k > 1$ for the dynamic and the fixed lexicon, respectively. For $k = 1$, we run a total of 500 simulations each.

Hyperparameters. Training hyperparameters such as batch size, learning rate, lexicon initialization values, and samples per epoch were selected based on model performance after running a small grid search (for details see Appendix A.1). In general, we found the qualitative results to be very robust across different hyperparameter values. The fixed lexicon entries are initialized as $b_{i,j} = 0.001, 1 \leq i, j \leq 100$, and the dynamic lexicon entries as $b_{i,j} = 0.1, 1 \leq i, j \leq 2$. In every epoch, 1000 word-object pairs are sampled randomly from the current training set. Agents are trained with vanilla stochastic gradient descent (SGD) on batches of 32 examples and with learning rate $\gamma = 0.1$, with one exception where we reduce the learning rate (see below).

Evaluation

We evaluate long-term learning and online inference. Word learning progress can be monitored by the average reward achieved per epoch. As the reward for each trial is either zero or one, the average reward directly corresponds to the proportion of words that are mapped onto

[3] The Zipfian input distribution is arguably a natural assumption about the relative frequency of meanings to be communicated. In previous work, we showed that pragmatic agents develop an ME bias regardless of whether words follow a uniform or a Zipfian distribution. However, we found the advantage of pragmatic reasoning and the resulting ME bias in terms of learning speed to be stronger for a Zipfian input distribution.

the correct object. When evaluating online inference, we are interested in the ME bias. We consider those time points in the learning process where the agent encounters a novel word for the first time. This happens every k epochs as determined by the exposure interval. By tracking the agent's inference at all these time points, we can analyze the ME bias throughout development.

Mutual exclusivity index. We use the ME index (Ohmer et al., 2020), I_{ME}, to quantify the ME bias formally. An ME bias exists if the probability of selecting a novel object upon receiving a novel word is greater than chance:

$$I_{ME} = \frac{p(\text{new object selected} \mid \text{new word}) - p(\text{new object})}{p(\text{familiar object})}.$$

If the probability of selecting a novel object given a novel word is at chance level the ME index is equal to zero and if the entire conditional probability mass is on the new object(s) it is equal to one. We add samples incrementally and evaluate the ME index separately for each new word. With a fixed lexicon, there are various novel objects that the agent can select, and their number decreases as a function of epochs. With a dynamic lexicon, there is only one novel object. The exact formulas for both cases are provided in Appendix A.2.

General versus specific referential contexts We use two different settings to evaluate the ME bias. In the *general-context evaluation*, the referential context C comprises all known words and objects, as well as any novel ones presented currently. Consequently, the agent's selection policy encompasses all objects in the lexicon. The ME bias strength, based on this policy, can be taken to quantify the agent's general assumption that a novel word must refer to a novel object rather than an old one. In contrast, the *specific-context* simulates the classical ME paradigm. The agent is presented with a novel word in the context of just one familiar object (distractor) and one novel object (target), and the agent's policy only encompasses these two objects. At the same time, we limit the pragmatic reasoning process to the novel word, and words that the agent considers plausible for the familiar object. These candidate words are determined by sampling 25 times (independently, with replacement) from the policy of a pragmatic speaker given the familiar object as input. The agent's reasoning process then includes all words that were sampled at least once.[4] Thus, not only the objects but also the words under consideration are contextualized in the specific-context. The results for the specific-context evaluation are obtained by aggregating: for every new word-object pair we create a specific context with every other object in the lexicon as distractor. Differences between the two evaluations provide insights into the role of contextualization in the ME bias phenomenon.

[4] This sampling procedure implements a probabilistic version of applying a relative threshold criterion to the word probability under the speaker's policy. One can think of it as the speaker constructing a context model (fixing which words and objects are salient alternatives for pragmatic reasoning) by collecting a number of candidate words that easily come to mind.

2.4.5 Analyses and results

Does pragmatic reasoning lead to ME in a long-term associative learning process?

Pragmatic reasoning can explain the ME bias and has been implemented successfully using the RSA model. We start by testing whether both our RSA-based agent models, the fixed lexicon implementation and the dynamic lexicon implementation, successfully realize this ME bias mechanism in a long-term word learning context. To measure whether the agents have an ME bias throughout learning, we calculate the ME index distributions over the course of training.

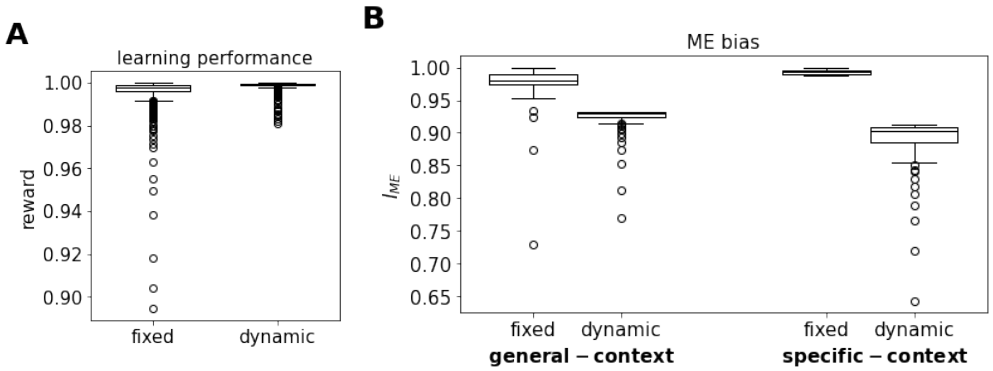


Figure 2.2: Distributions of rewards (**A**) and ME indices (**B**) across training. The distributions are averaged across 100 simulations with maximal exposure ($k = 15$). Results are shown for the fixed lexicon implementation as well as the dynamic lexicon implementation, and the ME index is evaluated in general-context and specific-context. The outliers in both plots represent low values that arise at the beginning or at the end of training (see Figure 2.3).

Fig. 2.2.A shows that agents with both types of lexica achieve very high rewards throughout the training process, which means that they manage to consolidate all the words they encounter. Fig. 2.2.B shows the ME indices for the general-context and the specific-context evaluation. Agents with a fixed lexicon have a stronger ME bias than agents with a dynamic lexicon in both types of evaluation. Still, the ME bias is very strong for either lexicon type, with all mean ME indices lying above 0.88. All distributions lie completely above zero, which means that the agents display an ME bias throughout the entire training process. The results extend our earlier finding that pragmatic agents with a fixed lexicon have an ME bias in a general-context evaluation (Ohmer et al., 2020), to agents with a dynamic lexicon and to a specific-context evaluation simulating the classical ME paradigm.

How does vocabulary size influence the ME bias?

We are interested in the predictions of the two implementations regarding the developmental trajectory of the ME bias. To test whether they are in line with the empirical observation that the ME bias increases as children grow older and have a larger vocabulary, we look at the agents' word learning performance and ME bias over time. Under the near-optimal learning conditions considered here, the agents map novel words onto the correct referents almost as soon as they have been added to the training set. We can therefore approximate the words that are part of

the agent's vocabulary by the words in the training set. As a new sample is introduced every 15 epochs, the vocabulary size grows monotonically with the number of training epochs.

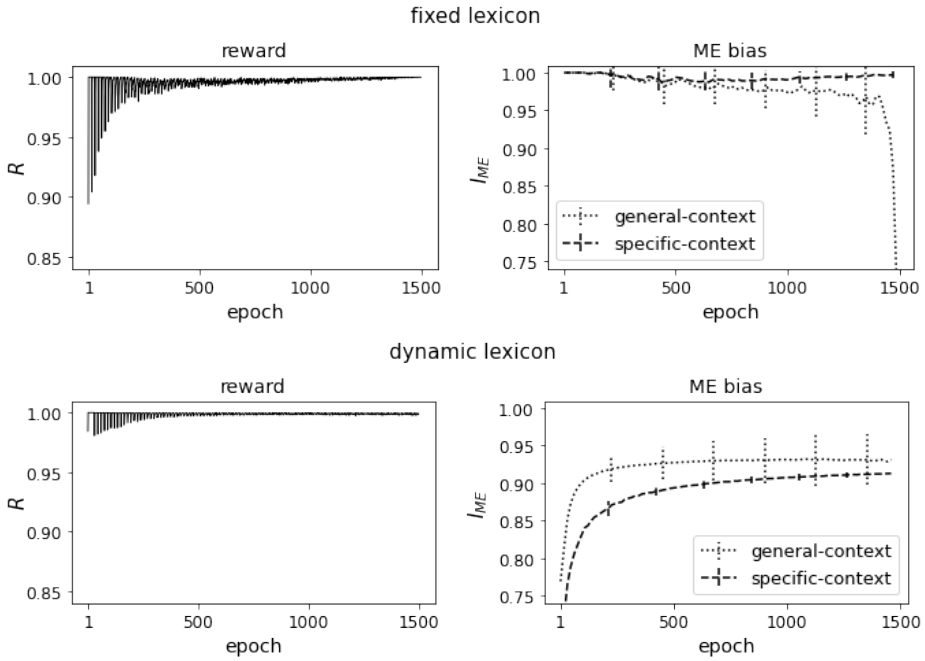


Figure 2.3: Reward and ME bias strength over the course of training for both lexicon types. Shown are rewards (left) and ME indices for the general- and the specific-context evaluation (right). For the specific-context evaluation, we average the ME index across all referential games with the same target. Reported are mean values across 100 runs with exposure interval $k = 15$, and for the ME indices, we include standard deviations.

Fig. 2.3 shows average rewards and ME indices over the course of training. We begin by clarifying technical peculiarities that arise with a fixed lexicon size. Early in the training process, rewards drop more strongly for agents with a fixed lexicon (top left) than for agents with a dynamic lexicon (bottom left) when new words are added to the training set. At training onset, agents with a fixed lexicon have far more selection options than agents with a dynamic lexicon. Over time, as more associations are established, the lexical ME bias reduces the number of potential referents and the meaning of novel words can more easily be inferred. Looking at the ME indices over time, the ME bias is relatively stable except for the initial and the final training phases. With a fixed lexicon (top right), the ME bias drops for the final samples in the general-context but not the specific-context evaluation. The sudden drop is due to a ceiling effect of the ME index calculation. For a maximal ME index, the selection probability for each new sample must be equal to $1/l$ where l is the number of free slots. When the number of free slots reduces to one, an extreme increase in the selection probability of the remaining novel samples is required for the ME bias to be maximal. In the specific-context evaluation, in contrast, only the relative difference in selection probability between the new sample and a randomly selected old sample is important. Both effects, the increasing rewards as well as the drop in ME bias, are an effect of the fixed lexicon design and are not conceptually relevant.

Let us now return to the question of interest. What predictions do the two implementations make about the relationship between vocabulary size and ME bias? For the fixed lexicon, we find that the ME index is high throughout training in both evaluations (ignoring the ceiling

effect) (Fig. 2.3, top right). In essence, the size of the lexicon does not influence the ME bias. In contrast, for the dynamic lexicon, the ME bias increases with the size of the lexicon following a curve with decreasing incline (Fig. 2.3, bottom right). While this effect only holds for the initial training phase in the general-context evaluation (dotted line), it holds throughout the entire training process in the specific-context evaluation (dashed line). An additional test reveals that if the policy from the general-context evaluation is used to contrast the selection probability of the novel object separately against the selection probability of each familiar object, the ME bias converges and does not increase continuously. Hence, for the ME bias to keep increasing, not only the target-distractor contrast itself is important, but also the contextualization of the pragmatic reasoning process, in that the agent only considers the objects present in the scene and only alternative words that would be a sensible choice for these objects. As associations between familiar words and objects are learned increasingly well, the agent considers fewer alternative words for the distractor, which facilitates the correct inference. In summary, agents with a fixed lexicon size have a constant ME bias, regardless of vocabulary size, while for agents with a dynamic lexicon the ME bias increases with vocabulary size, in particular when the inference process is contextualized.

How does the amount of linguistic exposure influence the ME bias?

The ME bias has also been shown to increase with the amount of exposure to familiar words, which can be measured by varying the child's familiarity with the distractor label in the classical ME paradigm. Children's linguistic exposure is based on a developmental history of word learning experiences, whereas direct manipulation in the lab is only possible over a short time span (e.g., Lewis et al., 2020). In our simulations, in contrast, we can directly regulate the amount of exposure by varying the exposure interval k and can do so throughout the long-term learning process. The higher the exposure interval, the more time the agents have to reinforce the relation between familiar words and referents, and thus a relation between word knowledge level and ME bias can be established.

Fig. 2.4 shows the results for different levels of exposure, $k \in \{3, 6, 9, 12, 15\}$, with larger intervals corresponding to darker shades. With respect to the agents' word learning performance (left column), we find that performance is lower when the agents have little exposure to the training samples (light-colored lines) as compared to high exposure to the training samples (dark-colored lines). Hence, the exposure interval can, in fact, be used to regulate the word-level knowledge for familiar words. Looking at the rewards over time, agents with a fixed lexicon improve to near-optimal performance regardless of exposure level. As explained for the exposure interval $k = 15$ above, the increase in reward stems from a continuous elimination of potential referents that arises when a lexical ME bias acts on a fixed lexicon. The performance of agents with a dynamic lexicon, however, decreases under low exposure, with a faster decrease for smaller intervals. With respect to the agents' ME bias (center and right column), there is a consistent pattern for both types of lexica as well as both evaluations: More exposure to familiar samples

increases the ME bias for novel samples. Visually, this pattern is reflected in the monotonically darker shades of red toward higher ME indices.

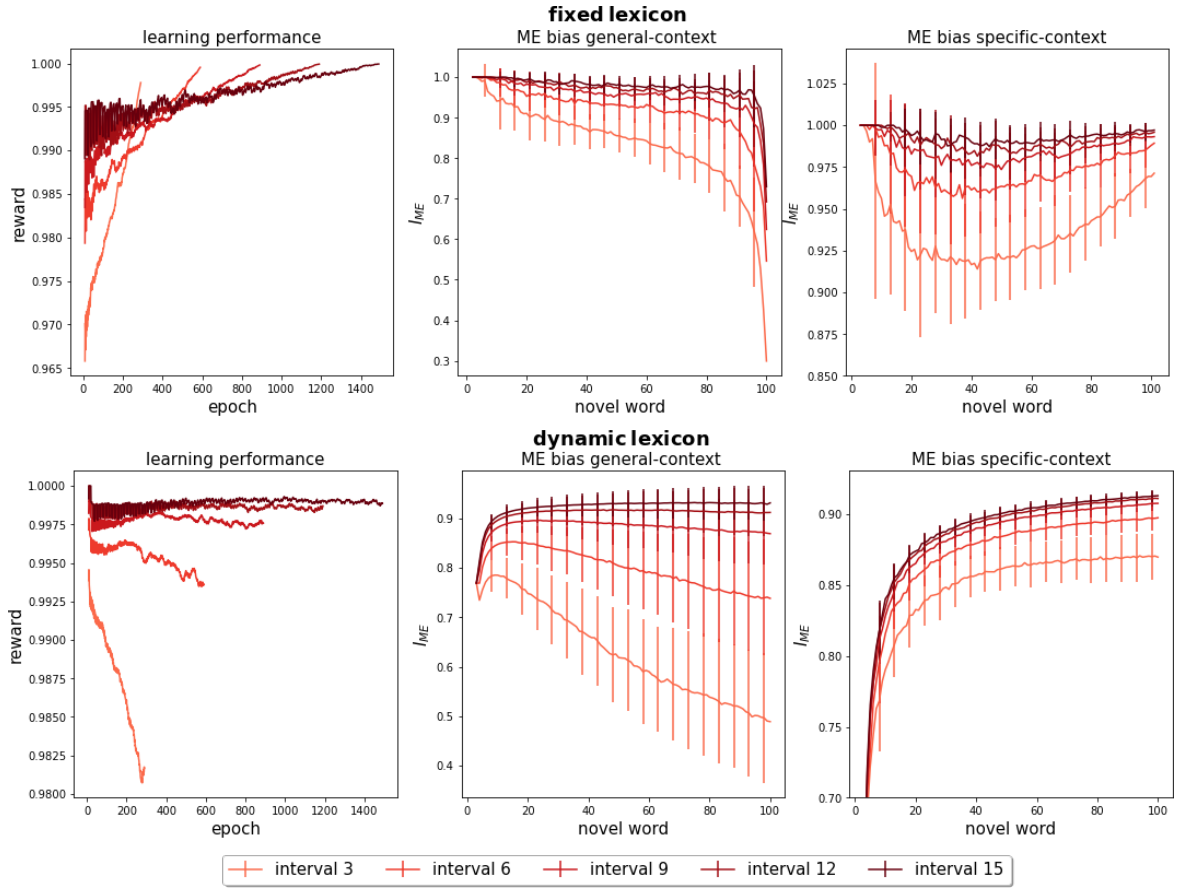


Figure 2.4: Performance and ME bias strength over the course of training for different amounts of linguistic exposure: $k \in \{3, 6, 9, 12, 15\}$. Plotted are mean and standard deviation across 100 runs for each exposure level. The darkest red lines repeat the results in Fig. 2.3 apart from different scaling on the x-axis. Shown are performance (left), bias strength as given by the ME index in the general-context evaluation (center), and the specific-context evaluation (right). Rewards are smoothed by calculating the moving average across 19 epochs.

We examine the role of linguistic exposure in more detail by comparing the agents' selection probabilities when tested for ME under maximal ($k = 15$) and minimal ($k = 1$) exposure levels, as shown in Fig. 2.5. In particular, we are interested in why the agents sometimes map novel words onto familiar objects. For the general-context, the policy provides a full selection distribution across all objects. For the specific-context, we calculate the policy for a referential game with each familiar object as distractor, respectively. Differences in the target selection probability indicate how much each object competes with the target object. Given a high amount of exposure, the samples in the training set are learned almost perfectly before novel samples are encountered. Therefore, selection probability mass is concentrated almost exclusively on the novel object(s). In the general-context evaluation, the high selection probability for novel objects is indicated by the probability mass lying on the upper right triangular matrix for the fixed lexicon, or on the diagonal for the dynamic lexicon (2.5.A, left column). In the specific-context evaluation, the target selection probability is consistently very high across all referential games (2.5.B, left column). Given a small amount of exposure, the samples in the training set have not

been learned perfectly by the time a novel sample is introduced. Accordingly, in an ME bias evaluation, agents may select the novel object but they may also select any of the objects they have not learned. In the general-context evaluation, some of the selection probability mass is allocated to other recently introduced objects (2.5.A, right column), and in the specific-context evaluation, agents perform much worse if the distractor is one of these more recent objects (2.5.B, right column). Overall, the ME bias increases with the amount of exposure to familiar samples and insufficient exposure makes the selection probabilities leak from novel to unconsolidated objects.

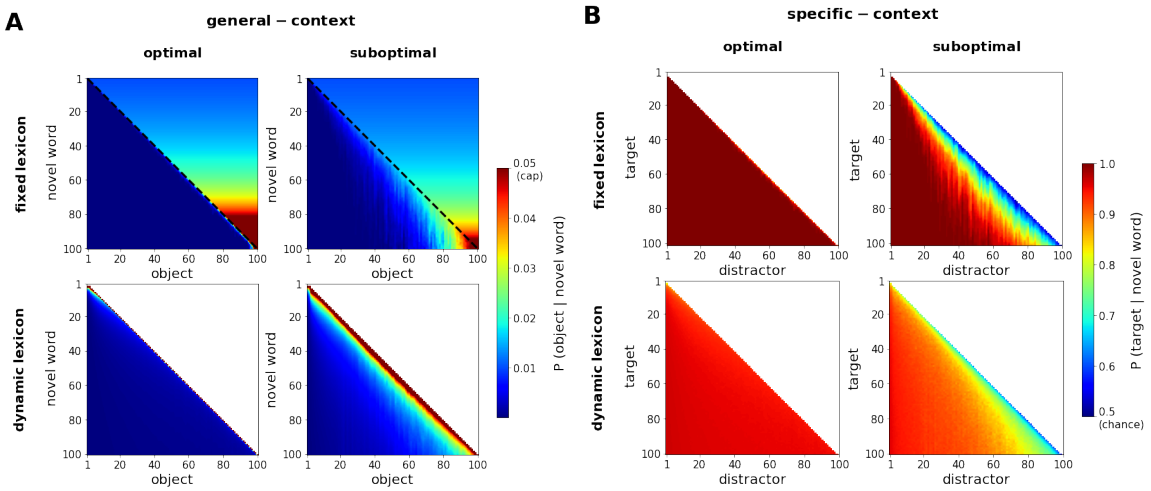


Figure 2.5: Comparison of the ME bias in terms of agents' selection policies, averaged across 100 runs. Selection probabilities are evaluated for every novel word-object pair that is added to the training set. In both figure parts, results are shown for the fixed (top) and the dynamic (bottom) lexicon when learning is optimal ($k = 15$; left) or suboptimal ($k = 1$; right). **A** For the general-context, we display the agents' selection probabilities given the novel word as input. We cap the probabilities at 0.05 to make differences below that value more visible. The y-axis indicates the introduction of a new sample to the training set. In the *optimal* learning condition, this happens every 15 training epochs and in the *suboptimal* learning condition every epoch. For every sample on the y-axis, the selection probabilities in the ME bias evaluation are plotted along the x-axis. **B** For the specific-context evaluation, we display the target selection probability for each referential game, with the novel object as target on the y-axis, and each (familiar) distractor object on the x-axis.

Does an ME bias during online inference support long-term learning?

The ME bias supports the fast mapping of words to objects during online inference. We investigate whether this inferential advantage also supports long-term learning, by relating differences in ME bias strength to differences in learning success. For this purpose, we divide the training data into two categories, words for which the agent has a weak ME bias ($0 < I_{ME} < 0.5$) upon first encounter, and words for which the agent has a strong ME bias ($I_{ME} \geq 0.5$) upon first encounter.[5] We consider a scenario with little linguistic exposure ($k = 1$) to make sure that not all samples are learned immediately, such that there is enough variation in learning success. For the simulations with a fixed lexicon, we additionally reduce the learning rate to $\gamma = 0.0001$ to achieve that. We then measure learning success, in terms of whether a word-object association is learned in the long run, and learning duration, in terms of the number of epochs it takes until

[5] The two categories cover all words since $I_{ME} > 0$ without exception.

the association is stable. A word-object association counts as learned when the word is mapped onto the correct object in more than 99% of the cases in all remaining epochs.

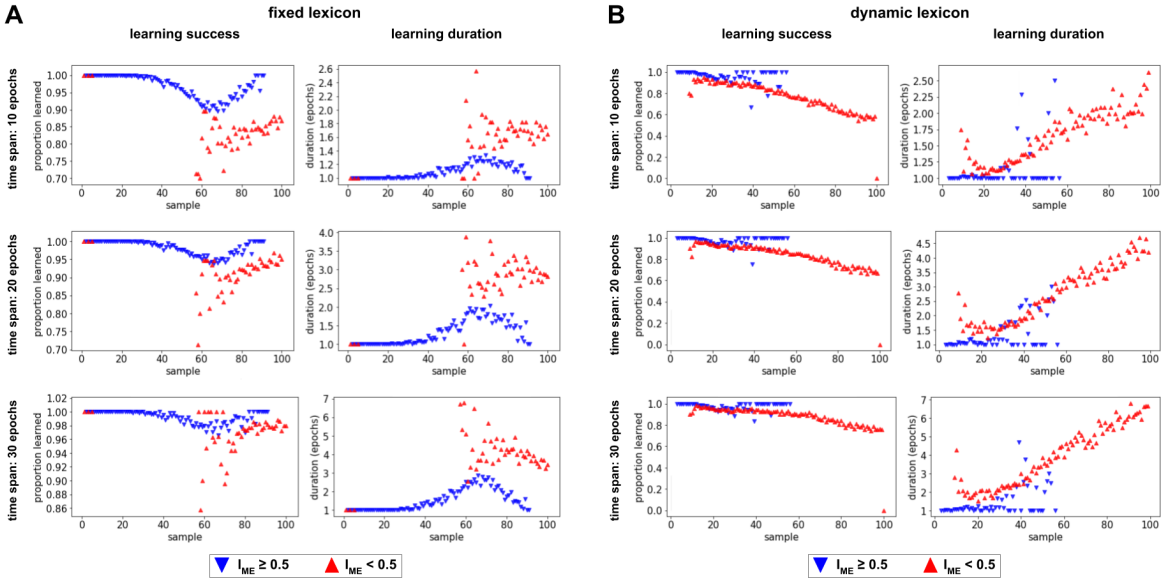


Figure 2.6: Learning success and learning durations with respect to ME bias strength. Shown are averages across 500 simulations with an exposure interval of $k = 1$ for the fixed lexicon implementation (**A**) and the dynamic lexicon implementation (**B**) respectively. For the fixed lexicon the learning rate was reduced to $\gamma = 0.0001$ to obtain a significant number of samples that are not learned immediately. The rows indicate different time spans in which the samples had to be learned after they were first encountered: 10, 20, or 30 epochs. Learning success is measured as the percentage of samples learned across simulations and learning duration as the average number of epochs until samples were learned. For a sample to be learned it must be mapped correctly 99% of the time in all remaining epochs. To display meaningful statistics, we only calculate learning success and duration for a minimum of five samples (1% of the simulations) with either strong or weak ME bias.

Fig. 2.6 shows the results for different time spans in which the agent must learn the samples, with each row corresponding to one time span (10, 20, and 30 epochs). Results for the fixed lexicon are given in Fig. 2.6.A and results for the dynamic lexicon in Fig. 2.6.B. Blue triangles correspond to words with a strong bias and red triangles to words with a weak bias. When children infer a certain word-meaning mapping in context, they do not necessarily remember this association (Horst & Samuelson, 2008). We find a similar behavior in our model. Even words for which the agent has a strong ME bias ($I_{ME} > 0.5$, blue triangles) are not always learned, in the sense that they are not consistently mapped onto the correct referent in the long run. The difference between inference and learning success arises because changes to the lexicon, based on the associative learning process, are incremental and operate on much slower time scales than the inference process. If the agent does not map the novel word sufficiently often onto the correct referent, the association is not reinforced strongly enough, even though the mapping upon first exposure was correct. But the longer the available time span, the more objects are learned. Looking at the relation between ME bias and long-term learning success, we find that a strong ME bias increases the learning success rate and decreases the learning duration. These improvements can be observed for both implementations and for all time spans.

Experiments with adults show that word learning is less successful and slower when referential uncertainty is high (Smith et al., 2011). The current analysis allows us to establish a link between

uncertainty and learning success via the ME bias. There is high referential ambiguity under suboptimal learning conditions: the agent does not know whether the novel word refers to the novel object or any of the old objects that it has not learned. This uncertainty, in turn, has a negative impact on the ME bias, and as a consequence on long-term learning.

How do lexical and inferential pressures influence the ME bias?

As discussed in Section 2.4.3, pragmatic reasoning can cause an ME bias via lexical and inferential pressures. The inferential pressure exists regardless of lexicon type but the lexical pressure arises only in the fixed lexicon because associations of unknown words and objects are updated in the learning process. Still, the dynamic lexicon accumulates evidence for one-to-one mappings via its initialization mechanism. Our goal is to identify how differences in lexical and inferential pressures for ME influence the developmental trajectory of the ME bias. Considering the results above, the main difference between the two implementations is that the dynamic lexicon predicts an increase in ME bias strength across development, whereas the fixed lexicon does not. We perform an ablation study for the fixed lexicon implementation by removing the agent's pragmatic reasoning ability during either learning or inference. The pragmatic reasoning process is replaced by that of a literal listener in the RSA model, as in (1b). A literal learner performing pragmatic inference will only have an inferential ME bias, while a pragmatic learner performing literal inference will only have a lexical ME bias. This allows us to disentangle the role of lexical and inferential pressures toward ME from other factors.

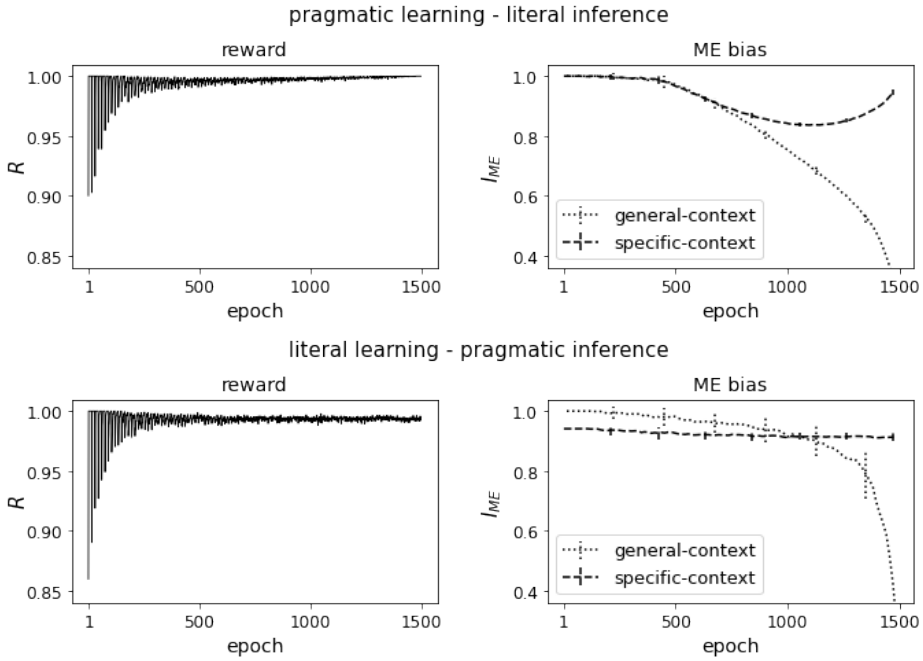


Figure 2.7: Average rewards and ME indices across training for an agent with a fixed lexicon applying pragmatic reasoning during learning but not inference (top row) or applying pragmatic reasoning during inference but not learning (bottom row). Training was conducted with an exposure interval of $k = 15$ epochs. We show average results across 100 runs, including standard deviations for the ME indices.

Fig. 2.7 (top row) shows the word learning performance (left) and ME bias strength (right) across training for an agent that uses pragmatic reasoning during long-term learning but not for inference. On average, word learning is successful throughout training as rewards remain constantly near-optimal. Initially, ME indices of both evaluations are also maximal, due to lexical constraints induced by pragmatic considerations during learning. However, further into the training, both ME indices start to decrease. The general-context ME index continues to decrease monotonically, whereas the specific-context ME index recovers. This pattern is very similar to the results for a fully pragmatic agent with a fixed lexicon under suboptimal learning conditions (see Fig. 2.4, fixed lexicon, $k = 3$). For the fully pragmatic agent, the decrease in general-context ME bias arises because one-to-one mappings between familiar words and objects are not strengthened sufficiently to fully exclude familiar objects as referents. For the pragmatic-literal agent, the decrease arises because familiar objects cannot be excluded as confidently in a literal reasoning process, which does not take into account that familiar objects are already "occupied" by a familiar word. The increase in specific-context ME bias is due to the fixed lexicon size, and not conceptually relevant. Even though the lexical constraint account does not commonly assume the involvement of pragmatic reasoning, the simulations underline that while lexical constraints can cause an ME effect, additional assumptions must be made to explain why it increases in strength.

Fig. 2.7 (bottom row) shows word learning performance (left) and ME bias strength (right) across training for an agent that uses pragmatic reasoning only in the online inference process but not during learning. High rewards indicate that learning is also successful for this combination. We compare the agent's ME bias to that of an agent with a dynamic lexicon (see Fig. 2.3, dynamic lexicon). The general-context ME bias is relatively constant throughout development for both implementations. Yet, while the dynamic lexicon predicts an increasing ME bias in the classical ME paradigm, the literal-pragmatic combination predicts a constant ME bias. Even though both agents largely rely on inferential pressures toward ME, agents with a dynamic lexicon collect evidence for one-to-one mappings via the initialization mechanism, whereas agents with a fixed lexicon do not. In conclusion, pragmatic online inference can cause an ME bias throughout development without a need for additional lexical pressures; but the developmental trajectory of the ME bias can only be accounted for if the pragmatic inference process includes the increasing evidence for one-to-one mappings in the lexicon.

2.5 Mutual exclusivity in pragmatic neural network agents

The models discussed in Section 2.4, from now on called *explicit lexicon models*, are limited by the simplicity of their input representations. They neither capture how human or artificial agents can learn new words from raw visual and linguistic input nor how the ME bias arises from such inputs. There are two different parts to the ME bias phenomenon. In a first step, the agent must recognize that visual and linguistic input represent novel *types* and not novel instances of familiar types. This process is closely related to the problem of out-of-distribution detection in

machine learning models (e.g., DeVries & Taylor, 2018; Hendrycks & Gimpel, 2017; Liang et al., 2018). In a second step, the agent must map the novel word to the novel object, which is the actual ME effect. The explicit lexicon models can capture the second step, but not the first one, which must rely on perceptual similarities and possibly common-sense knowledge.

The explicit lexicon models can in principle be combined with neural network modules as they rely on the same gradient-based learning mechanism. For example, objects displayed in images and words recorded as text could be processed by dedicated networks mapping them onto the corresponding row or column in the lexicon. This approach faces two immediate problems. First, end-to-end training is difficult as mapping onto specific slots of the lexicon requires using the non-differentiable argmax function. So, either training is not end-to-end or the argmax operation must be approximated, for example, using the *Gumbel-softmax trick* (Jang et al., 2017). The second problem is specific to modeling the ME bias. If neural networks are trained to process the visual or linguistic input, they will fail to recognize novel inputs due to the diagnosed anti-ME bias. For example, an image classification network will map objects from novel categories with high confidence onto familiar categories. However, if the agent does not recognize an object as novel, it cannot use the ME bias.

To overcome these problems, we use continuous word and object representations that can exploit similarity relations in the input space. To perform pragmatic reasoning on these representations, their association strength (corresponding to the lexicon entries) is determined by calculating the similarities between these representations in a joint embedding space. With this architecture, end-to-end training is possible. We expect pragmatic reasoning to cause an ME bias also in this setup. As certain word and object representations become very similar over the course of training, pragmatic reasoning makes the use of novel words for these objects unlikely. We run an experiment to test this hypothesis and additionally examine the influence of negative sampling on our model.

2.5.1 Neural pragmatic agent model

The model, as shown in Fig. 2.8, consists of three main components: a vision module (orange), a language module (blue), and a pragmatic reasoning module (green). The agent learns new word-object mappings by receiving a word and trying to select the correct referent from several objects given by the context. The vision and the language module map their respective inputs into a joint embedding space, where lexical association strength is determined by the similarity of the embeddings. In the pragmatic reasoning module, the RSA model is used to calculate a policy for the different objects under the given input word, taking into account alternative input words that could have been used.

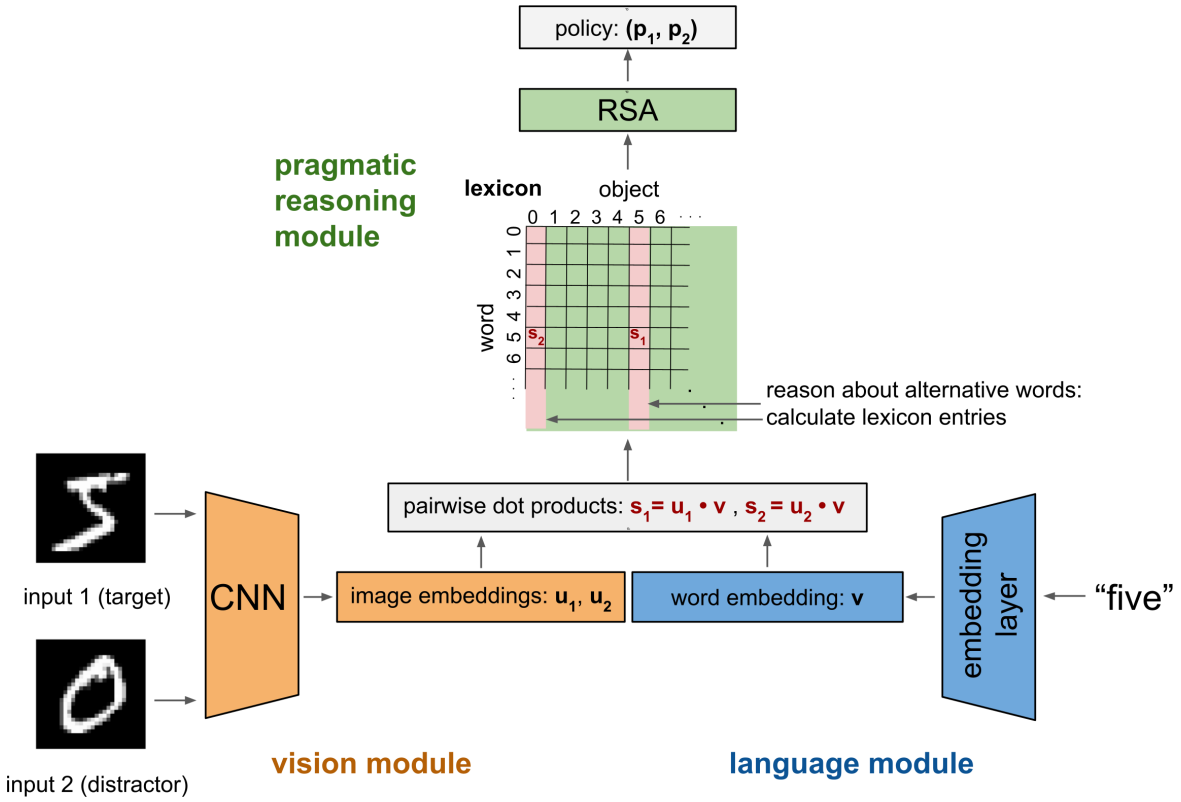


Figure 2.8: Visualization of architecture and training setup for the neural network model.

Vision and language modules

Vision module. The vision module maps raw pixel input onto an image embedding. We pretrain a convolutional neural network (CNN) on the input data using supervised learning.[6] The CNN consists of two convolutional layers followed by a dense layer and the final output layer (see Appendix A.3 for details on model and training hyperparameters). We use the activations of the fully connected layer to extract the image features. These features are mapped into the joint embedding space by an additional fully connected layer with sigmoid activation function.

Language module. The language module consists of an embedding layer mapping the integer symbol inputs onto continuous vectors. An additional fully connected layer with sigmoid activation function maps these word representations into the joint embedding space.

Bounding the embedding space. Often representations are unbounded in a joint embedding space. When learning a lexicon, associations between learned words and objects can take on extreme values over time. In the dynamic lexicon implementation, we initialize novel lexicon entries with the lexicon's mean value. Here, in contrast, we cannot control how novel slots are initialized, that is, what values the embeddings of unknown words and objects take on. As

[6] Pretraining the CNN facilitates training the remaining model parameters. It can be done without loss of generality, given that we are interested in how the ME bias arises when associations between words and objects are learned, independent of how the visual features of these objects are extracted.

a consequence, also the similarities (associations) between unknown words and objects are unconstrained. We find that bounding the embedding space by using a sigmoid output function, instead of a linear one, is in our case sufficient to provide a working initialization.

Pragmatic module

Again, our agent is implemented as a pragmatic listener in the RSA framework. As the pragmatic reasoning process involves the literal listener and the pragmatic speaker, we need to express these formulas based on our neural network architecture. We assume that the agent receives a single input word, w, in a context with multiple objects, $C = \{o_1, ..., o_k\}$. Given word embedding $\mathbf{v} \in \mathbb{R}^m$ and image embeddings $\mathbf{u}_1, ..., \mathbf{u}_k \in \mathbb{R}^m$, we can calculate the similarity values between the word embedding and each image embedding $s_1, ..., s_k \in \mathbb{R}$ with $s_i = \exp(\mathbf{u}_i^T \mathbf{v})$. With a generic optimality parameter α, this leads to the following reformulation of (1b)–(3b):

$$P_{LL}(o \mid w, C) \propto s \,, \tag{1c}$$

$$P_{PS}(w \mid o, C) \propto P_{LL}(o \mid w, C)^{\alpha} \,, \tag{2c}$$

$$P_{PL}(o \mid w, C) \propto P_{PS}(w \mid o, C) \,. \tag{3c}$$

2.5.2 Methods

Again, agents and training were implemented with Tensorflow 2.0.

Training

Main setup. We train our agents on a referential game (see Fig. 2.8). During each round of the game, the agent receives a word, the target object (referred to by the word), and a distractor object as input. The agent outputs a selection probability for the two objects, and the actual selection is sampled from this policy. If the agent selects the target object, it receives a positive reward, $R = 1$; otherwise it receives zero reward, $R = 0$. Again, the agent is trained with REINFORCE using Equations (4) and (5), with the only difference that the parameters to be optimized, θ, correspond to the neural network weights instead of the lexicon entries. We use the images of the MNIST data set (LeCun et al., 2010) as objects. These images contain $70,000$ handwritten examples of the digits 0–9, with a train/test ratio of $60,000/10,000$ and a size of 28×28 pixels. In our setup, the world consists of 10 different objects, corresponding to the different digits and 20 possible words, corresponding to 20 distinct symbols (0–19), both of which are uniformly distributed. Our training and test sets contain digits 0–8, and nine randomly selected, distinct words, which are assigned to these objects. Selection and assignment of words vary between, but not within runs. The test set is used to measure how well the network generalizes to novel examples of digits 0–8. The remaining object, digit 9, and the remaining words are reserved for evaluating the ME bias and form a separate data set. Reserving multiple words and a novel

object for the evaluation simulates a world in which there are many potential names for an object. By holding out the images of digit 9, we have train/test sets of $54,051/8,991$ images. To generate the referential games, each image is used as the target once and paired with a random distractor showing a different digit. At every training trial, the agent's input consists of one of the nine words in the training set, an image of the digit that word refers to (target), and an image of a different digit.

Hyperparameters. The embedding layer of the language module as well as the two fully connected sigmoid layers mapping word and object representations into the joint embedding space each have dimensionality 32. All network parameters are initialized randomly. The network is trained with Adam optimizer, learning rate $\gamma = 0.0001$, and batch size 64. Training proceeds for 100 epochs. All parameters were selected by hand.

Evaluation

To evaluate the ME bias, we compile referential games with a novel input word and number 9 as the target. We measure the ME bias as the correct selection probability in this test setup. Precisely, the correct selection probability is calculated by pairing each of the examples of digit 9 with a random distractor from the test set as well as a random novel input word, and averaging the results across these test games. Pairing the novel object with different potential novel names provides a more robust ME bias estimate as random differences in the embeddings of novel words are averaged out.

Experiments

We train the agent on the referential games as described above and evaluate whether it has an ME bias. By default, the agent's pragmatic reasoning step encompasses the word and objects given by the context as well as all other words in its lexicon, so the remaining words in the training set (assigned to digits 0–8). As the number of words and objects is small, taking into account all alternative messages is not too costly; for more complex worlds sampling may become necessary. For the explicit lexicon models, we know $\alpha = 5$ to be a suitable optimality parameter from earlier work, and the grid search further showed that results are not very sensitive with respect to optimality. Without such information for the neural network implementation, we test different optimality values, $\alpha \in \{5, 10, 15\}$. In addition, to evaluate whether the ME bias can be attributed to the pragmatic reasoning ability of the agent we run the same experiment with a literal agent as given by (1c) and compare the results. For the pragmatic agent with optimality $\alpha = 5$, we test modified versions of the negative sampling strategies employed by Gulordava et al. (2020). For negative sampling of words, the agent takes into account *all* possible words in its reasoning process, not just the ones in the training set. For negative sampling of objects,

the images of number 9 already appear as distractors during training. We also test negative sampling of both words and objects. In total, we run 25 simulations for each variation.

2.5.3 Results

Fig. 2.9 shows the results for the pragmatic neural network agents. The top row shows the training rewards and the bottom row the strength of the ME bias, both over time. All runs converge to maximal accuracy on the training data and reach test accuracies (not in the figure) greater than 99.8%. Looking at the rewards (top left), it is not surprising that pragmatic agents learn faster than literal agents. For a literal listener, the learning update only affects the representations of the current training input. The pragmatic listener samples alternative words the speaker could have used. Accordingly, the learning update not only affects the representations of the currently present word and objects but also the representations of these alternative words. Pragmatic agents with different optimality parameters (top left) as well as different negative sampling strategies (top right) learn approximately equally fast, with a slight advantage for higher optimality parameters.

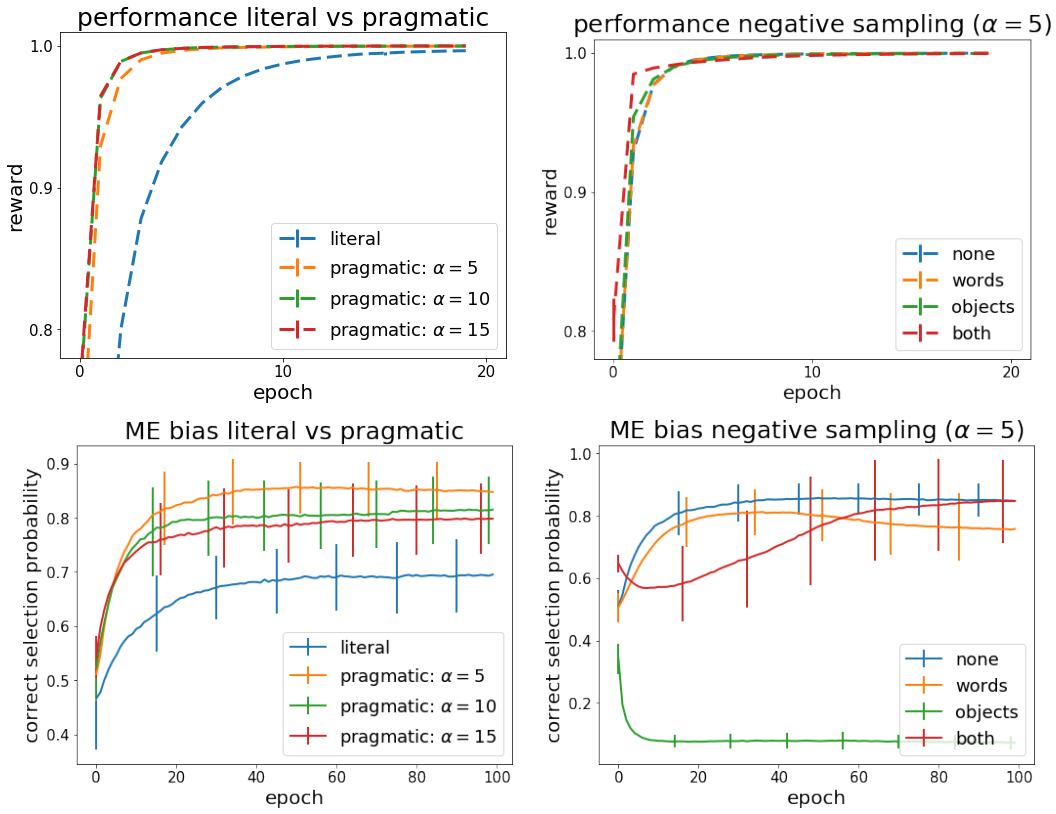


Figure 2.9: Rewards (top) and ME bias (bottom) for the neural network architecture. All values are averaged across 25 runs, and standard deviations are displayed by error bars; no error bars are visible for the rewards due to very little variation. The left column compares pragmatic agents with different optimality parameters and a literal agent. The right column compares different negative sampling strategies for a pragmatic agent with optimality $\alpha = 5$: no negative sampling, negative sampling of words, negative sampling of objects, or negative sampling of words and objects (both).

There is no clear relationship between performance and ME bias. While all agents achieve maximal rewards, the correct selection probabilities in the ME paradigm vary strongly. The different agents trained without negative sampling (left column) all display an ME bias. The strength of the bias varies with the optimality parameter, with higher optimality leading to a lower bias. With higher optimality parameters, small distances in the embedding space are amplified strongly by the exponentiation with a large α. Then both word–target and word–distractor similarity may become zero, such that pragmatic reasoning cannot take effect. Interestingly, even the literal agents have a weak bias and make correct selections on average 69.5% of the trials at the end of training. The ME bias of the literal agent arises from the structure of the embedding space. Looking at the literal agent's lexicon (see Appendix A.4), it turns out that representations of the novel words lie closer to representations of the novel object than to those of familiar objects. Further research is needed to understand why this pattern arises. Still, the pragmatic agents have a consistently higher bias, with 79.8% ($\alpha = 15$), 81.5% ($\alpha = 10$), and 84.7% ($\alpha = 5$) at the end of training.

Looking at the different sampling strategies (bottom right), training without negative sampling surprisingly leads to the strongest bias, with negative sampling of words and negative sampling of both words and objects being close or equal. If negative sampling is used only for objects, the agent develops a strong negative ME bias. Appendix A.4 shows visualizations of the learned lexica using different sampling strategies. Overall, it seems that pragmatic reasoning alone induces enough competition between novel and familiar objects. As discussed by Gulordava et al. (2020), negative sampling of objects introduces an anti-polysemy bias, while negative sampling of words introduces an anti-synonymy bias. Given that the ME bias is an anti-synonymy bias, this distinction explains why negative sampling of words or both words and objects leads to a much higher ME bias than negative sampling of objects. When negative sampling of objects is used, the ME bias drops far below chance. The representations for these objects move so far away in the embedding space that novel words eventually lie closer to the distractors than the target. In conclusion, our main result is that pragmatic reasoning alone is sufficient for the agent to develop an ME bias and negative sampling leads to no further increase in bias strength.

2.6 Discussion

We provide a new computational pragmatic model of the ME bias that combines insights from cognitive models of language use and modern machine learning techniques. We use agent models with explicit lexical representations to demonstrate that pragmatic reasoning not only leads to an ME bias in the classical ME paradigm but can also capture important aspects of the relation between ME bias and long-term word learning. In line with empirical findings, our model makes the following predictions: a) the ME bias increases with the agent's exposure to familiar words and objects, b) the ME bias increases with the agent's vocabulary size, and c) correct inference does not guarantee long-term learning. The different implementations with fixed and dynamic lexica allow the modeler to choose between different assumptions on how

pragmatic reasoning causes ME—only via an inferential bias or also via a lexical bias. Further analysis of these competing pressures reveals that an inferential ME bias alone is sufficient to predict an increase in ME bias strength across development if evidence for one-to-one mappings—collected throughout learning—is used in the inference process. We additionally show that a strong ME bias during online inference positively influences learning success and duration. Pragmatic reasoning may therefore constitute a useful ME bias mechanism for machine learning models. As a proof-of-concept, we demonstrate a transformation of our approach to a deep neural network architecture working with raw visual inputs and show that pragmatic reasoning also leads to an ME bias in such deep neural network agents. Together, our results open up new possibilities for research on the ME bias in word learning and deep neural networks.

2.6.1 Word learning

If pragmatic reasoning processes as formalized by the RSA framework play a role in infant, child, or adult word learning, they can be captured by our model, at least at the *computational theory* level of explanation (Marr, 1982). Over the past years, there has been growing evidence on the pervasive role of pragmatics in (early) language learning. It has been shown that preverbal infants already understand the communicative nature of language (Martin et al., 2012; Vouloumanos et al., 2012). A recent review by Bohn and Frank (2019) maps out how young children use pragmatic inferences in word learning and how language understanding becomes increasingly more subtle as these inferences grow more complex over time. On the contrary, several studies have found that even at 5 years of age children often fail to perform certain types of pragmatic inferences (Huang & Snedeker, 2009). In the classical ME bias paradigm, alternative utterances can be derived from the context, C. The novel word, w_{n+1}, is contrasted with the label of the familiar object, $w_i \in \{w_1, ..., w_n\}$. It turned out that in cases where pragmatic inference fails, children struggle with generating alternative utterances because they are less clear from the context, rather than computing the inference per se (Barner et al., 2011). Supporting this, several experiments by Frank and Goodman (2014) suggest that children and adults do indeed make RSA-like inferences to infer novel word meanings in context.

Next to pragmatic accounts of ME, constraint and bias accounts form a major strand of theories. They propose that infants have an innate or early emerging lexical bias toward one-to-one mappings between words and meanings. In principle, this bias can be specific to word learning or result from domain-general processes (Markman, 1992). In the dynamic lexicon implementation, pragmatic reasoning affects the agent's inference process. As such it can be seen as a computational model for a pragmatic inference account. In the fixed lexicon implementation, pragmatic reasoning not only affects the inference process but also induces a lexical ME bias. This lexical bias emerges at learning onset and explains why agents with a fixed lexicon but not agents with a dynamic lexicon have a strong ME bias already at the beginning of training. So, with a fixed lexicon implementation lexical constraint accounts and pragmatic inference

accounts can be accommodated by the same general principle of pragmatic reasoning—applied to learning and inference.

Our ablation analysis provides important insights into the role of lexical and inferential pressures toward ME. In line with the probabilistic pragmatic model by Lewis and Frank (2013), we find that both lexical and inferential pressures are sufficient but not necessary for ME. In addition, innate or early emerging lexical biases alone cannot account for the fact that children's ME bias increases with their vocabulary size. It follows that lexical constraint accounts must identify additional factors responsible for this development (e.g., Halberda, 2003). But inferential pressures must also use an evidence accumulation mechanism that reflects increasing certainty about the justification of a one-to-one assumption to model the increasing bias (c.f., Lewis & Frank, 2013).

By using gradient-based learning in pragmatic agents, our model combines aspects of probabilistic pragmatic and associative word learning models (e.g., Kachergis et al., 2012; McMurray et al., 2012; Regier, 2005). Our agents use pragmatic reasoning to infer the meaning intended by the speaker among different alternatives but use gradient-based learning. While associative models typically hard code a competition mechanism to achieve ME (Yurovsky et al., 2013), in our case, such competition arises naturally from the consideration of alternative meanings and utterances. Compared to probabilistic models, gradient-based learning allows our model to separate online inference and long-term learning and thereby to account for interactions between them. At the same time, our model loses the ability to learn from few examples, a disadvantage that is shared by many gradient-based word learning models (e.g., McMurray et al., 2012; Najnin & Banerjee, 2018; Vong & Lake, 2020). From a technical perspective, it opens up the integration of a pragmatic reasoning module with neural network components for processing visual or linguistic input. In sum, our approach might inspire new pragmatic word learning models as it differentiates between long-term learning and online inference and can operate on raw inputs when implemented with a deep neural network architecture.

Aside from dedicated models, general learning theory in the form of the Rescorla-Wagner (R-W) model has also been applied to word learning (e.g., Ramscar et al., 2013; Ramscar et al., 2010). The R-W model can be considered a reinforcement learning model where learning is driven by reward prediction errors. Importantly, associations between referents (cues) and words (outcomes) are updated for both words that are present and words that are not present. However, it is unclear how a learner identifies relevant non-outcomes (Hollis, 2019). Both, our pragmatic model and R-W models of word learning have the effect that children reason about the informativity of a word compared to other words (Ramscar et al., 2013), but the pragmatic reasoning process provides a natural explanation of how relevant alternative utterances can be identified in terms of what a listener thinks a speaker could have said.

This paper set out to explain ME bias behavior as selecting the novel referent o_{n+1} given a novel word w_{n+1}. What is left unaddressed is how the set of relevant referents is to be construed in the first place by the learning agent. If the learning agent knows the word "dog" and also knows that the dog in front of them is called "Fido", a novel word for a novel object (e.g., a

cat) could contrast with the known object at the level of kinds or at the level of individual names. Just like knowledge of how to construe relevant utterance alternatives is crucial for ontogenetically developing pragmatic reasoning abilities (see above), so, too, is it necessary to construe which meaning distinctions are relevant in the given context. Linguistic theory models relevance of meaning distinctions as (possibly implicit) questions under discussions (e.g., Roberts, 2012), essentially using partitions of objects into equivalence classes based on which distinctions matter to the conversation. Recent natural language applications similarly have started to integrate such partition-based approaches to modeling discourse relevance (e.g., Nie et al., 2020). By extending the work presented here to include different levels of partitioning objects into relevance-guided equivalence classes to which novel words might refer, the present approach could be extended to go beyond considering one-to-one relationships between words and objects, thereby capturing a hierarchical organization of word meanings at different levels of granularity. Further challenges for extending the approach in this paper to the full flexibility of natural language lexical meanings include dealing with polysemy, ambiguity and context-dependence, vagueness, and, though arguably very infrequent (since no two expressions are absolutely equivalent in meaning and use), synonymy.

2.6.2 Deep neural networks and outlook

Apart from the pragmatic reasoning module, our deep neural network implementation is very similar to existing deep word learning models (Gulordava et al., 2020; Vong & Lake, 2020). Gulordava et al. (2020) even try a pragmatic inference based approach at test time. Still, these models rely on negative sampling during training to induce competition and achieve an ME bias. As children can map entirely novel words to entirely novel objects in the ME paradigm, the use of negative sampling undermines the explanatory potential of these models. We demonstrate that using pragmatic reasoning at training and test time is sufficient to cause an ME bias such that negative sampling is not necessary.

In future work, it is important to establish if and how our approach can scale to more complex data sets and learning scenarios. Various works apply the RSA framework to deep learning problems. The resulting neural RSA models are used for different supervised learning tasks, such as generating and interpreting referential expressions (Andreas & Klein, 2016; Cohn-Gordon et al., 2018; Monroe et al., 2017; Monroe & Potts, 2015; Zarrieß & Schlangen, 2019), generating and following instructions (Fried, Andreas, et al., 2018; Fried, Hu, et al., 2018), machine translation (Cohn-Gordon & Goodman, 2019), and text generation (Shen et al., 2019). Most of these models are different from ours in that they pretrain a literal listener or a literal speaker on a labeled data set, and then add pragmatic reasoning on top of this "base agent" at test time, whereas our agent applies pragmatic reasoning during training. None of these publications addresses the ME bias challenge. Notably, Zarrieß and Schlangen (2019) consider the problem from the speaker's perspective and use pragmatic reasoning to create better referring expressions for scenes including novel objects. Without negative sampling, regulating the embeddings of novel words and objects becomes crucial for pragmatic reasoning to induce an ME effect (Gulordava

et al., 2020). Our results suggest that bounding the embedding space is one option to achieve this, but it may at the same time limit the model's flexibility. Even though pragmatic reasoning in neural networks is successful in complex domains, and can, in principle, induce an ME bias, the question of whether the proposed ME mechanism generalizes to such applications remains.

Several points should be considered when working with a more complex deep learning setup. First, given the limited number of words in the lexicon, our agent can iterate over all of them in its reasoning process. In a more realistic scenario with an ever-growing vocabulary size, this iteration is computationally too demanding. Many of the approaches mentioned above apply sampling techniques to limit the search space of speaker and listener. Yuan et al. (2018) show that using only the most promising word or object candidates in the pragmatic reasoning process even improves the agent's success. Second, a more challenging training setup can be investigated. How does the agent behave when facing multiple words or several distractor objects? Third, the behavior of the embedding space in relation to architecture and parameter choice must be better understood, such that more general solutions to regularizing the embeddings of novel words and objects can be developed. Our setup is well suited as proof of concept but any research trying to push these ideas to a full word learning model or to a deployable machine learning architecture must factor in these points.

In general, our work is in line with a trend toward building artificial agents with pragmatic reasoning abilities. This trend can be observed in language emergence research (e.g., Choi et al., 2018; Kang et al., 2020; Yuan et al., 2020), amongst others. Language emergence research often employs cooperative games that require a speaker and listener to develop a communication protocol. Our model can also be applied in a multi-agent language emergence setting (Ohmer et al., 2020). Given that the pragmatic listener reasons about the speaker, a speaker agent is already part of the model. In addition, language emergence models typically also use reinforcement learning. Although the focus of this paper is on the ME bias, future work should apply our pragmatic agent models to other language learning and language emergence phenomena.

2.7 Conclusion

We have developed a model of learning in pragmatic agents, which can be parameterized by lexicon entries or neural network weights. We show that pragmatic inference combined with learning can account for the ME bias phenomenon and (at least qualitatively) its developmental trajectory, also under the influence of modulating factors. The neural network model demonstrates how pragmatic reasoning in semantic learning can implement an ME bias mechanism in deep word learning models. In future work, we would like to investigate the model behavior for many-to-one and one-to-many associations between words and objects, include a mechanism for determining relevant meaning distinctions, and find more general solutions to structuring the embedding space in the neural network model.

Data availability

Materials and code are openly available at the Open Science Framework: https://osf.io/2wz9x/?view_only=154564daa91c4ce9a6398ea641ae598d

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—GRK 2340. We would like to thank the two anonymous reviewers and the editor for their extremely helpful comments and suggestions.

Open Access funding enabled and organized by Projekt DEAL.

Conflicts of interest

The authors have no conflicts to disclose.

A Appendix

A.1 Hyperparameter search – Explicit lexicon

To find good hyperparameters for the models with an explicit lexicon, we conducted a grid search for the fixed and dynamic lexicon simulations, respectively. We ran the search for an intermediate exposure interval of $k = 10$, and varied the following hyperparameters:

- data set size: $\{100, 1000\}$
- batch size: $\{16, 32\}$
- learning rate: $\{0.001, 0.01, 0.1\}$
- lexicon initialization: $\{0.0001, 0.001, 0.01, 0.1\}$

So, for each lexicon, we tested $2 \times 2 \times 3 \times 4 = 48$ different combinations. Out of data set size, batch size, and learning rate, we used the hyperparameters that worked best across both lexicon types, to allow for a more direct comparison. The initial lexicon size is very different between the two lexicon types, which is why we used the lexicon initialization only for within implementation comparison. Choosing some parameters based on best performance across implementations is not problematic, since the final parameter combinations achieve (with negligible differences) the same performance as the best parameters for each implementation. The full list of results and details of the selection procedure can be found in our OSF project.

A.2 ME index formulas

Words and objects are indexed $1 \leq i \leq N$, respectively, with words and objects of the same index belonging together. New word-object pairs are added to the training set in order of their indices. In the following formulas, the ME index is evaluated with respect to the word and object with index j, so w_j and o_j. For the fixed lexicon, until the last word-object pair is added, there are always several novel objects the agent can select:

$$I_{ME}(B_{PL}, w_j) = \frac{\sum_{i=j}^{N} PL(o_i \mid w_j, B_{PL}) - \frac{N-(j-1)}{N}}{\frac{j-1}{N}},$$

In case of a dynamic lexicon, there is only one novel object the agent can select leading to the following simplification:

$$I_{ME}(B_{PL}, w_j) = \frac{PL(o_j \mid w_j, B_{PL}) - \frac{1}{M}}{\frac{M-1}{M}},$$

where $M <= N$ is the current lexicon size (after being extended for w_j and o_j).

A.3 Convolutional neural network for feature extraction

The vision module of the deep neural network implementation has the following architecture. First, there are two convolutional layers, each with 32 filters of size 3×3. Then a fully connected layer with 64 units follows and finally the output layer with 10 units. Hidden units use a relu activation function and output units a softmax activation function. Each convolutional layer is followed by a max-pooling layer with pooling size 2, and every layer apart from the output layer uses dropout with a probability of 0.3. Weights were initialized randomly, and training proceeded until an early stopping criterion with patience 3 was reached on the validation loss. We used a batch size of 32 and the Adam optimizer with a learning rate of 0.001. The model achieved the following training, validation, and test accuracies: 97.97%, 98.87%, 99.08%.

A.4 Example lexica of the neural network agent

The matrices in Fig. 2.10 show the literal agent's lexicon after training, for three different random seeds. Learned associations are given by high lexicon entries, corresponding to strong similarities between word and object representations in the embedding space. It can be seen that the agent always learns the one-to-one correspondences between familiar objects (0–8) and familiar words (nine randomly selected words). The bottom row of the lexica shows how strongly the novel object, digit 9, is associated with each word. It turns out that associations with novel words are not necessarily stronger than with familiar words. However, novel words tend to lie closer to novel object 9 than to familiar objects, that is, lexicon entries for novel words (e.g., $\{0, 1, 2, 3, 7, 10, 12, 13, 15, 18, 19\}$ in the first example) are relatively high in the bottom row.

The structure of the joint embedding space illustrates why also the literal agent displays an ME bias in our simulations. After training, embeddings of novel words happen to be more similar to embeddings of the novel object than to embeddings of the familiar objects.

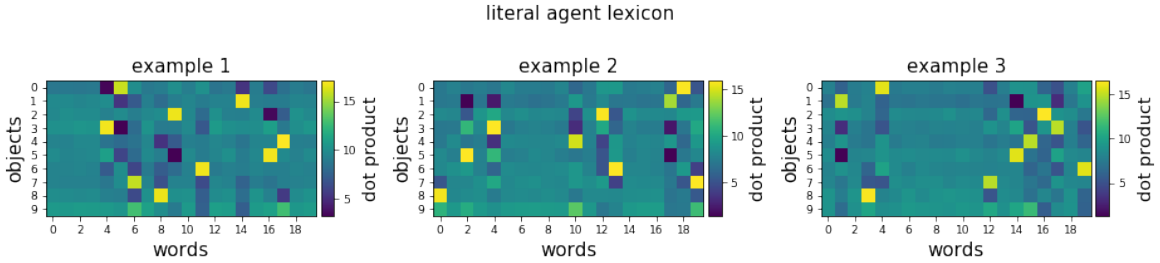


Figure 2.10: Lexica of the literal agent from three randomly selected runs. Lexicon entry ij is calculated as the average dot product between the embeddings of training examples showing object i and the embedding of word j.

The matrices in Fig. 2.11 visualize the pragmatic agent's lexicon after training, for different negative sampling strategies, and for three different random seeds. Again, the learned one-to-one mappings between familiar words and familiar objects are clearly visible. The lexica without negative sampling (first column) and with negative sampling of words (second column) are similar to the lexica of the literal agent. With negative sampling of objects (third column), however, the agent learns that the novel object is not associated with any of the familiar words. Driving the embedding of the novel object away from the embeddings of familiar words, simultaneously increases the distance to the embeddings of novel words, resulting in low values throughout the bottom row. This side effect can be mitigated by negative sampling of both words and objects (fourth column). Embeddings of novel words/objects move away from those of familiar objects/words. Within the bounded embeddings space, embeddings of novel words and novel objects stay close together in the process.

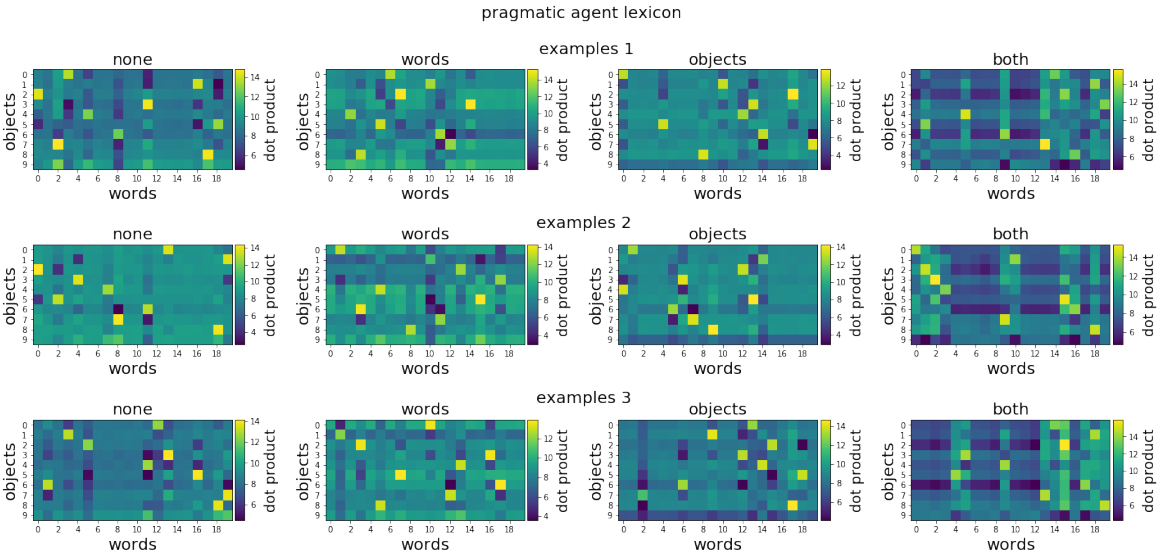


Figure 2.11: Randomly selected lexica of the pragmatic agent ($\alpha = 5$) trained with different negative sampling strategies. Each row contains one example for each sampling strategy. Lexicon entry ij is calculated as the average dot product between the embeddings of training examples showing object i and the embedding of word j.

3 Case study 2: Referring to objects at different levels of specificity

This chapter presents case study 2. The chapter starts with a lay summary, which is followed by the content of the publication:

Ohmer, X., Duda, M., and Bruni, E. (2022). Emergence of hierarchical reference systems in multi-agent communication. *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, pp. 5689–5706. https://aclanthology.org/2022.coling-1.501/

3.1 Lay summary

We can refer to the same entity at different levels of specificity. For example, a dog can be called "Fido" (his name), "dalmatian", "dog", or "animal". We can also be more specific by adding descriptions as in "black-and-white dog". Multiple levels of specificity can be said to define a *hierarchy* from the more concrete expressions (Fido) at the bottom to more abstract expressions at the top (animal). The level of specificity we choose when referring to an object largely depends on the context. For example, "dog" is unambiguous if we see a dog and a cat, but not if we see two dogs. In this project, we build a computational model of communicating agents and study whether they, too, develop hierarchical reference systems, allowing them to be more or less specific depending on the context.

To do so, we design a language emergence game. Language emergence games are used to simulate the origins and evolution of language. The agents in the game have different information about the world and have to develop a language, from scratch, to exchange that information. We call our game *hierarchical reference game*. The game is played by a sender and a receiver agent. The sender sees an object and has information about which object properties are relevant in the given context. Together, the object and the relevant properties define a target concept, which the sender has to communicate. The concept is more concrete if many properties are relevant (e.g. color, size, and shape) and more abstract if only a few properties are relevant (e.g. color). The sender sends a message to the receiver. Based on that message, the receiver must select an object that matches the target concept among many mismatching distractor objects. For example, if the sender sees a *big red triangle*, but only color and shape are relevant, a red triangle of any size is considered to match the target concept.

Our results show that the agents learn to play the game over time, which means that they can communicate hierarchically structured concepts. We further test whether the agents consistently use the same message (e.g. "aba") for the same concept (e.g. *red*). It turns out that they use messages very consistently for concrete concepts but there is some variation for abstract concepts.

Unlike concrete concepts, abstract concepts (e.g. *red*) are represented by many different examples (e.g. *big red triangle*, *small red circle*, ...). At times, the agents communicate irrelevant information which reduces consistency. Comparing different input data sets reveals that agents are more consistent for abstract concepts if objects are more diverse, so if they come in many different shapes, colors, and so on. In that case, using a single term for an abstract concept like *red* is more efficient because a greater number of objects can be described by this term (red triangles, red circles, red squares, . . .). Thus, beyond developing dedicated messages for concrete objects, the agents also develop dedicated messages for abstract concepts, if these messages allow them to describe many different objects.

In principle, the agents could describe each concept by using a distinct and arbitrary message. However, the agents can also communicate about new concepts, which hints at a more systematic strategy. For example, they can communicate about novel combinations of properties: having learned to communicate about blue triangles and red circles they can also communicate about red triangles or blue circles without having seen them before.

The concepts in our setup are defined compositionally. The same basic properties are recombined to generate various objects. The emergence of compositional language is of great research interest, as it is thought to lie at the heart of our ability to produce and understand a possibly infinite number of sentences. We follow up with an analysis of the agents' language to determine how they achieve abstraction and whether the use of compositional messages plays a role. It turns out that the compositional input structure is reflected in the agents' language. Although the relationship is not perfect, the agents seem to use specific symbols to encode specific properties and recombine these symbols to communicate combinations of these properties.

In natural language, abstraction can be achieved by leaving out irrelevant information ("triangle") or by explicitly stating that some aspect is irrelevant ("triangle of any color"). The agents in our simulations use both of these strategies. They use shorter messages for more abstract concepts, indicating that they leave out some of the irrelevant information. But there are also a few symbols that they systematically use across abstract concepts. We interpret these symbols as *abstraction operators*, similar to "any" as in "any color". In short, hierarchical reference is achieved through compositional messages, where symbols encode specific properties or the irrelevance of properties.

Our work shows that computational models can develop hierarchical reference systems when taking into account information about the context. Dedicated expressions for abstract concepts emerge especially if objects are very diverse. Our language analysis reveals that systematicity is, in part, achieved through compositionality and the use of abstraction operators.

3.2 Abstract

In natural language, referencing objects at different levels of specificity is a fundamental pragmatic mechanism for efficient communication in context. We develop a novel communication game, the *hierarchical reference game*, to study the emergence of such reference systems in artificial agents. We consider a simplified world, in which concepts are abstractions over a set of primitive attributes (e.g., color, style, shape). Depending on how many attributes are combined, concepts are more general ("circle") or more specific ("red dotted circle"). Based on the context, the agents have to communicate at different levels of this hierarchy. Our results show that the agents learn to play the game successfully and can even generalize to novel concepts. To achieve abstraction, they use implicit (omitting irrelevant information) and explicit (indicating that attributes are irrelevant) strategies. In addition, the compositional structure underlying the concept hierarchy is reflected in the emergent protocols, indicating that the need to develop hierarchical reference systems supports the emergence of compositionality.

3.3 Introduction

Humans excel at using language to convey information efficiently in context. A speaker does not have to communicate every detail. Rather, a listener can infer the intended meaning of an utterance by assuming that sufficient information was provided. This idea was first explicitly formulated by Grice (1975) in his conversational maxims, in particular the *Maxim of Quantity*: "1. Make your contribution as informative as is required (for the current purposes of the exchange). 2. Do not make your contribution more informative than is required." An illustration of this mechanism can be given in the form of a simple referential context. In a scene with a red circle and a green triangle, "circle" is enough information to identify the referent, whereas more complex scenes may require the speaker to name both object attributes—shape and color—to allow for an unambiguous interpretation. The Maxim of Quantity requires a hierarchical reference system, that allows the selection of the most appropriate level of specificity for a given context.

In this paper, we follow the proposal by Higgins et al. (2018) and define concepts as compositional abstractions over a set of primitive attributes (e.g., color, style, shape), see Figure 3.1.a. The concepts are maximally specific at the leaf nodes, where all attribute values are determined. Moving from the subordinate levels up to the superordinate levels, the number of concept-defining attribute values decreases. Thus, each parent concept is an abstraction (i.e. a subset) over its children and over the original set of attribute values. Given this definition of a concept hierarchy, we use a language emergence paradigm with artificial agents to study whether a corresponding reference system can emerge given a structured perception of the world.

In most language emergence simulations, a sender and a receiver agent are trained on a reference game (e.g., Dagan et al., 2021; Havrylov & Titov, 2017; Lazaridou et al., 2018; Rodríguez Luna

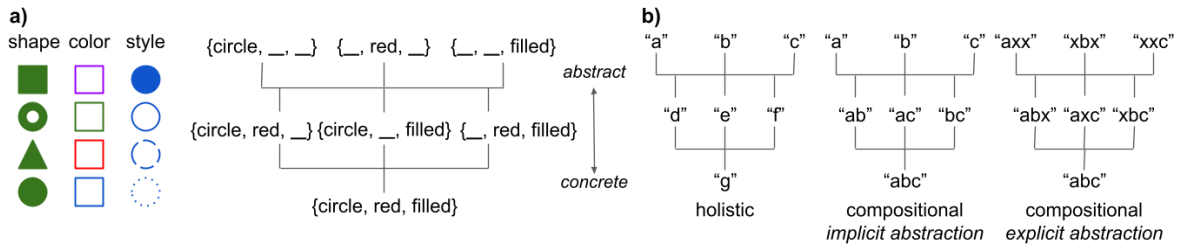


Figure 3.1: a) Example of a concept hierarchy. Shown are all attribute values and the concept hierarchy constructed from the concept "red filled circle". **b)** Example languages for the concept hierarchy in part a). Possible abstraction strategies include holistic and compositional languages. In compositional languages, abstraction can further be indicated implicitly or explicitly.

et al., 2020), based on the signaling game originally developed by Lewis (1969). The sender sees a target object and sends a message to the receiver. Using that message, the receiver tries to identify the target among a set of distractor objects. Crucially, in the current form, reference games completely ignore that different contexts may require referential expressions at different levels of abstraction. Having no access to the distractors, the sender cannot choose relevant object attributes in a context-dependent way. Moreover, random sampling of the distractors typically encourages the sender to communicate all object attributes. Therefore, the standard reference game cannot account for the emergence of hierarchical concepts in communication.

We develop a *hierarchical reference game* to address this shortcoming. Instead of an object, the sender receives a *concept* as input. The concept is defined by an attribute vector (object) and a relevance vector (context). The relevance vector indicates for each attribute whether it is relevant in the current context or not. Based on the sender's message, the receiver must identify an object that instantiates the target concept among a set of distractors. The input concepts have a compositional and hierarchical structure. While the game is designed to encourage communication at different levels of abstraction, it does not regulate how this abstraction is realized; in particular, there is no explicit pressure on the emergent language to reflect the compositional input structure.

First, we evaluate if the agents can successfully play the game, i.e. communicate specific contextually relevant object attributes. Second, to measure whether the agents' strategies are systematic, we test whether they can generalize to novel concepts, and also whether they consistently use the same expressions for the same concepts at all levels of abstraction. Third, we investigate the emerging protocols to study the mechanisms by which systematic abstraction is achieved, see Figure 3.1.b. In natural language, there is *holistic* abstraction as in "Dalmatian" ⊆ "dog" ⊆ "animal"; but also *compositional* abstraction as in "filled red circle" ⊆ "red circle" ⊆ "circle". Abstraction can further be *implicit*, by omitting irrelevant attributes, and *explicit*, by indicating that certain attributes are irrelevant (as in saying "a circle of any color"). We evaluate which, if any, of these abstraction strategies are used by the agents.

Our work makes several contributions. We develop the hierarchical reference game and show that it can be used to model the emergence of referential expressions at different levels of abstraction. We also provide novel metrics to examine how the agents achieve abstraction.

Working with different data sets, i.e. different concept hierarchies, allows us to disentangle data set specific and general effects. Not least, our results suggest that communication about concept hierarchies supports the emergence of compositionality.

3.4 Related work

Referring expression generation (REG). There has been a long history of research on understanding how people generate referring expressions, dating back to Winograd (1972). The most influential work on REG in both the eighties (Appelt, 1985; Appelt & Kronfeld, 1987) and nineties (Dale & Reiter, 1995; Reiter & Dale, 1992) integrated the Gricean maxims into their systems. The latter developed the *iterative algorithm*, which was used and extended to model various aspects of REG (Krahmer & van Deemter, 2012). Like Dale and Reiter (1995), we define objects as sets of attribute-value pairs and consider the subset of referring expressions whose single purpose is to identify an object. However, our main focus is not to generate human-like referring expressions but rather to build artificial agents capable of hierarchical reference. Hence, we ignore many effects that play a role in human REG, for example basic level categories (Rosch & Mervis, 1975; Rosch et al., 1976).

By now, several large-scale data sets of referring expressions for complex real-world images have been collected and are used to train deep neural networks (DNNs) (e.g., Kazemzadeh et al., 2014; Luo et al., 2020; Luo & Shakhnarovich, 2017; Mao et al., 2016; Yu et al., 2018; Yu et al., 2016). The data sets are collected in a reference game setup: one participant has to refer to a target entity in a given image, and the other participant has to identify the target. Models are often trained on both components, expression generation and comprehension (e.g. Luo & Shakhnarovich, 2017; Mao et al., 2016). Several REG models try to integrate deep learning with computational accounts of pragmatic reasoning (e.g., Andreas & Klein, 2016; Le et al., 2022; Monroe & Potts, 2015), such as the Rational Speech Act framework (Frank & Goodman, 2012). Our model also implements expression generation (sender) and comprehension (receiver) using DNNs but hard codes pragmatic inferences in the relevance vector. Most importantly, the agents are not trained on a labeled data set but develop their own referring expressions in a language emergence game.

Emergent multi-agent communication. Language emergence simulations are popular in evolutionary linguistics as well as AI research. In evolutionary linguistics, they are used to study the origins and evolution of human and animal communication (e.g., Cangelosi & Parisi, 2002; Kirby, 2002b; Wagner et al., 2003). In AI research, they are used with the aim of building artificial agents capable of flexible and goal-directed language use, which arguably relies on grounding language in interaction (e.g., Lazaridou & Baroni, 2020; Steels, 2001, 2003).

Starting with Foerster et al. (2016) and Lazaridou et al. (2017), there has been an increasing interest in language emergence simulations with DNN agents (for a review, see Lazaridou &

Baroni, 2020). These approaches stand in contrast to the currently dominant DNN models in NLP, which learn passively by being exposed to large amounts of text (Bisk et al., 2020). As discussed above, in many implementations, hierarchical reference systems cannot emerge because the sender does not have access to information about the context. Even in the rare cases where it does (e.g., Dessi et al., 2021; Lazaridou et al., 2017), the emergence of hierarchical reference has not yet been investigated.

3.5 Setup

3.5.1 Concept representation

We use symbolic, disentangled input representations. A concept is composed of an object vector and a relevance vector. Objects have n attributes and each attribute can take on k values.[1] The relevance vector $r \in \{0,1\}^n$ indicates which attributes are relevant (1) and which ones are irrelevant (0). E.g., if the sender's input is $(4,3,1)(1,0,0)$, the concept in question is $(4,\_,\_) := \{(4,x,y) \mid x,y \in \mathbb{N}, 1 \leq x,y \leq k\}$. Object $(4,3,1)$ could instantiate the attributes *shape*, *color*, and *style* with specific values such as *circle*, *red*, and *filled* (see Figure 3.1.a). Relevance vector $(1,0,0)$ would then indicate that only the first attribute value, *circle*, is relevant and must be communicated.

3.5.2 Hierarchical reference game

Like the classical reference game, the hierarchical reference game is played by a sender, S, and a receiver, R. However, rather than communicating the input object as it is, the sender must abstract a concept from this object based on the relevance vector. One round of the hierarchical reference game proceeds as follows (see Figure 3.2):

1. An object, o, and a relevance vector, r, are sampled randomly and passed to S.
2. Based on this input, S generates a message, m. The message is a concatenation of symbols from vocabulary V, $s_i \in V$, and has maximal length L, such that $m = (s_i)_{i \leq L}$.
3. R receives the message m, as well as a set of objects containing one target, t, and several distractors. t has the same attribute values as the input object o for relevant attributes (as defined by r), while the values of irrelevant attributes are sampled randomly. The distractors are constructed by sampling object instances of concepts that would arise from o in combination with other relevance vectors than r.
4. Based on m, R selects one object among target and distractors.

By our choice of distractors, we simulate an environment in which the relevance vector matches pragmatic needs: the speaker tries to be as specific as necessary in a given context. To further

[1] We present objects as n-hot encodings to the agents, such that each object $o \in \{0,1\}^{nk}$.

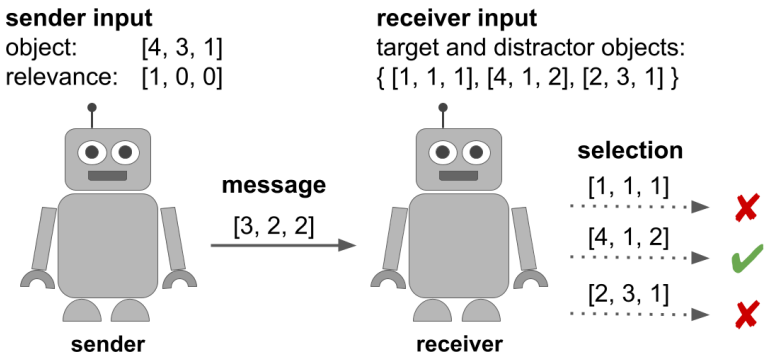


Figure 3.2: Schematic illustration of the hierarchical reference game.

discourage communication of irrelevant attributes, we choose distractors that are more abstract than the target concept but still similar, by replacing exactly one 1 (*relevant*) in the relevance vector with a 0 (*irrelevant*). Additional experiments, where we sample distractors with equal probability from all levels of the concept hierarchy, can be found in Appendix B.1.[2]

3.5.3 Architecture

Both agents are implemented as single-layer GRUs. The sender input is processed by two dense layers, one receiving the object vector and one the relevance vector, followed by a dense layer mapping a concatenation of these two representations to the sender's initial hidden state. The message is produced incrementally. At each time step, the sender generates a probability distribution over the vocabulary which is used to sample a symbol from the Gumbel-softmax distribution (Jang et al., 2017). The GS distribution is a continuous distribution that approximates categorical samples, and whose parameter gradients can be easily computed via a reparameterization trick. The receiver processes the incoming message. An additional dense layer maps target and distractor objects onto embeddings. These embeddings have the same dimensionality as the receiver's hidden state. The receiver's selection probabilities are determined by applying a softmax function to the dot products between object embeddings and hidden state.

3.6 Experiments

Our implementation is based on PyTorch, and uses the EGG toolkit (Kharitonov et al., 2019).[3] Our code and results are available at https://github.com/XeniaOhmer/hierarchical_reference_game.

[2] In that case, the agents still learn to play the game successfully and to form abstract concepts but they have a stronger tendency to convey also irrelevant information.
[3] https://github.com/facebookresearch/EGG

3.6.1 Data sets

In order to investigate how the number of attributes and the number of values per attribute influence the formation of abstract concepts, we use a set of different data sets, $D(n, k) \coloneqq \{(m_1, \ldots, m_n) \mid m_i \in \mathbb{N}_k\}$ with $\mathbb{N}_k = \{1, ..., k\}$ (see Table 3.1).

Table 3.1: Input data sets with n attributes and k values. Data sets are labeled as $D(n, k)$.

	$k = 4$	$k = 8$	$k = 16$
$n = 3$	$D(3, 4)$	$D(3, 8)$	$D(3, 16)$
$n = 4$	$D(4, 4)$	$D(4, 8)$	
$n = 5$	$D(5, 4)$		

We sample relevance vectors with equal probability from each level of the concept hierarchy.[4] We repeat that procedure until there are 10 samples for each input object and each number of relevant attributes in the data set. In addition, we create 10 distractors per sample. We reserve 20% of the data for zero-shot testing (see Section 3.6.3), and split the remaining data randomly into training and validation sets at a ratio of 0.75/0.25.

3.6.2 Hyperparameter selection and training

In our simulations there is always exactly one target object for the receiver, i.e. only that object—and none of the distractors—is an instance of the target concept. The agents minimize the cross-entropy loss between target and selection. During training, a message is given by the GS distributions across symbols, whereas during testing the argmax values are used. Hence, it is possible to jointly update the weights of sender and receiver by backpropagating through the approximated "discrete" messages.

We conducted a hyperparameter search to identify model and training parameters leading to high performance on the validation set for the range of different data sets we use (for details see Appendix B.2). Agents have an embedding layer with 128 units and a hidden layer with 256 units. The discrete messages are approximated using GS with an initial temperature of 1.5, decaying exponentially at a rate of 0.99. We train for 300 epochs using Adam optimizer with batch size 32 and learning rate 0.0005. For all data sets, we use the number of attributes as maximal message length L. The minimal vocabulary sizes in Table 3.2 allow the sender to generate a distinct message for each input concept. They correspond to the number of attribute values plus one additional symbol to indicate irrelevance. The agents have an additional end-of-sequence symbol to terminate the messages before L is reached. We run our experiments with a factor $f = 3$ of the minimal vocabulary size. Additional experiments with other values for f be found in Appendix B.1.[5] We conduct 5 runs per data set.

[4] If relevance vectors are sampled uniformly from the set of all relevance vectors, the amount of 0 and 1 entries follows a binomial distribution. Sampling relevance vectors with equal probability from each level of the hierarchy ensures that all abstraction levels occur equally often.

[5] Smaller factors make the task more difficult and performance decreases, while larger factors do not yield any

Table 3.2: Minimal vocab size for each data set.

	$k = 4$	$k = 8$	$k = 16$
$n = 3$	5	9	17
$n = 4$	5	9	
$n = 5$	5		

3.6.3 Evaluation

We are interested in different aspects of the emergent language, and use existing as well as novel metrics to evaluate these.

Zero-shot evaluation. We generate two different zero-shot test sets. The first test set is used to evaluate whether the agents can generalize to novel *objects*. It contains combinations of attribute values that do not occur in the training and validation data, and is reserved for testing after the data generation process. The second test set is used to evaluate whether the agents can generalize to novel *abstractions*. We withhold abstractions from one value per attribute from the training data. The agents are trained from scratch on the remaining data and evaluated on the held-out data.

Message consistency and effectiveness. To measure whether agents consistently use the same messages for the same concepts we employ information-theoretic metrics. Let C be the set of concepts, and M be the set of messages. The conditional entropy of messages given concepts,

$$\mathcal{H}(M \mid C) = -\sum_{c \in C,\, m \in M} p(c, m) \log \frac{p(c, m)}{p(c)},$$

measures how much uncertainty about the messages remains after knowing the concepts. Low values indicate that the agents consistently use the same messages for the same concepts, i.e. the language does not contain many synonymous expressions. $\mathcal{H}(C \mid M)$, in turn, measures how much uncertainty about the concepts remains after knowing the messages and should therefore negatively correlate with the agents' performance. Low values indicate that agents effectively use messages that uniquely identify the target concept, i.e. the language does not contain many polysemous expressions. On this basis, we define a *consistency* an *effectiveness* score, using the marginal entropies $\mathcal{H}(C)$ and $\mathcal{H}(M)$ for normalization:

$$\text{consistency}(C, M) = 1 - \frac{\mathcal{H}(M \mid C)}{\mathcal{H}(M)}$$
$$\text{effectiveness}(C, M) = 1 - \frac{\mathcal{H}(C \mid M)}{\mathcal{H}(C)}.$$

---

further improvements.

Finally, the normalized mutual information,

$$\mathcal{NI}(C,M) = \frac{\mathcal{I}(C,M)}{0.5 \cdot \big(\mathcal{H}(C) + \mathcal{H}(M)\big)} = \frac{\mathcal{H}(M) - \mathcal{H}(M \mid C)}{0.5 \cdot \big(\mathcal{H}(C) + \mathcal{H}(M)\big)},$$

is a symmetric measure that combines the two conditional entropies into one score.

Symbol redundancy. We develop this metric to approximate whether agents repeat information about attribute values in their messages. It assumes that each attribute value, a_v (e.g a=*color*, a_v=*red*), is encoded by a specific symbol and counts how often that symbol is repeated given that a_v is being encoded. The preferred symbol for each attribute value is defined $s_v^a := \arg\max_s \mathcal{I}(a_v, s)$, where we code for each message whether s occurs at least once (the position of s is irrelevant). Symbol redundancy is defined as the average number of occurrences of s_v^a per message given that a_v is part of the target concept.

Topographic similarity. Topographic similarity (*topsim*) measures to what degree similar inputs are described by similar messages and is frequently used as a measure of compositionality. The metric calculates the pairwise distances between the inputs, as well as the pairwise distances between the corresponding messages, and then correlates the two distance vectors (Brighton & Kirby, 2006). In the hierarchical reference game, we need to calculate the topographic similarity between messages and concepts. We use an n-hot encoding of the concepts (n being the number of attributes) and treat abstraction from each attribute as an additional attribute value. If an attribute is relevant, that value is zero (no abstraction), if an attribute is irrelevant this value is one (abstraction) and overwrites the original attribute value. Assuming that each attribute can take on $k = 4$ different values, the input encoding for the example in Figure 3.2 becomes:

$$\begin{aligned} \text{sender input} &: [\,4\,3\,1\,]\,[\,1\;0\;0\,] \text{ (object + relevance)} \\ \text{encoding} &: [\,0\,0\,0\,1\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,1\,] \end{aligned}$$

Analogously to Lazaridou et al. (2018), we calculate the pairwise distances of the inputs using the cosine distance, and the pairwise distances between the messages using the edit distance. The topsim score is calculated as the Spearman correlation between the two resulting distance vectors.

Disentanglement. Positional disentanglement (*posdis*) and bag-of-symbols disentanglement (*bosdis*) are used to measure different types of compositionality (Chaabouni et al., 2020a). For both metrics, concepts are encoded as for the topsim score. Posdis measures whether symbols in specific positions encode the values of a specific attribute, i.e. whether the compositional

structure is order-dependent. Let s_j be the j-th symbol of a message, then posdis is defined as

$$\text{posdis} = \frac{1}{L} \sum_{j=1}^{L} \frac{\mathcal{I}(s_j, a_1^j) - \mathcal{I}(s_j, a_2^j)}{\mathcal{H}(s)},$$

where L is the maximal message length, and a_1^j and a_2^j are the attributes that achieve the highest and second-highest mutual information with s_j ($a_1^j = \arg\max_a \mathcal{I}(s_j, a)$; $a_2^j = \arg\max_{a \neq a_1^j} \mathcal{I}(s_j, a)$). Bosdis measures whether symbols refer to specific attribute values independent of their position. In that case, the language is permutation-invariant and only symbol counts matter. Let n_j be a counter of the j-th symbol in a message, then bosdis is defined as

$$\text{bosdis} = \frac{1}{|V|} \sum_{j=1}^{|V|} \frac{\mathcal{I}(n_j, a_1^j) - \mathcal{I}(n_j, a_2^j)}{\mathcal{H}(n_j)},$$

where V is the vocabulary size, and a_1^j and a_2^j achieve the highest and the second-highest mutual information with n_j.

3.7 Results

In this section, quantitative and aggregated results will be presented. Random examples of concepts and messages, together with a qualitative analysis can be found in Appendix B.4. The first part of Appendix B.4 shows example mappings between abstract concepts and messages and the second part highlights different abstraction strategies.

3.7.1 Performance and generalization

Figure 3.3 shows the mean accuracies on training, validation, and zero-shot test sets for all data sets. Training accuracies (top left) and validation accuracies (top right) are very high for each data set, considering that chance performance is $< 10\%$. Thus, the agents learn to refer to objects at different levels of abstraction, and their strategies do not overfit the training data.

Accuracies for novel combinations of attribute values (bottom left) are consistently higher than accuracies for novel combinations of abstraction and attribute value (bottom right), except for $D(3, 8)$. Accordingly, generalizing to novel abstractions of attribute values is harder than generalizing to novel objects. Both types of generalization tend to improve with the number of attributes as well as the number of values, which may be due to an increase in input space size (Chaabouni et al., 2020b). Similar to training and validation accuracies, generalization to novel objects reaches almost perfect accuracies, if there are many attributes. While generalization to novel abstractions is more difficult, accuracies strongly exceed chance performance and are still very high for $D(3, 16)$ with 84.76% and $D(4, 8)$ with 94.38%. A large number of attribute values seems to be more important for generalizing to novel abstractions than for generalizing

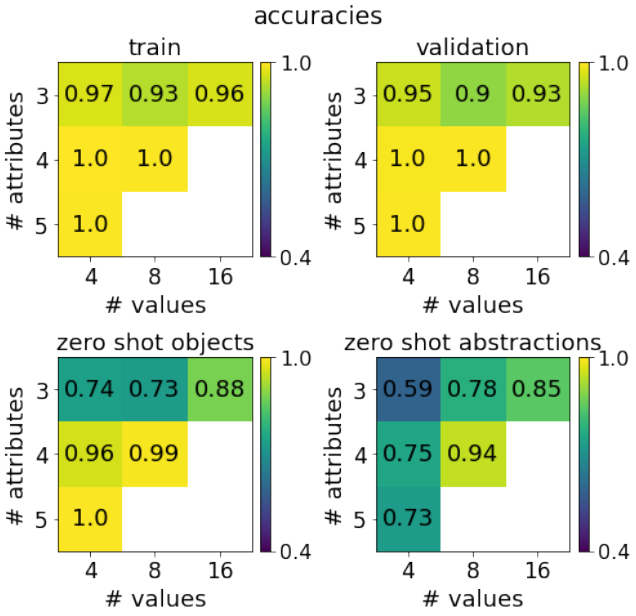


Figure 3.3: Mean accuracies across five runs for each of the training data sets. Shown are accuracies on the training set, the validation set, and the two zero-shot test sets.

to novel objects, possibly because it is more useful to learn systematic abstraction if there are many attribute values. A strategy that abstracts from a certain attribute can be applied to more concepts if that attribute has many values (i.e. has more children on the concept hierarchy). Overall, the agents develop hierarchical reference systems and, with enough attributes and values, these systems generalize well to novel objects and novel abstractions.

3.7.2 Mapping between concepts and messages

We determine the structure of correspondences between messages and concepts. Figure 3.4 shows the mean effectiveness and consistency scores. The effectiveness score measures how much information about the target concept is contained in the message. It follows that the agents can only achieve high performance if the language is effective. The results show this interrelation, in that the pattern of effectiveness scores matches the pattern of training and validation accuracies across the different data sets. The consistency score, on the other hand, measures whether a concept is consistently mapped onto the same message, and high consistency is not necessary to achieve high performance. The score is higher for a larger number of attribute values, supporting the finding above that many values per attribute increase the pressure to develop systematic abstraction strategies.

For each data set, the normalized mutual information lies between the effectiveness and the consistency score. It is generally high ($0.902 \leq \mathcal{NMI} \leq 0.945$), indicating that messages and concepts are strongly predictive of each other. A one-to-one correspondence between words and messages is not enforced by the setup because the message space is far larger than the concept space. The high entropy scores mean that a systematic mapping between concepts and messages, and therefore also systematic abstraction emerge nonetheless.

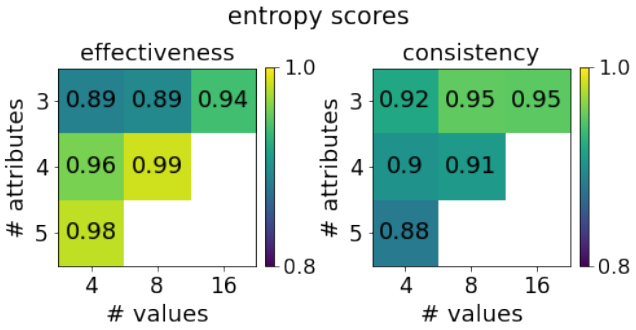


Figure 3.4: Mean effectiveness and consistency scores. We display the mean scores across five runs for each of the training data sets.

To analyze where the languages deviate from a one-to-one correspondence between concepts and messages, we consider the relation between entropy scores and level of abstraction (see Figure 3.5). The mutual information between messages and concepts is higher for more concrete concepts. This effect is largely driven by an increase in consistency, while effectiveness is relatively constant across all levels of abstraction. Thus, deviations from the one-to-one correspondence between concepts and messages occur mostly for abstract concepts. These deviations arise because different messages map to the same concept, not vice versa. In other words, the languages contain synonymy but no polysemy.

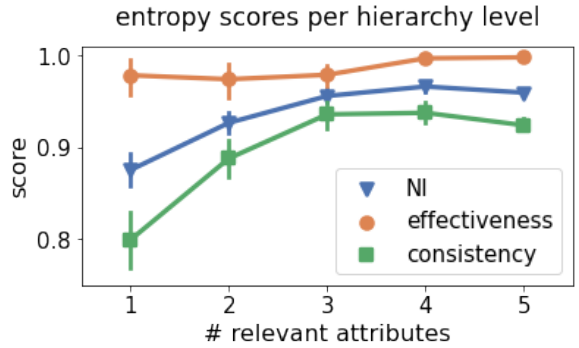


Figure 3.5: Mean entropy scores across all data sets for different numbers of relevant attributes: from left to right concepts become more concrete. Error bars indicate bootstrapped 95% confidence intervals.

3.7.3 Linguistic abstraction strategies

Here, we look more closely at the types of internal message structures used to create a hierarchical reference system.

Implicit versus explicit abstraction. In natural language, there are implicit and explicit ways of communicating that attributes are irrelevant. A commonplace implicit strategy is to simply omit information about irrelevant attributes, e.g. one might say "car" rather than "red car" if sufficient. Since the maximal message length corresponds to the maximal number of relevant attributes, the agents could achieve a similar effect by using shorter messages for more abstract

concepts or by using messages that contain more redundancies. Figure 3.6 shows message length and symbol redundancy averaged across data sets for each level of abstraction. The agents indeed use implicit abstraction strategies and this is captured by both metrics. The message length decreases for more abstract concepts while symbol redundancy increases. For abstract concepts, symbols that encode irrelevant attributes are either omitted or replaced by repetitions of symbols encoding relevant information.

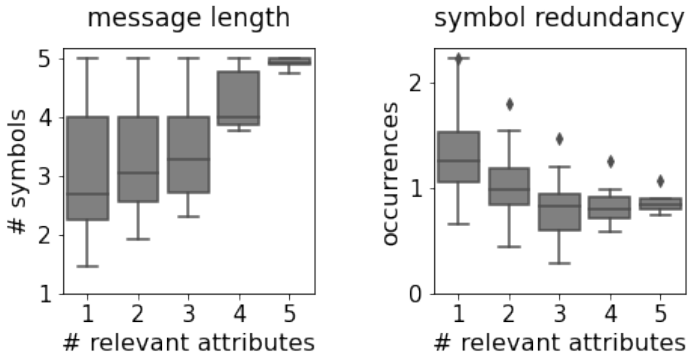


Figure 3.6: Average message length and symbol redundancy across data sets for different numbers of relevant attributes: from left to right concepts become more concrete.

Explicit abstraction would mean that the agents dedicate symbols to express that information is irrelevant. Such *abstraction operators* should co-occur frequently with abstract concepts. We calculate the average number of symbol occurrences per message for each level of abstraction. We rank the symbols by their occurrences for the most abstract concepts to identify candidate symbols. Figure 3.7 shows the results for the top ten candidates, averaging multiple runs for each ranked symbol. The ranking is visible in the left-most columns, where the number of occurrences per message decreases monotonously from the highest to the lowest rank. Strikingly, for all data sets except $D(3, 4)$ only 1–3 symbols occur very frequently together with very abstract concepts and the occurrence values decrease rapidly when going further down the ranks. Importantly, these symbols do not occur frequently at every level of the concept hierarchy. Rather, their usage decreases continuously as concepts become more concrete, as indicated by the gradient from left to right in the top rows. Thus, it seems likely that the agents use one or a few symbols to explicitly communicate information about the irrelevance of one or more attributes. The formation of abstraction operators is surprising since the message space is large enough to encode irrelevance differently, for example by combining symbols or using different symbols for different attributes.

Compositional versus holistic abstraction. The hierarchical reference game requires the agents to repeatedly communicate the same attribute values but for different concepts—different because of the values of other attributes (traversing the hierarchy horizontally) or because of the level of abstraction (traversing the hierarchy vertically). Although the agents could develop holistic protocols, this repeated reference across contexts might encourage them to develop "reusable" mappings from attribute values to symbols, i.e. compositional expressions.

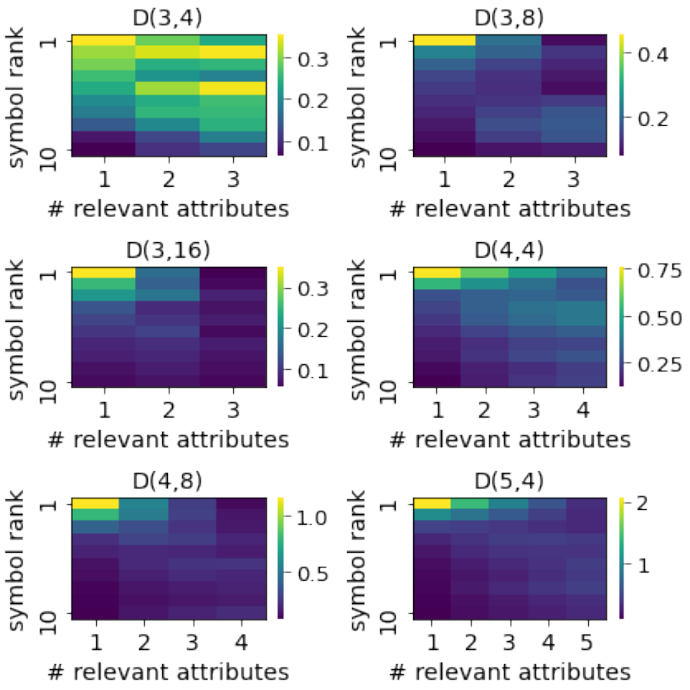


Figure 3.7: Average number of symbol occurrences per message for each level of abstraction. Symbols are ranked based on their occurrences for the most abstract concepts (i.e. with the fewest relevant attributes). Results are averaged across runs based on ranked symbols and shown only for the top ten ranks.

We use the different compositionality metrics to quantify the degree and nature of compositionality in the messages. Mean scores for each metric and data set can be found in Appendix B.3. The mean topsim score across data sets is 0.424. The score is even higher, with 0.501, if only concrete concepts are taken into account (as in a standard reference game). The mean posdis score across data sets is 0.115 and the mean bosdis score 0.406. So, there is compositional structure in the messages, and the agents prefer to use specific symbols per attribute value, independent of their position in the messages.

In additional experiments (see Appendix B.1), we trained the agents on $D(4, 8)$ with different vocabulary sizes, using factors $f \in \{1, 2, 3, 4\}$ of the minimal vocab size in Table 3.2. While a fully positional encoding can be achieved with a smaller vocabulary ($f = 1$), a fully position-independent encoding requires a larger vocabulary. Mean training accuracies for $f = 1$ are 0.936, and for all other factors > 0.99. Figure 3.8 shows the compositionality scores for each factor. Surprisingly, all scores tend to increase with the vocabulary size, regardless of whether the corresponding type of compositionality requires a large vocabulary size or not. Usually, vocabulary size is reduced to increase the pressure for compositional solutions (Kottur et al., 2017). In our case, compositionality probably increases with vocabulary size because the emerging compositional structure is largely non-positional.

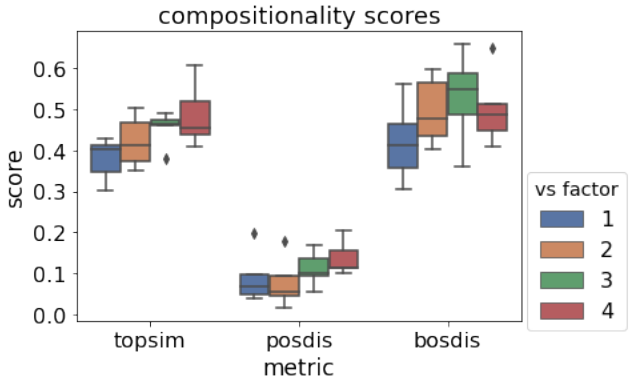


Figure 3.8: Boxplots of the compositionality scores for $D(4, 8)$ and different vocabulary sizes.

3.8 Conclusion

In this work, we developed a hierarchical reference game to study the emergence of hierarchical reference systems. In the game, concepts are defined as abstractions over a set of attributes. To refer to these concepts, our agents developed abstract terms and used these terms systematically, in the sense that they could generalize to novel objects and novel abstractions. It seems that, aside from more obvious strategies such as leaving out irrelevant information, the agents developed abstraction operators to explicitly indicate the irrelevance of certain attributes. Even more surprisingly, for some data sets, they used *the same* few symbols to indicate irrelevance across attributes, rather than a dedicated symbol per attribute. While the game design encourages the emergence of abstract concepts, the use of specific abstraction operators emerged without any explicit pressure.

In addition, our results suggest that compositional language may emerge as part of a hierarchical reference strategy. In the classical reference game, the sender typically tries to communicate the union of all object attributes. Without additional pressures, the emerging languages are not compositional (Dagan et al., 2021; Kottur et al., 2017; van der Wal et al., 2020). In the hierarchical reference game, in contrast, the sender must pick out specific attributes for communication, which potentially stimulates disentanglement. This interpretation is in line with the finding that the emergence of compositionality is supported by an increasing number of relevant events that can be referred to (Nowak et al., 2000). In the hierarchical reference game, cross-situational reuse is increased, as reference to attribute values occurs not only across objects but also across levels of abstraction.

We envision two main directions for future work. First, we would like to implement a hierarchical reference game with raw visual inputs instead of symbolic input vectors. Higgins et al. (2018) have developed a neural network (SCAN) that not only learns disentangled visual primitives in an unsupervised manner but also abstractions over such primitives from very few symbol-image pairs that apply to a particular concept. Combining our language emergence game with such a network would allow us to study the simultaneous emergence of abstract visual and linguistic concepts, as well as interactions between these two processes. Second, instead of hard-coding

the relevance vector, the relevance of certain attributes should arise from the agents' intentions. Ideally, the agents would play a more complex game and determine themselves which properties of the environment are relevant for their objectives in the current context. Besides, sender and receiver could use pragmatic reasoning (as for example in Choi et al., 2018; Kang et al., 2020; Yuan et al., 2020) to encode and decode which attributes should be emphasized to communicate certain concepts.

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—GRK 2340. We would like to thank Dieuwke Hupkes, Leon Schmid, Lucas Weber, and the anonymous reviewers for their valuable feedback and suggestions.

B Appendix

B.1 Varying distractor sampling and vocab size

Setup

We conduct control experiments, changing the vocabulary size, and changing the distractor sampling strategy. In the original experiments, the message space is much larger than the space of concepts that need to be communicated. By reducing the vocabulary size, we aim to test whether a smaller message space increases the probability of one-to-one associations between concepts and messages. In addition, the distractors are sampled from concepts that are one level more abstract than the target concept on the concept hierarchy. Here, we relax this assumption by sampling distractors from all levels of the concept hierarchy with equal probability. Together, these additional experiments allow us to extend our results to different vocabulary sizes, and more general distractor distributions.

We focus on a single data set. We use the data set with four attributes and eight values per attribute, $D(4, 8)$, which achieved the highest mean validation accuracies and normalized mutual information scores in the original setup. In the original experiments, we used a factor of 3 of the minimal vocabulary size 9 (8 for each value plus 1 for coding irrelevance). Now, we run the same experiment for factors of 1, 2, and 4; and in addition, we repeat the experiment for each factor with the alternative sampling strategy. Again, we conduct five runs for each factor and sampling strategy.

Results

Figure 3.9 shows the accuracy scores for the different vocabulary size factors, and the different distractor sampling strategies, where *unbalanced* refers to the original strategy of selecting distractors from more abstract concepts, and *balanced* refers to the control strategy of sampling distractors with equal probability from all levels of abstraction. For both sampling strategies, performance is higher if the vocabulary size is large, likely because having a larger message space increases the number of solutions. A larger vocabulary size seems to be particularly important if distractor sampling is balanced.

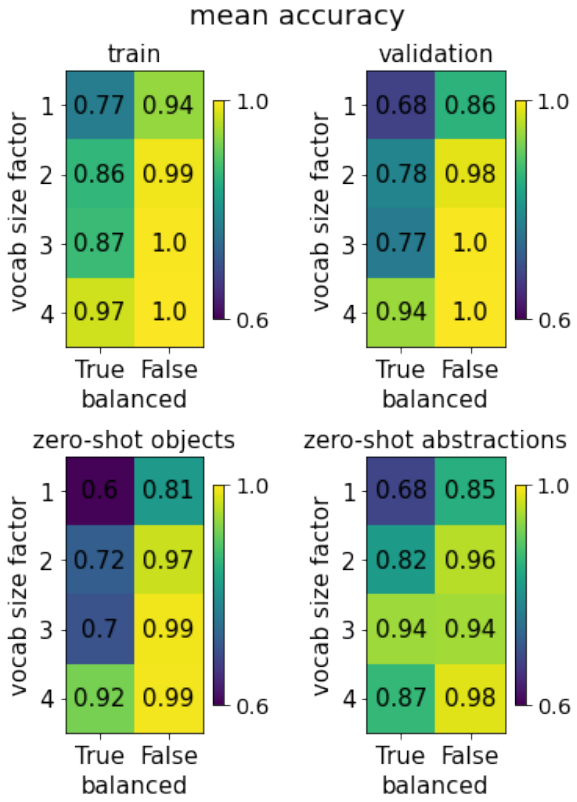


Figure 3.9: Mean accuracies for the control experiments across five runs, on the training data, the validation data, and the two zero-shot test sets. The y-axis gives the factor used to determine the vocabulary size, vocab size = factor $\times$ minimal vocab size, and the x-axis indicates whether distractors are sampled from concepts that are one level more abstract than the target concept (unbalanced), or sampled from all levels of the concept hierarchy with equal probability (balanced).

The original, unbalanced sampling strategy achieves higher performance than the control strategy on all data sets. So, choosing distractors very similar to the target facilitates learning, and probably also abstraction as suggested by the zero-shot evaluation with new abstractions. To make sure that the unbalanced sampling strategy only facilitates learning but does not make the task easier, we run an ablation test. We evaluate each sender-receiver pair on the validation set of the sampling strategy that was *not* used for training. For all vocabulary sizes and runs, the agents perform better on the balanced validation set compared to the unbalanced validation set, regardless of the sampling method used during training. In conclusion, sampling

distractor concepts that are very similar to the target concept makes the task more difficult but improves learning by increasing the pressure to communicate only relevant aspects, and thereby to develop abstract concepts.

These results are confirmed by the entropy-based evaluation metrics shown in Figure 3.10. Effectiveness and consistency are consistently lower for the balanced distractor sampling strategy. However, while the level of abstraction does not have a strong effect on the difference in effectiveness scores, the difference in consistency scores decreases continuously with the level of specificity. In line with the generalization ability, this suggests that the unbalanced sampling strategy supports the formation of abstract concepts by reducing the probability of successful target selection if irrelevant attributes are communicated.

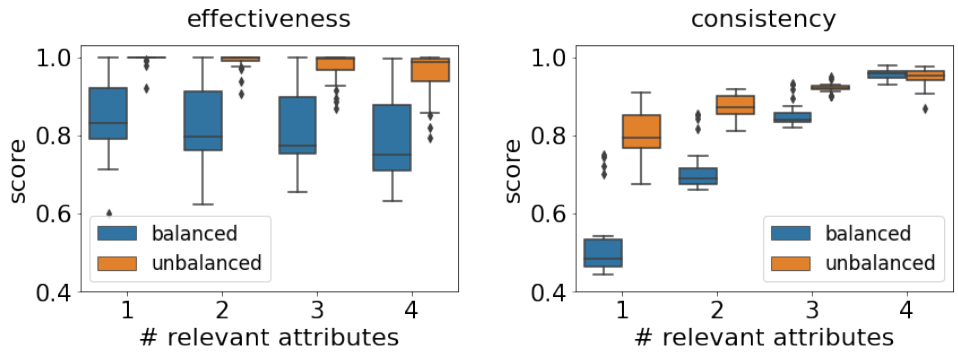


Figure 3.10: Effectiveness and consistency scores for balanced and unbalanced distractor sampling, separated for each level of abstraction. Distributions show the results across the different vocabulary sizes and runs. The level of abstraction is given on the x-axis: from left to right the concepts become more concrete.

B.2 Hyperparameter search

We ran our hyperparameter search for the three data sets spanning up the space of all data sets we use, $D(3, 4)$, $D(5, 4)$, and $D(3, 16)$ (see Table 3.1). We expected that hyperparameters working across all these extreme cases should also work for interpolations between them. Certain hyperparameters were fixed across the search. We used GRUs with Adam optimizer, and a GS temperature of 1.5 with an exponential decay rate. Message length cost was 0, and vocab size factor 3. We varied the following hyperparameters:

- batch size: $\{32, 64, 128\}$
- learning rate: $\{0.0005, 0.001\}$
- hidden layer dimension: $\{128, 256\}$
- embedding layer dimension:
  always half of the hidden layer dimension
- GS temperature decay rate: $\{0.97, 0.99\}$

For the grid search we stopped the training process after 60 epochs. All results can be found in our repository.

B.3 Compositionality scores

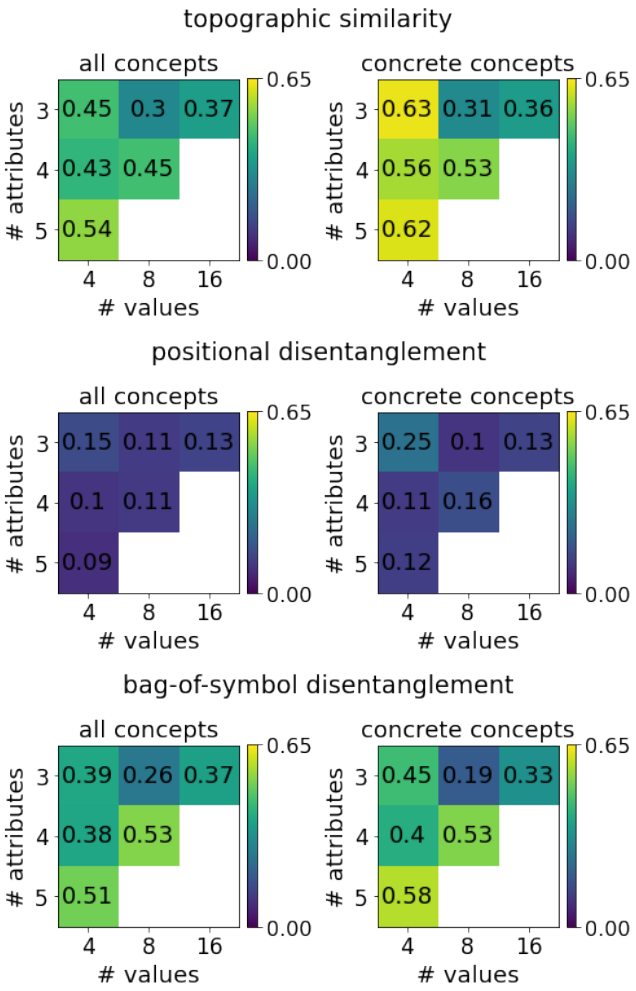


Figure 3.11: Mean compositionality scores per data set.

B.4 Qualitative examples

This section provides qualitative examples of concept-message pairs. Examples were randomly selected from the first run of each data set. Interestingly, this microcosm of random examples reflects all communication patterns that were identified in the quantitative analyses.

Mappings between concepts and messages

We are interested in whether the agents use the same message to refer to abstract concepts regardless of how these concepts are instantiated. Figure 3.12 shows the messages for a randomly selected concept at the highest level of abstraction (only one attribute is relevant), instantiated by different attribute vectors for each data set. Shown are twenty randomly selected instances each, and the examples are sorted by message.

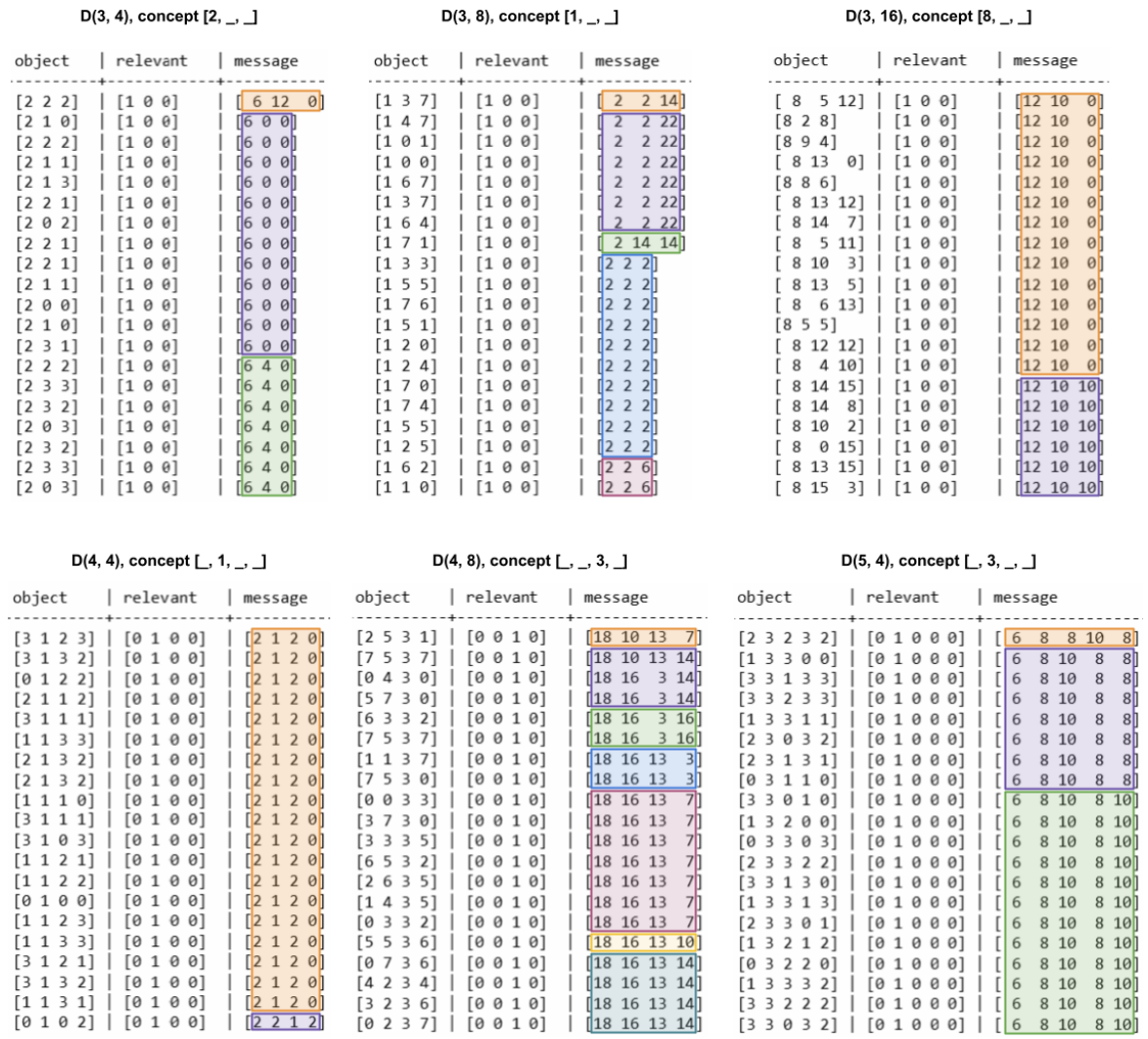


Figure 3.12: Example messages for one abstract concept per data set. For each data set, we randomly select a concept at the highest level of abstraction. We then randomly select 20 instances of that concept in the training data and display these instances together with the corresponding messages (from the first run). The same messages are grouped together in colored boxes.

Abstraction is relatively systematic. For all data sets, the agents group together different concept instances in their messages. For some data sets, the instances are grouped under very few messages. For example, the sender trained on $D(4, 4)$ groups together all example instances of the concept $(\_, 1, \_, \_)$ under just two different messages ($(2, 1, 2, 0)$ and $(2, 2, 1, 2)$). Across data sets, 2, 3, 5, or 7 different messages are used to describe the 20 example instances. In line with the quantitative results, there is no perfect one-to-one correspondence between *abstract* concepts and messages. How many different messages are used also depends on the abstraction strategy (see below).

Abstraction strategies

To visualize the agents' abstraction strategies, we randomly selected an object (i.e. attribute vector) for each data set, and show the messages for that object across the concept hierarchy, so for each abstraction in the training set. Because of their large number the examples are split into two figures, Figure 3.13 ($D(3,4)$, $D(3,8)$, and $D(3,16)$) and Figure 3.14 ($D(4,4)$, $D(4,8)$, and $D(5,4)$), which will be analyzed together. The messages will first be analyzed for compositional structure, and then for implicit versus explicit abstraction.

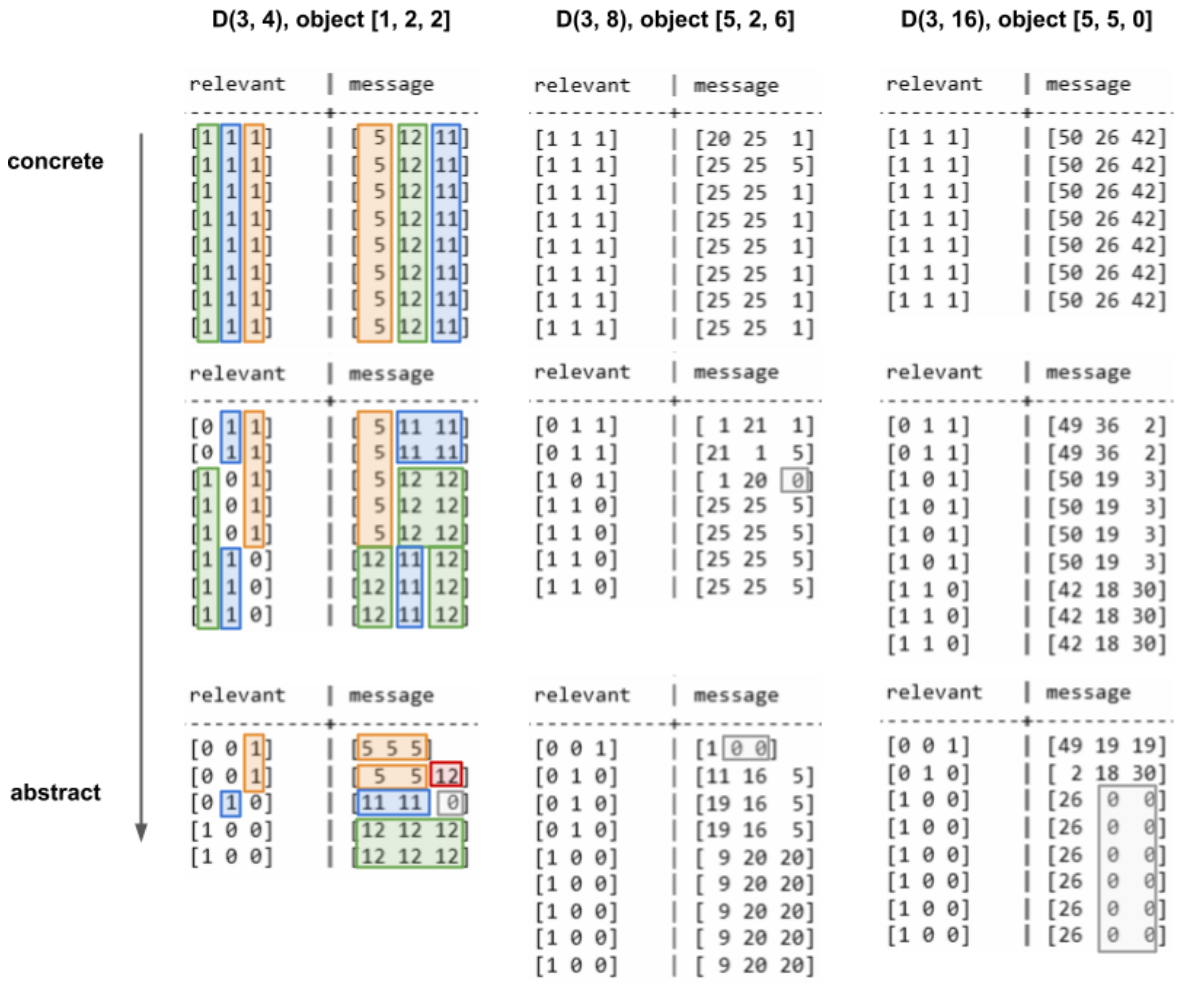


Figure 3.13: Messages for a random object at each level of abstraction available in the training data. The corresponding messages are shown for the first run of each data set: $D(3,4)$, $D(3,8)$, and $D(3,16)$. The highlighted patterns are explained in the text.

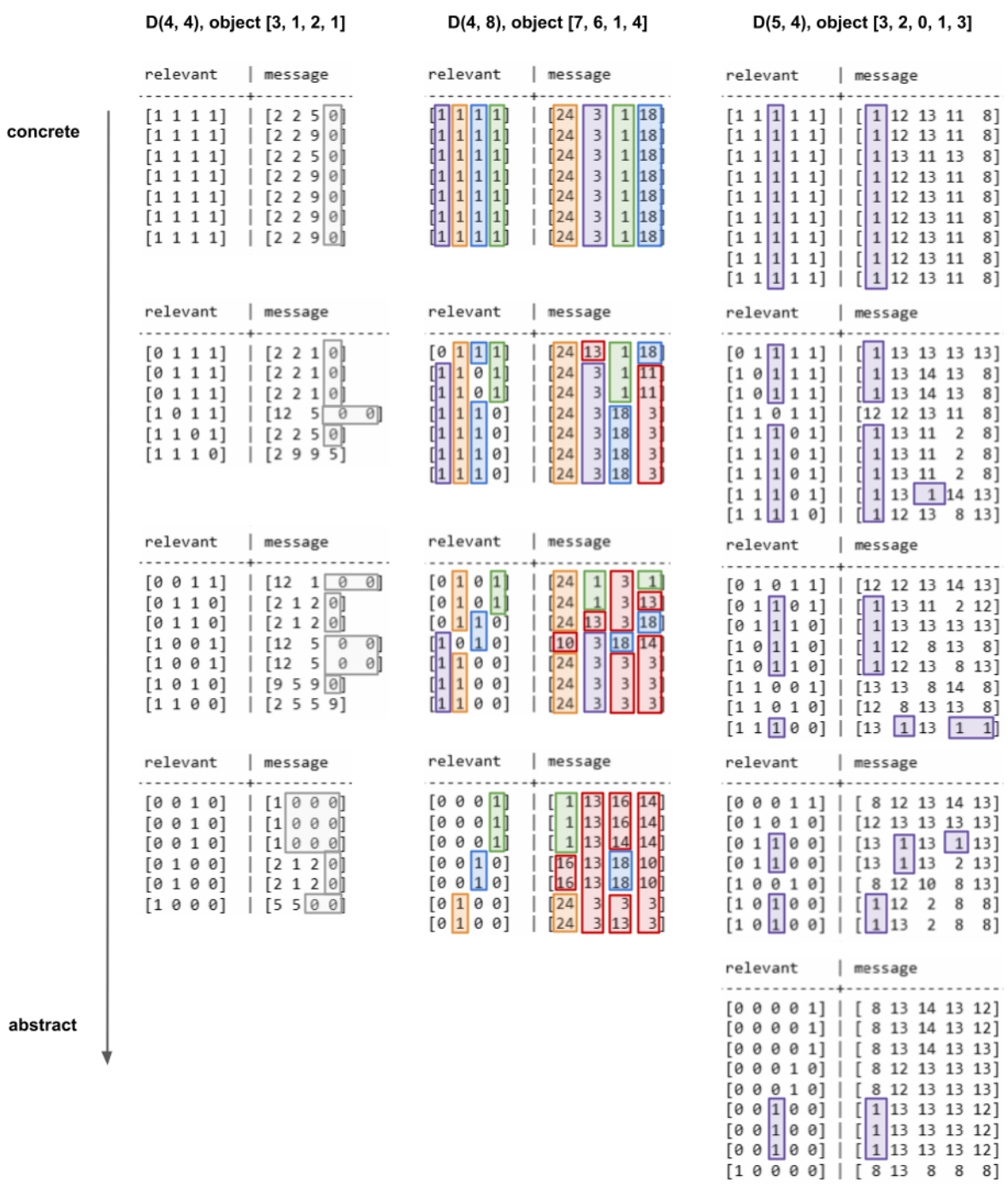


Figure 3.14: Messages for a random object at each level of abstraction available in the training data. The corresponding messages are shown for the first run of each data set: $D(4,4)$, $D(4,8)$, and $D(5,4)$. The highlighted patterns are explained in the text.

Compositional versus holistic abstraction. For some data sets, the agents seem to use *trivially compositional* messages, i.e. messages whose meaning corresponds to the intersection of meanings of their constituents. An unambiguous pattern can be identified for $D(3,4)$ and $D(4,8)$, where a mapping between each attribute value and a specific symbol can be established (color-coded in orange, green, blue, and purple). In the case of data set $D(4,8)$, the number of additional

"filler" symbols increases with the level of abstraction (color-coded in red). These might serve as abstraction operators (see below). For other data sets, like $D(5, 4)$, such mappings can only be identified for specific attribute values (color-coded in purple). Here, symbol 1 occurs if and only if the third attribute has the value 3. For all identified mappings, the symbols are used to encode specific attribute values relatively independent of their position, which is in line with the high bosdis and low posdis scores in our quantitative analyses.

For the remaining data sets, the messages are not unstructured but no one-to-one correspondences can be identified. For example, looking at the messages for $D(3, 16)$, symbol 26 might encode value 5 at position 1 for the most concrete and most abstract concepts but is not used at the intermediate level of abstraction. So, while the abstraction strategies are almost perfectly compositional in some cases, there are large variations between data sets, and potentially also runs and concepts.

Implicit versus explicit abstraction. The examples show instances of implicit and explicit abstraction strategies. Implicit abstraction is identified through shorter messages and more symbol redundancy for higher levels of abstraction; explicit abstraction through the use of abstraction operators. At least for some data sets, the messages tend to become shorter with increasing abstraction. E.g. messages become shorter in the case of $D(3, 16)$ and $D(4, 4)$ (the end-of-sequence symbol 0 is color-coded in gray).

Symbol redundancy and abstraction operators can best be identified in reference systems with compositional structure. $D(3, 4)$ is a perfect example of increasing symbol redundancy. Each symbol corresponds to a specific attribute value, and symbols are repeated to fill up the messages for more abstract concepts. E.g., the concept (1, 2, 2) is encoded as (5, 12, 11), the concept (1, _, 2) as (5, 12, 12), and the concept (1, _, _) as (12, 12, 12).

$D(4, 8)$, on the other hand, is a perfect example of explicit abstraction. As messages become more abstract the frequency of symbols that do not encode an attribute value ($\{3, 10, 11, 13, 14, 16\}$, marked in red) increases. Note that symbol 3 seems to serve both roles, encoding an attribute value as well as encoding abstractions. To confirm the intuition that these additional symbols serve as abstraction operators, we look at other abstract concepts for $D(4, 8)$. Figure 3.15 shows the messages for 20 random examples. Indeed, at least two of the abstraction operators occur in each message. Only symbol 11 does not occur and might serve a different function. The quantitative analyses suggest that usually less abstraction operators are used than in this specific example. Less compositional protocols may also use explicit abstraction operators or symbol redundancy but these cannot easily be identified in a qualitative analysis.

D(4, 8)

object	relevant	message
[0 2 3 5]	[0 0 0 1]	[8 16 13 14]
[3 4 5 2]	[0 1 0 0]	[14 16 14 14]
[6 0 5 4]	[0 0 1 0]	[16 13 9 9]
[0 7 0 5]	[0 0 1 0]	[16 7 16 14]
[1 0 7 4]	[1 0 0 0]	[10 3 17 14]
[5 3 5 4]	[0 1 0 0]	[10 16 14 14]
[2 5 3 4]	[1 0 0 0]	[10 3 14 21]
[7 5 5 0]	[0 0 0 1]	[19 16 3 16]
[6 3 0 6]	[0 0 0 1]	[23 13 14 14]
[0 6 4 2]	[0 0 0 1]	[20 13 14 14]
[2 0 7 5]	[1 0 0 0]	[10 3 14 21]
[7 7 0 2]	[0 0 1 0]	[16 7 13 10]
[3 0 6 0]	[0 0 0 1]	[19 16 3 16]
[7 6 5 4]	[0 0 0 1]	[1 3 16 14]
[6 4 5 7]	[0 0 0 1]	[4 13 14 14]
[6 2 2 1]	[0 0 0 1]	[22 13 22 14]
[3 0 5 3]	[0 0 1 0]	[16 13 9 9]
[5 0 7 3]	[0 1 0 0]	[24 16 14 14]
[6 7 7 2]	[0 0 1 0]	[16 13 6 6]
[2 0 3 6]	[0 1 0 0]	[24 16 14 10]

Figure 3.15: Messages for 20 randomly selected concepts at the highest level of abstraction, for the first run of $D(4, 8)$. The highlighted patterns are explained in the text.

4 Case study 3: Interactions between language and perception

This chapter presents case study 3. The chapter starts with a lay summary, which is followed by the content of the publication:

Ohmer, X., Marino, M., Franke, M., and König, P. (2022) Mutual influence between language and perception in multi-agent communication games. *PLOS Computational Biology* (accepted). https://arxiv.org/abs/2112.14518

4.1 Lay summary

Natural language is influenced by perception. Expressions for concrete concepts, like color terms, trivially depend on perception. Moreover, it has been argued that also abstract concepts can be understood by connecting them to concrete (perceptual) experiences. For example, we can reason about "time" in terms of a moving object ("Time flies", "The time has come ...") or about love as a journey ("It's been a long bumpy road."). But language also influences perception. For example, we can detect an object more easily if we are told what to look for. In particular, many studies show that language induces *categorical perception*, which means that objects are perceived as more similar if they have the same label, and as more different if they have different labels. We study these bidirectional influences in language emergence simulations with artificial agents.

We simulate the emergence of language with a reference game. The game is played by a sender and a receiver agent. In each round of the game, the sender sees a target object and sends a message to the receiver. Based on that message, the receiver has to identify the target object among a set of distractor objects. The objects in the game are abstract 3D shapes, defined by their color, scale, and shape (e.g. *tiny red sphere*). These three attributes are task-relevant, as the receiver has to distinguish between target and distractors based on these attributes. The agents receive a reward if they succeed. They start out with random messages and random guesses but they can use the reward signal to develop a working communication system (called "protocol") over time.

Our agents are implemented as deep neural networks. Deep neural networks are popular machine learning models but are also increasingly used to study human visual processing as well as the emergence of language. Our work extends existing research by investigating *interactions* between emergent language and visual representations with these models. The agent model consists of a vision module that processes images of the objects, and a language module that generates or interprets messages. We apply systematic manipulations to the agents' a) visual

object representations to study the effects on the emergent language, and b) communication protocol to study the effects on their visual representations.

In the first experiment, we study the influence of perception on language. We manipulate how agents perceive the objects in their world by changing their vision modules. It turns out that different visual biases lead to different emergent protocols. For example, agents that are good at perceiving color but not so good at perceiving shape tend to develop the same label for a green cube and a green sphere. Agents that are good at perceiving shape, in contrast, tend to develop different labels. In short, the agents tend to use the same labels for objects they perceive as similar and different labels for objects they perceive as different.

In the second experiment, we study the influence of language on perception. We test how the agents' visual representations change when learning or using different languages. In the language learning scenario, we fix the language and train only the receiver on the communication game. In the language emergence scenario, we pair agents with different visual representations, such that the emergent language forms a compromise between their preferences. In both scenarios, the agents' visual representations adapt to the language they learn or use. In particular, we find categorical perception effects: objects that are grouped under the same label become perceived as more similar, and objects with different labels as more different.

The mutual influence between language and perception can also lead to mutual improvement. Agents that accurately perceive all task-relevant attributes (color, scale, and shape) develop more successful communication protocols. Interestingly, playing the communication game generally improves the agents' visual representations. Agents are more successful if they can communicate all task-relevant attributes, and hence they learn to better represent these attributes regardless of their initial biases. This raises the question of whether our perceptual system might be shaped, in part, to facilitate communication. Additional analysis shows that *if* communication about certain object attributes is very important for the success of an agent in the world, the perceptual system will be structured to accurately represent these attributes. Our analysis remains agnostic about whether the perceptual system will arrive at these representations through learning processes in the individual or through evolutionary processes in a population.

In conclusion, our results can account for co-adaptation effects between language and perception. Besides, they point out ways to improve perception and communication in computational models. Visual representations of task-relevant information can become more accurate through communication; more accurate visual representations, in turn, lead to more successful communication protocols.

4.2 Abstract

Language interfaces with many other cognitive domains. This paper explores how interactions at these interfaces can be studied with deep learning methods, focusing on the relation between language emergence and visual perception. To model the emergence of language, a sender and a receiver agent are trained on a reference game. The agents are implemented as deep neural networks, with dedicated vision and language modules. Motivated by the mutual influence between language and perception in cognition, we apply systematic manipulations to the agents' (i) visual representations, to analyze the effects on emergent communication, and (ii) communication protocols, to analyze the effects on visual representations. Our analyses show that perceptual biases shape semantic categorization and communicative content. Conversely, if the communication protocol partitions object space along certain attributes, agents learn to represent visual information about these attributes more accurately, and the representations of communication partners align. Finally, an evolutionary analysis suggests that visual representations may be shaped in part to facilitate the communication of environmentally relevant distinctions. Aside from accounting for co-adaptation effects between language and perception, our results point out ways to modulate and improve visual representation learning and emergent communication in artificial agents.

4.3 Author summary

Language is grounded in the world and used to coordinate and achieve common objectives. We simulate grounded, interactive language use with a communication game. A sender refers to an object in the environment and if the receiver selects the correct object both agents are rewarded. By practicing the game, the agents develop their own communication protocol. We use this setup to study interactions between emerging language and visual perception. Agents are implemented as neural networks with dedicated vision modules to process images of objects. By manipulating their visual representations we can show how variations in perception are reflected in linguistic variations. Conversely, we demonstrate that differences in language are reflected in the agents' visual representations. Our simulations mirror several empirically observed phenomena: labels for concrete objects and properties (e.g., "striped", "bowl") group together visually similar objects, object representations adapt to the categories imposed by language, and representational spaces between communication partners align. In addition, an evolutionary analysis suggests that visual representations may be shaped, in part, to facilitate communication about environmentally relevant information. In sum, we use communication games with neural network agents to model co-adaptation effects between language and visual perception. Future work could apply this computational framework to other interfaces between language and cognition.

4.4 Introduction

Language is not an isolated system. Language is grounded in the physical world and serves to coordinate and achieve common objectives (Clark, 1992; Lewis, 1969). Under this functional perspective, it becomes obvious that language interfaces with many areas of cognition, among others, perception, action and embodiment, and social cognition (Bisk et al., 2020). To understand the origins and evolution of language it is important to take these connections into account. In this paper, we demonstrate how deep learning models of interactive language emergence can be used to study the relationship between language and other areas of cognition, focusing on the interface between language and visual perception.

Deep neural networks (DNNs), even though originally developed for engineering purposes, have been used to study human cognition in various fields. In terms of language emergence and language evolution, simulations with neural network agents have been used to model, for example, the emergence of color naming systems (Chaabouni et al., 2021; Kågebäck et al., 2020), contact linguistic phenomena (Harding Graesser et al., 2019), the emergence of word learning biases (Ohmer et al., 2020; Portelance et al., 2021), or the emergence of compositional structure (Choi et al., 2018; Li & Bowling, 2019; Ren et al., 2020). In terms of visual perception and representation learning, DNNs have been used to model brain activations in the visual cortex (Cichy et al., 2016; Khaligh-Razavi & Kriegeskorte, 2014; Kriegeskorte, 2015) and judgments of image similarity (Jozwik et al., 2017; Peterson et al., 2018). Our work extends existing research by studying *interactions* between language emergence and visual representation learning in neural network agents.

In human cognition, the influence between language and perception is bidirectional. Expressions for concrete concepts like colors depend on perception (Regier et al., 2007). But also abstract concepts can be understood and represented via metaphoric mappings to concrete concepts grounded in sensorimotor experience, for example in reasoning about time as a moving object ("The time will come when ...", "Time flies") (Lakoff & Johnson, 1980). Similarly, the effects of language on perception can be observed for high-level cognitive processes such as recognition as well as low-level processes such as discrimination and detection (Lupyan et al., 2020). In particular, language affects perceptual processing by imposing categorical structure (Forder & Lupyan, 2019; Winawer et al., 2007). We aim to analyze such bidirectional influences systematically, by studying the effects of variations in visual representations on emergent communication and vice versa.

More precisely, this paper looks at three questions: (i) how does perceptual bias affect language emergence, (ii) how does exposure to a particular linguistic input influence perceptual representations, and relatedly (iii) could perceptual representations be shaped by an optimization process towards successful communication of environmentally relevant distinctions. We use a conventional language emergence setup with two agents, a sender and a receiver, playing a reference game, based on the signaling game originally developed by Lewis (1969). The sender sees a target object and sends a message to the receiver. Using that message, the receiver tries to

identify the target among a set of distractor objects. By choosing this kind of game, we study the emergence and effects of *referential labels*, with denotations as sets of real-world objects. Reference is arguably a core function of language around which more complex functions are organized (Jackendoff, 1999). The agents have a vision module to process input images, and a language module to generate (sender) or interpret (receiver) messages. In line with many existing models (Havrylov & Titov, 2017; Lazaridou et al., 2020; Rodríguez Luna et al., 2020), the vision modules are implemented as pretrained convolutional neural networks (CNNs) and the language modules as recurrent neural networks (RNNs). The following three paragraphs enlarge on how this setup is adjusted to address each question.

(i) To study the influence of perception on language, we design agents with different visual biases, such that object representations vary between agents. We fix these biases and combine different agents to quantify differences in the emergent communication protocols. Given that concept formation in humans depends on perceptual similarity (Sloutsky, 2003), our manipulations target the similarity relationships between object representations. By applying a new method called *relational label smoothing* to the CNN pre-training we modify the class labels, such that the resulting representational similarities between objects vary for different conditions. Thereby, we can test how language groundedness is influenced by these differences, and how certain perceptual predispositions can benefit communication.

(ii) To study the influence of language on perception, we allow agents to adapt their visual representations (CNN weights) while playing the communication game. We measure how perception adapts to fixed languages in language learning, or to different communication partners in language emergence. To analyze changes in perception we again rely on similarity relationships between visual representations. Several studies concerning categorical perception have shown that language affects perceptual similarity (Lupyan et al., 2020). Moreover, developing a system of similarity relationships along *relevant* perceptual dimensions (e.g., color, shape, magnitude, texture) is a major achievement in child development (Smith, 1989). In our case, relevance is determined by the communication game. Thus, our setup not only allows us to study how language influences perceptual similarity but also how a system of similarity relationships with respect to task-relevant dimensions can evolve via communication.

(iii) Finally, an evolutionary analysis explores whether an agent's perceptual system might be optimized over time to facilitate communication about relevant aspects of the environment. As in (i), we consider agents with different, fixed perceptual biases. We train an extensive variety of agent combinations on the reference game and derive a payoff matrix for a symmetric population game. We subject this payoff matrix to a simple analysis in terms of evolutionary stable states (ESSs) (Maynard Smith, 1974). Thereby, we can determine whether certain perceptual representations (biases) are more likely to prevail in an adaptation process to the demands of linguistic interaction, which in our case defines the agents' environment. Importantly, ESS-analysis does not entail a commitment to an underlying process of biological evolution. ESSs can also be considered the rest points of other (agent-internal) optimization processes.

4.4.1 Related work

Communication games have been used to study the emergence and evolution of language theoretically (Crawford & Sobel, 1982), experimentally (Blume et al., 1998; Crawford, 1998), and computationally (Kirby, 2002b). Artificial intelligence research has also emphasized the importance of learning to communicate through interaction for developing agents that can coordinate with other, possibly human agents in a goal-directed and intelligent way (Mikolov et al., 2015). It has been shown that by playing communication games, artificial (robotic) agents can self-organize symbolic systems that are grounded in sensorimotor interactions with the world and other agents (Bleys et al., 2009; Steels, 1998, 2001; Steels & Belpaeme, 2005). For example, in a case study with color stimuli, simulated agents established color categories and labels by playing a (perceptual) discrimination game, paired with a color reference game (Steels & Belpaeme, 2005). Bleys et al. (2009) extended these findings to robotic agents, demonstrating that successful color naming systems emerge in spite of differences in the agents' perspective. These studies are mainly interested in how a categorical repertoire can become sufficiently shared among the members of a population to allow for successful communication. Our analyses, in contrast, assume that successful communication will emerge, and focus on how visual representations and language shape each other.

Over the past years, research using communication games to study language emergence in DNN agents has been gaining popularity (Lazaridou & Baroni, 2020). Some of these models skip any form of perceptual processing by using symbolic input data (Bouchacourt & Baroni, 2019; Chaabouni et al., 2020a; Kharitonov & Baroni, 2020). Even though other models implement a visual processing system and work with image data (Havrylov & Titov, 2017; Lazaridou et al., 2017), they have rarely been used to explore the relation between language and visual perception. Notably, Rodríguez Luna et al. (2020) examined the effects of natural differences in object appearance (such as frequency, position, and luminosity) on emergent communication. Apart from that, Bouchacourt and Baroni (2018) measured the alignment between agents' internal representations and conceptual input properties to determine whether emergent language captures such properties or relies on low-level pixel information. Still, these models usually extract object representations from fixed, pre-trained CNNs. As a result, they make claims about how the emergent language relates to the input, not the visual perception of that input. In our work, we exploit the flexibility of modern setups and introduce systematic variations in the agents' visual processing, such that we can establish a relationship between differences in visual processing and differences in emergent protocols.

4.5 Materials and methods

4.5.1 Data set

We use the *3dshapes* data set (Burgess & Kim, 2018). The data set contains images of 3D shapes in an abstract room, generated from six latent factors, which can vary independently: floor color (10 values), wall color (10 values), object color (10 values), object scale (8 values), object shape (4 values), and object orientation (15 values). We use a subset of four different object colors (red, yellow, turquoise, purple), and four different object scales (equally spaced from smallest to largest); amounting to 96000 different images. For our purpose, we define objects by color, scale, and shape of the geometric shape, such that there are $4^3 = 64$ different objects. The term "object" refers to an object class, such as "tiny red cube", with each image representing an instance of such an object. Consequently, if we say that two agents see the same object, e.g., a tiny red cube, they both see an object that agrees on the relevant attributes (object color, object scale, and object shape), but not necessarily on the irrelevant ones (floor color, wall color, object orientation), e.g., they might both see a tiny red cube, one against a yellow wall and another against a green wall. Similarly, when we say that two objects are different, they differ in at least one of the relevant attributes but may agree on all irrelevant ones.

4.5.2 Communication game

Two agents, sender S and receiver R, play a reference game where one round of the game proceeds as follows:

1. A random object is selected as the target.
2. S sees an image of the target and produces a message. Messages have length L and consist of a sequence of symbols $(s_1, ..., s_L)$ from vocabulary $V = \{0, \ldots, |V| - 1\}$.
3. R sees a possibly different image of the target and additionally k random distractor images, showing other objects. Based on the message from S, R tries to select the image showing the target.
4. If R succeeds, both agents receive a positive reward, $r = 1$, otherwise they receive zero reward, $r = 0$.

Three attributes—color, size, and shape—define what we call "object". Sender and receiver see potentially different images of the same target object, while the distractor images show different objects. Consequently, it lies in the nature of this game, that *conceptually relevant* (i.e. class-defining) attributes and *task-relevant* attributes coincide.

4.5.3 Model

The model components and their interactions in the communication game are shown in Fig 4.1. Sender and receiver each have a vision module to process images, i, and a language module to

generate (sender) or process (receiver) discrete messages, m. The sender maps the input image to a probability distribution over messages, $\pi_S(m \mid i)$, by sequentially generating a probability distribution across symbols conditioned on the symbols produced so far. The receiver maps the input message onto a probability distribution over (target and distractor) images, $\pi_R(i \mid m)$. These distributions define the agents' policies. During training, actions are sampled from the policies, whereas for testing the arguments of the maxima are used.

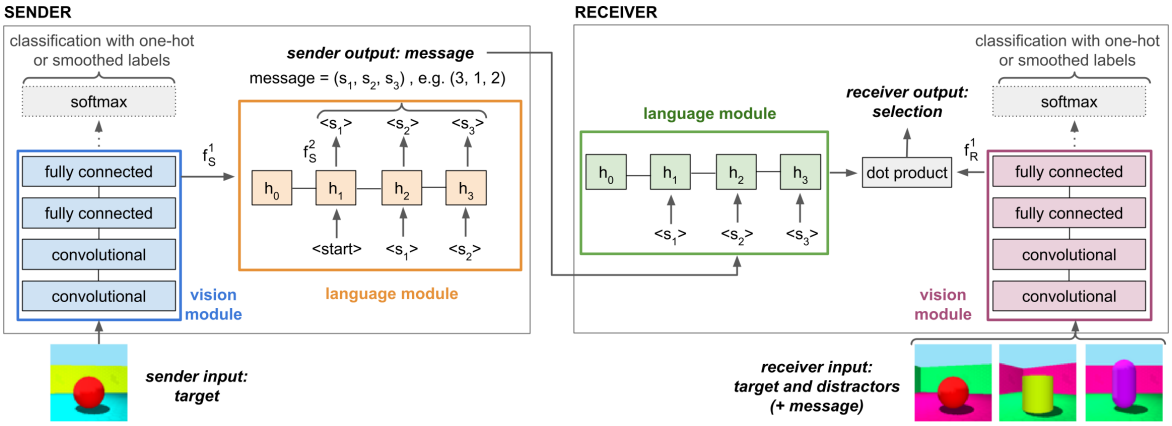


Figure 4.1: Schematic visualization of sender and receiver architecture and their interaction in one round of the reference game. The sender takes an image of the target object as input. The image is processed by the sender's vision module and the resulting activations are used to initialize the hidden state, h_0, of the sender's language module. The initial input to the sender's language module, ⟨start⟩, is a zero vector of the same dimensionality as the symbol embeddings, and at each time step a symbol is sampled from its output distribution. The generated message is processed by the receiver's language module. In addition, the target and the distractor images are processed by the receiver's vision module. The final selection probability is proportional to the dot product between the receiver's final hidden state and the image embeddings.

The vision module, $v(\cdot)$, is a CNN pretrained to classify the 64 different objects. The agents use the activations of the fully connected layer before the final softmax layer as object representation. The language module, $l(\cdot)$, consists of an embedding layer and a gated recurrent unit (GRU) layer (Cho et al., 2014). Each agent has an additional fully connected layer, $f^1(\cdot)$, mapping the visual representations onto the same dimensionality as the GRU hidden state. For the sender, the output of f_S^1 is used to initialize the hidden state of the language module. The sender has an additional fully connected layer, $f_S^2(\cdot)$, mapping the GRU hidden state onto a probability distribution across symbols at each time step, t, such that $\pi_S(m = (s_1, \ldots, s_L) \mid i) = \prod_{t=1}^{L} \pi_S(s_t \mid s_{k<t}, i)$, with $\pi_S(s_t \mid s_{k<t}, i) \propto f_S^2(h_t)$. For the receiver, the dot product between the output of layer f_R^1 and the final GRU hidden state defines the selection policy: $\pi_R(i \mid m) \propto \exp\left(f_R^1(v_R(i)) \cdot l_R(m)\right)$.

4.5.4 Introducing perceptual biases via relational label smoothing

In order to investigate the influence of differences in perception on emergent language, we develop a method called *relational label smoothing*, which allows us to systematically manipulate the CNN pretraining and thereby to create agents with different perceptual biases. We aim to have four conditions, next to the unmanipulated DEFAULT. Specific biases for either of the object-defining attributes—color, scale, and shape—make up three of these conditions. E.g., in

the COLOR condition, color similarities are amplified. In addition, we experiment with an ALL condition, where we amplify similarities for all three attributes simultaneously.

Relational label smoothing calculates the target at training time as a weighted sum of the usual one-hot target, $\mathbf{y}_0$, and a relational component, $\mathbf{y}_r$, according to

$$\mathbf{y} = \sigma \mathbf{y}_r + (1 - \sigma)\mathbf{y}_0 \,,$$

where $\sigma \in \mathbb{R}$ is the smoothing factor, controlling the strength with which the relationship(s) should be enforced.

To enforce object similarities along one specific attribute (or dimension), a, we use a single-level hierarchical version of relational label smoothing. If i is the true object class, we define superclass C_i as the set of object classes having the same value as i for a. Then $\mathbf{y}_r$ is given by

$$y_{r_{ij}} = \begin{cases} (n-1)^{-1} & j \in C_i \text{ and } i \neq j \\ 0 & \text{else} \end{cases} \,,$$

where n is the number of object classes in C_i. E.g., in the COLOR condition, if the training sample is a red object, the relational component, $\mathbf{y}_r$, is a uniform distribution of $1/(16-1)$ across the class indices of the other 15 red objects, see Fig 4.2.A, which increases the representational similarity between red objects, and analogously that of objects sharing other color values, see Fig 4.2.B.

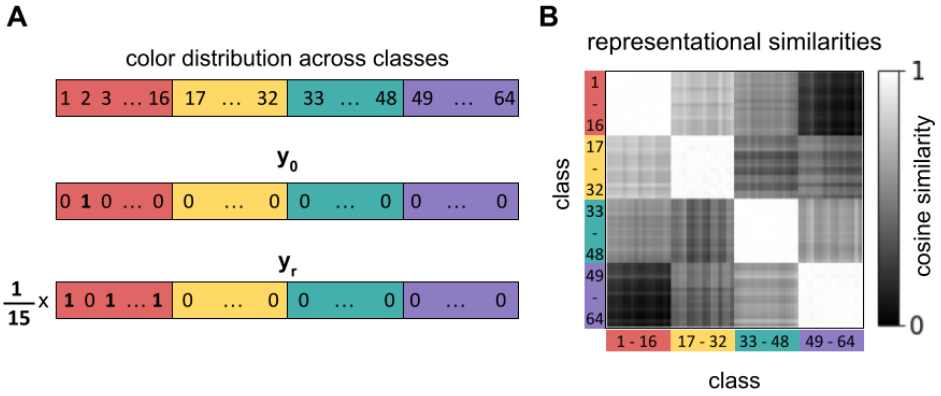


Figure 4.2: Creating perceptual bias with relational label smoothing. (A) Example of how the training targets (labels) are adapted to induce a COLOR bias. To generate a CNN with a COLOR bias, some of the target weight is spread across all other classes that *have the same color as the target object*. In our data set, there are 64 different object classes. The first sixteen classes comprise red objects (classes 1–16), followed by yellow objects (classes 17–32), turquoise objects (class 33–48), and purple objects (classes 49–64). For example, if the input image belongs to class 2 ("tiny red cylinder"), the usual target label, $\mathbf{y}_0$, is a one-hot vector where the entire weight lies on the true class index. The relational component, $\mathbf{y}_r$, spreads some of the target weight onto all other red objects. The target vector used for training is a weighted average of the original target and the relational component. Analogously, to introduce a scale/shape bias, some of the target weight is spread onto all other objects of the same scale/shape as the input object. (B) Representational similarity matrix for the COLOR CNN after training ($\sigma = 0.6$). Entries at position (i,j) correspond to the average cosine similarity between the CNN activations for images of class i and the CNN activations for images of class j (based on the penultimate fully-connected layer). The white 16×16 blocks on the diagonal indicate that objects of the same color are perceived as very similar to each other.

In order to enforce relationships for multiple attributes in a single model, we generalize the

previous definition to include $\mathbf{y}_r$ to be a sum over relational components,

$$\mathbf{y}_r = \frac{1}{N} \sum_{a=1}^{N} \mathbf{y}_{r_a} ,$$

where N is the number of attribute relationships, and $\mathbf{y}_{r_a}$ represents the relational component from attribute a. To calculate the relational component for the ALL condition, we average the relational components from the COLOR, SCALE, and SHAPE conditions.

4.5.5 Training and hyperparameters

We use a train/test split of 0.75/0.25.

General setup. The general training setup varies depending on which direction of influence between perception and language is being investigated. A schematic overview of these variations is shown in Fig 4.3. The agents' vision modules are always pretrained on a classification task, and different perceptual biases can be achieved via the different pretraining conditions explained above. Categories do not have to originate from language. Categories can also be formed through interactions with the world, and nonhuman animals as well as preverbal human infants can learn categories (Sloutsky & Deng, 2019). Of course, these categories can still be lexicalized later on. The classification task is motivated by this ability to form categories through interactions with the world. While we do not explicitly model such interactions we assume they take place nonetheless. To study the influence of differences in perception on communication (Fig 4.3, top row), we train a sender and a receiver with fixed vision module weights on the communication game. The evolutionary analysis uses the same setup. Here, multiple games between sender-receiver pairs are used to approximate the communicative success of agent populations with different perceptual dispositions. To study the influence of language on perception, we consider language learning and language emergence (Fig 4.3, center and bottom row). In the language learning scenario, the language is fixed—using a trained sender—and only the receiver is trained, while in the language emergence scenario, both agents are trained. Importantly, in both scenarios, not only the language module but also the vision module is trained, such that changes in perception can occur. When learning to communicate, visual representations may adapt but they are still constrained by the functions of the visual system. In our case, this function is limited to object recognition (classification). To ensure that the agents' perceptual ability does not deteriorate to processing only aspects relevant to the communication game, training on the classification task used for pretraining continues. The loss function is generated by adding the classification loss and the communication game loss together.

CNN pretraining. The CNN architecture consists of two convolutional layers with 32 channels, followed by two fully connected layers with 16 nodes, and a final softmax layer. The first convolutional layer is followed by a 2×2 max-pooling layer. For pretraining, we use stochastic

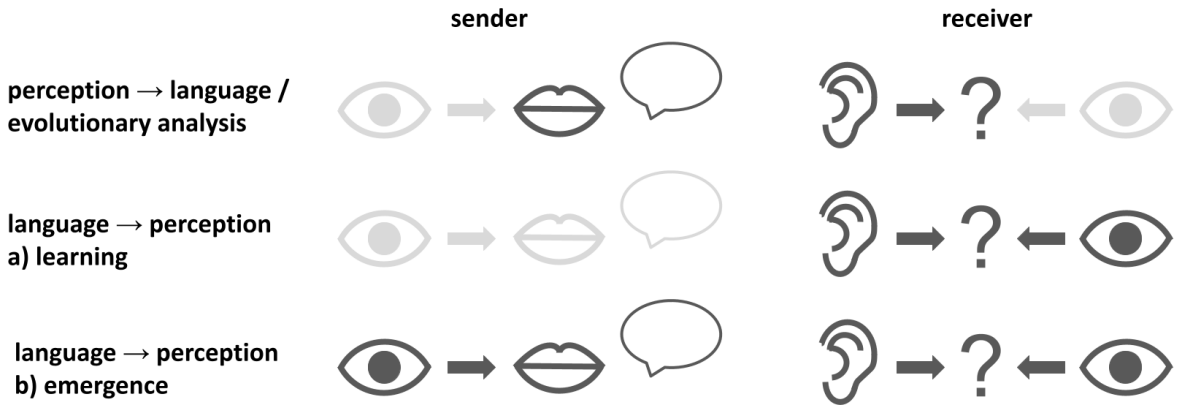


Figure 4.3: Illustration of the training setups. The vision module is represented by an eye, the language module by a mouth (sender) or an ear (receiver). The speech bubble represents the message, and the question mark the receiver's selection. Modules that are not trained, i.e. have fixed weights, are light gray. Modules that are trained are dark gray. Note that the vision modules in the two language emergence scenarios (center and bottom row) are trained on the communication game and simultaneously also on the original object classification task.

gradient descent (SGD) with learning rate 0.001 and batch size 128, and train for 200 epochs. We set smoothing factors as high as possible while keeping the classification accuracy close to maximal. For the COLOR, SCALE, and SHAPE conditions, we use a smoothing factor of $\sigma = 0.6$. For ALL, the weight is distributed across more classes, which allows for a higher smoothing factor of 0.8. All networks achieve test accuracies > 97%.

Communication game. For most simulations, we use vocabulary size $|V| = 4$, message length $L = 3$, and $k = 2$ distractors. In principle, this allows agents to use a distinct symbol for each object and thereby to achieve maximal reward. As there are only a few distractors, agents may achieve relatively high rewards with suboptimal strategies. It is in the variation of such local solutions that we hope to identify linguistic differences that reflect perceptual biases and vice versa. We also run control experiments with a larger vocabulary size and more distractors, as well as control experiments changing the task-relevance of individual attributes. The agents minimize the negative expected reward, $-\mathbb{E}[r]$, and their trainable weights are updated using REINFORCE (Williams, 1992), which is a basic policy gradient algorithm. We train all agents using Adam with learning rate 0.0005 and batch size 128. Embedding and GRU layer each have a dimensionality of 128. We add an entropy regularization term (Mnih et al., 2016) of 0.02 to sender and receiver loss to encourage exploration. The vision modules are initialized with the weights of the pretrained CNNs. When both agents are trained, training proceeds for 150 epochs, if only the receiver is trained (language learning) for 25 epochs.

4.5.6 Evaluation

We are interested in the mutual influence between perception and language. Accordingly, we devise metrics to quantify perceptual biases as well as linguistic biases.

Perception

Let $A = \{color,\ scale,\ shape\}$ be the set of object attributes, and V_a all values that attribute $a \in A$ can take on, e.g., $V_{scale} = \{tiny,\ small,\ big,\ huge\}$.

Given a set of inputs, *representational similarity analysis* (RSA) (Kriegeskorte et al., 2008) measures the similarity between two representational spaces, by calculating the pairwise distances (in our case similarities) of input representations in either space and then correlating the two distance matrices. We use the analysis in two different ways. In the first case, RSA quantifies how well an agent's visual representations capture conceptually relevant attributes. Here, the two spaces under comparison are the space of the agent's visual representations generated by $v(\cdot)$, and a symbolic space of k-hot encoded attribute vectors ($k = |A| = 3$). In the second case, RSA quantifies the degree of perceptual alignment between an agent and its communication partner, and the two spaces under comparison are the two different visual representation spaces. In a first step, we extract $N = 50$ random example images for each object (class) and generate a representational similarity matrix (RSM) for each space under comparison, by calculating the pairwise cosine similarities between the corresponding representations, $sim_{cos}(r_i, r_j) = \frac{r_i^T r_j}{\|r_i\|\|r_j\|}$. Fig 4.2.B shows an example of an RSM for a COLOR agent. In a second step, the actual RSA score is calculated as the Spearman correlation between the RSMs of the two spaces under comparison.

The RSA score with respect to the attribute template tells us how well differences in the underlying compositional object structure correlate with differences in the agent's visual representations. Fig 4.4.A shows the RSM calculated from k-hot encoded attribute vectors, which serves as a ground-truth template. We can also use RSA to quantify whether agents can represent similarity relationships for some attributes better than for others. In order to do so, we replace the k-hot attribute vectors above by one-hot vectors encoding the values V_a of a specific attribute a, and repeat the procedure for each attribute $a \in A$, resulting in separate RSA scores for color, scale, and shape. Fig 4.4.B shows the color RSM template. Notice, that the RSA scores for individual attributes attenuate each other, as the agent's representations cannot simultaneously match all three templates. If one score is higher than the others, the agent represents one attribute at the cost of the others and is said to have a perceptual bias for that attribute. We denote the general RSA score (including all attribute values) by RSA, and the scores for a specific attribute by RSA_a.

Language

We use an information-theoretic evaluation to quantify the linguistic bias. Communicative success is based on what information about the target objects, O, the sender encodes in the messages, M, but also what information the receiver decodes from the messages to determine its object selections, S. Communicative success depends on both these factors, suggesting a three-way analysis, see Fig 4.5 (left), which would allow us to quantify the shared and distinct

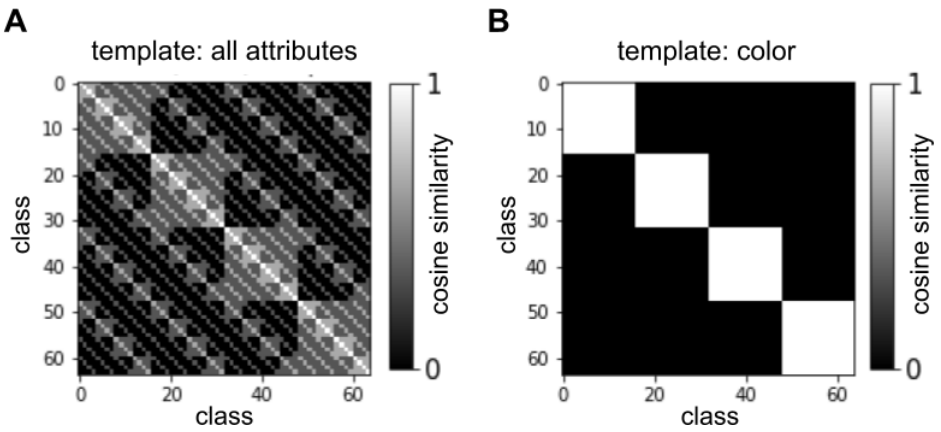


Figure 4.4: Quantifying perceptual bias. (A) Object similarities calculated from 3-hot encodings based on all three attributes. This template is used in the RSA calculation to measure how well conceptually relevant attributes are encoded. (B) Object similarities calculated from 1-hot encodings based on color value. This template is used to calculate RSA_{color}.

information between all combinations of objects, messages, and selections. However, in our experiments, the shared information between objects and selections is entirely predicted by the messages, since the receiver can only make selections based on message content (for details see Appendix C.1). Therefore, we can skip the object-selection interface, leading to separate analyses of the relation between objects and messages, and messages and selections Fig 4.5 (right).

The mutual information between two random variables, $I(X, Y)$, measures how predictive these variables are of each other

$$I(X, Y) = H(Y) - H(Y \mid X) = H(X) - H(X \mid Y)\,,$$

where $H(X)$ is the marginal entropy and $H(X \mid Y)$ the conditional entropy defined as

$$H(X \mid Y) = -\sum_{y \in Y,\, x \in X} p(y, x) \log \frac{p(y, x)}{p(y)}\,.$$

The conditional entropy indicates how much uncertainty about X remains (on average) after learning Y. It turns out that, in all our experiments, the analysis of sender and receiver are symmetric in that $H(O \mid M) \approx H(S \mid M)$, $H(M \mid O) \approx H(M \mid S)$, and accordingly also $I(O, M) \approx I(M, S)$. Therefore we limit our analysis to the sender.

The conditional entropy, $H(O \mid M)$, quantifies the degree of uncertainty about the objects when knowing the messages that were sent. In reverse, to measure how much information about the objects is encoded in the messages, we can define an effectiveness score by

$$E(O, M) = 1 - \frac{H(O \mid M)}{H(O)}\,,$$

with $E(O, M) \in [0, 1]$. To measure linguistic bias, we can define an effectiveness score for individual attributes. Let O_a be the values of attribute a for all objects, and M the generated messages as above, then we can measure how much information about a is encoded in the

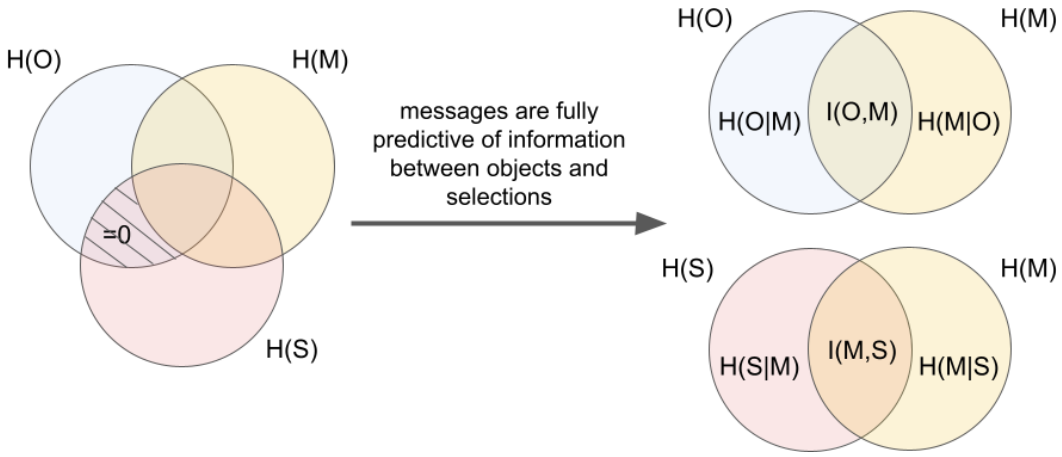


Figure 4.5: Schema of the information in the target objects, O, the corresponding messages, M, and objects selected by the receiver, S. H denotes entropy and I mutual information. The object-selection interface is entirely predicted by the messages as the mutual information between objects and selections given messages (shaded region on the left side) is zero. Therefore we can separate the analysis of sender (objects-messages) and receiver (messages-selections) as shown on the right. Note, the schema is not an actual set-theoretic representation and serves illustrative purposes only.

messages as $E(O_a, M)$. It follows, that

$$\overline{E(O_a, M)} = \frac{1}{|A|} \sum_{a \in A} E(O_a, M)$$

measures how well all conceptually relevant attributes are communicated. Unlike the RSA scores for individual attributes, $E(O_a, M)$, can be maximal for all attributes at the same time.

4.6 Results

This section presents analyses and results. At first, a validity check of label smoothing as a method to induce selective visual biases is performed. Then, each of the three questions under investigation is treated separately.

4.6.1 Perceptual biases generated via label smoothing

Relational label smoothing can systematically manipulate perception. In order to test the validity of our manipulations, we check whether relational label smoothing induces the intended biases. As the agents' vision modules use object representations from the penultimate CNN layer, we quantify the biases for that layer using RSA. t-SNE plots (van der Maaten & Hinton, 2008) and pairwise class similarities of object representations can be found in Appendix C.2 and Appendix C.3. Table 4.1 shows the RSA scores for each of the five pretraining conditions. Surprisingly, the DEFAULT CNN represents differences in color values much more accurately than differences in other attributes. This inherent color bias may be due to the networks' direct access to color information via the RGB channel input (Hill, Clark, et al., 2020). COLOR, SCALE,

and SHAPE networks mostly capture differences in the respective attribute. The ALL network represents differences in all three attributes, which can be seen from relatively high RSA scores per attribute, as well as a higher overall RSA score. Note, maximum values per attribute are smaller than in the other conditions due to mutual attenuation. In conclusion, by default, object representations extracted from CNNs are biased towards representing color information but relational label smoothing can shift this bias to other attributes as well as improve coverage of the entire input topology.

Table 4.1: RSA between visual object representations and object attributes for each pretraining condition. Scores are calculated between object representations and k-hot attribute encodings, RSA (bottom row), as well as for each individual attribute a, RSA_a.

	default	color	scale	shape	all
RSA_{color}	0.633	0.750	0.019	0.021	0.440
RSA_{scale}	0.101	0.019	0.750	0.025	0.319
RSA_{shape}	0.056	0.017	0.015	0.748	0.424
RSA	0.439	0.437	0.437	0.442	0.675

4.6.2 Influence of perception on language

To quantify the influence of different visual biases on emergent communication, we trained agents with different visual biases (and fixed vision module weights) on the communication game. For all CNNs (DEFAULT, COLOR, SCALE, SHAPE, ALL) we trained a sender-receiver pair where both agents used the same vision module and thus had the same bias. In addition, to evaluate the impact of sender versus receiver bias we ran experiments combining a DEFAULT receiver with each type of sender, and combining a DEFAULT sender with each type of receiver. We conducted twenty runs per agent combination. All agents learned to play the game, with mean test rewards ranging between 0.914–0.968 (details about the agents' performance follow later in this section).

Perceptual biases systematically shape emergent language. We begin by analyzing the effect of perceptual biases on emergent language when both agents have the same bias. We use the effectiveness score to measure how much information about specific attributes is contained in the messages. The results for each type of bias and each attribute are shown in Fig 4.6.A. The five blocks on the x-axis show the perceptual bias conditions, with each bar representing one of the three attributes. In the DEFAULT condition (left) the messages are strongly grounded in object color, which can be attributed to the inherent color bias of the DEFAULT CNN. Agents with a color, scale, or shape bias (central three blocks), ground their messages to a large extent in the attributes they have a perceptual bias for. Overall, the effectiveness across conditions is significantly higher for biased attributes ($M = 0.868$) than unbiased attributes ($M = 0.468$), as indicated by a bootstrapped 95% confidence interval (CI) for the difference in means of $[0.355, 0.444]$. Qualitatively, the observed patterns prevail also if the vocabulary size and the number of distractors are increased, both of which encourage the agents to communicate

more information about each attribute (see Appendix C.4). It seems that if agents are good at perceiving object similarities along specific dimensions, they prefer to communicate these dimensions over others.

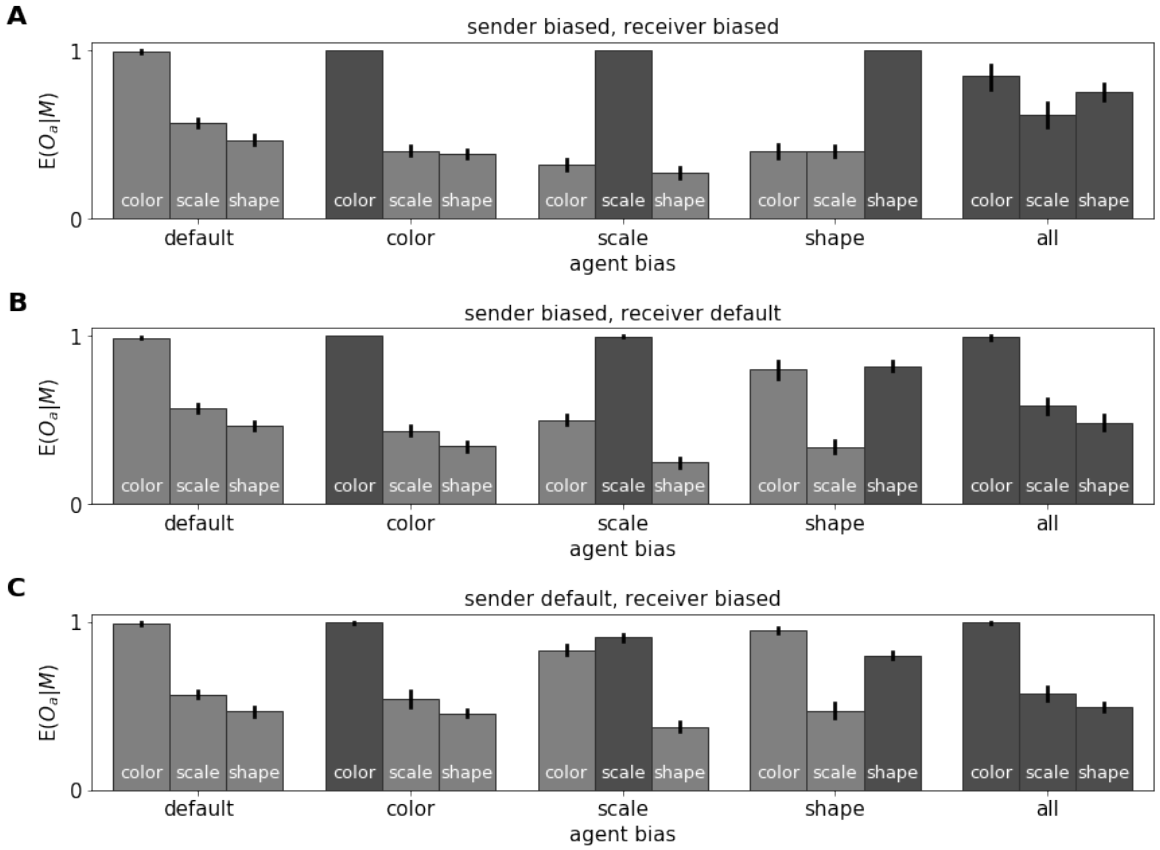


Figure 4.6: Effectiveness per attribute for different pairings of senders and receivers. Pairings are (A) biased sender and biased receiver, (B) biased sender and DEFAULT receiver, and (C) DEFAULT sender and biased receiver. The x-axis shows the agents' perceptual biases. The bars are labeled with the attribute a used for calculating $E(O_a|M)$, with attributes enforced via label smoothing in dark gray. We report means and bootstrapped 95% CIs of twenty runs each.

Sender bias is more influential than receiver bias. Effectiveness scores for varying the sender bias in combination with a DEFAULT receiver are shown in Fig 4.6.B, and for varying the receiver bias in combination with a DEFAULT sender in Fig 4.6.C. The results for DEFAULT from part (A) are repeated as a reference. Comparing part (B) to part (A) of the figure, and singling out the effects of color, scale, and shape biases, biasing only the sender has similar effects as biasing both agents. For each of these biases, the language is grounded largely in the corresponding attribute. Still, the color bias of the DEFAULT receiver leads to an increase in color effectiveness when the sender itself does not have a color bias. Comparing (C) to (B), also a receiver bias is carried over into the emergent language, even though its influence is weaker and the color bias of the DEFAULT sender dominates. We calculate the mean absolute difference (MAD) between the average effectiveness scores in (B) and (A), as well as (C) and (A), for COLOR, SCALE, and SHAPE condition, to quantify the relative influence of biasing one versus both agents. The imbalance between sender and receiver bias is reflected in a higher MAD for biased receivers (0.194) than

biased senders (0.103). Looking at the ALL condition, an interesting pattern emerges. If both agents have an ALL CNN as in (A), the message information is more evenly distributed across all attributes than in the DEFAULT condition. However, if either of the agents uses a DEFAULT CNN, as in (B) or (C), this effect is reversed and the messages are mostly grounded in color, which is likely because the "flexible" ALL agent adapts to the inherent color bias of the DEFAULT agent. In line with this interpretation, the MAD between average effectiveness scores in ALL condition and DEFAULT condition is very small, both when the sender is biased (0.012) and when the receiver is biased (0.013). In sum, perceptual biases of both sender and receiver are reflected in the emergent language, but due to the asymmetry of communication, the sender bias is more influential. Further, agents that rely strongly on all conceptually relevant object dimensions for perceptual categorization can flexibly adapt their language to suit communication partners with more narrow perceptual discrimination abilities.

Perception of relevant similarity relationships improves communication. Table 4.2 displays the training rewards, test rewards, and average effectiveness across attributes for all five conditions (sender and receiver biased). Results for pairing biased with DEFAULT agents can be found in Appendix C.5. The mean test rewards range between 0.914–0.968 across all conditions, at a chance level of 0.33. We are particularly interested in the ALL versus DEFAULT comparison, so whether sharpening the agents' perception with respect to conceptually relevant dimensions improves emergent communication in comparison to default processing. According to all three metrics, ALL agents achieve the best values, and DEFAULT agents the second-best values. The strong perceptual bias for individual attributes seems to bias the communication to a degree that is harmful to performance. Still, the differences between ALL and DEFAULT are significant based on the bootstrapped 95% CIs for the difference in means with respect to training rewards ($[0.007, 0.017]$), test rewards ($[0.005, 0.014]$), and average effectiveness ($[0.040, 0.083]$). The higher average effectiveness in the ALL condition suggests that enforcing conceptually relevant similarities helps the agents to overcome categorization biases, such that they can better communicate all relevant attributes—instead of forming semantic categories based on individual attributes—and as a consequence achieve higher performance.

Table 4.2: Training rewards, test rewards, and average effectiveness across attributes for sender-receiver pairs with the same bias. Reported are means and bootstrapped 95% CIs calculated from twenty runs per condition. The best values across conditions are highlighted.

	default	color	scale	shape	all
train reward	0.956 ± 0.003	0.928 ± 0.008	0.910 ± 0.006	0.937 ± 0.008	**0.968** ± 0.004
test reward	0.959 ± 0.003	0.929 ± 0.009	0.914 ± 0.007	0.939 ± 0.008	**0.968** ± 0.004
$\overline{E(O_a, M)}$	0.676 ± 0.013	0.596 ± 0.015	0.532 ± 0.016	0.600 ± 0.020	**0.738** ± 0.017

4.6.3 Influence of language on perception

To study the influence of different linguistic biases on visual perception, we considered a language learning and a language emergence scenario. For the language learning scenario, we

used the trained senders from the agent pairs above (where both agents have the same bias) and trained DEFAULT receivers to learn their language. For the language emergence scenario, we ran experiments combining a DEFAULT receiver with each type of sender, and combining a DEFAULT sender with each type of receiver. We conducted ten runs per scenario and agent combination, with mean test rewards ranging between 0.919–0.973 (for details about training and test rewards see Appendix C.6).

Linguistic biases influence perception. In the language learning scenario, the language was fixed and learned by the receiver. Fig 4.7, top left, shows that the linguistic biases clearly influence the agent's perception: if message content is biased towards a specific attribute—as in the DEFAULT (color attribute), COLOR, SCALE, and SHAPE condition—the agent learns to better represent visual differences for this attribute. As the DEFAULT receiver starts out with a perceptual color bias (see Table 4.1), changes in visual perception are most clearly visible in the SCALE and SHAPE conditions, where the color bias is reduced, and scale or shape bias increases. Looking at the RSA scores between the sender's and the receiver's visual object representations (Fig 4.7, bottom left) we find that unless both agents start out with a color bias (DEFAULT and COLOR condition) the scores increase, so the receiver's representations adapt to those of the sender. The center and right columns of Fig 4.7 visualize the same analysis results for the language emergence scenario, once for a DEFAULT receiver paired with senders from different conditions (center), as well as for a DEFAULT sender paired with receivers from different conditions (right). The exact same qualitative patterns as in the language learning scenario emerge, with differences in amplitude suggesting that the receiver is more affected by the sender's bias than vice versa. The agents' biases are passed on through language, even if there is no fixed linguistic protocol to begin with.

Communication can improve perception of relevant similarity relationships. Color, scale, and shape information is relevant for the communication game. Therefore, it seems plausible that playing the game could improve visual object representations with respect to these attributes. Fig 4.8 shows the RSA scores of a DEFAULT agent after training in the language learning scenario (left), and the language emergence scenario as receiver (center) or sender (right). The CNN type of the communication partner is color-coded. Indeed, compared to the original RSA score, regardless of the scenario and the bias of the communication partner, the CNN of the DEFAULT agent better accounts for differences in the conceptually relevant attributes. The representational grouping of objects based on the inherent CNN color bias is reduced by playing the communication game.

We further analyzed the influence of scenario (learning, emergence - DEFAULT receiver, emergence - DEFAULT sender) and communication partner bias (DEFAULT, COLOR, SCALE, SHAPE, ALL) by looking at the bootstrapped 95% CIs for the differences in means. Mean RSA scores are lowest in the learning scenario ($M = 0.518$). They are higher in the emergence scenario with a DEFAULT receiver ($M = 0.543$), with a CI of $[0.017, 0.033]$, and even higher for the emergence

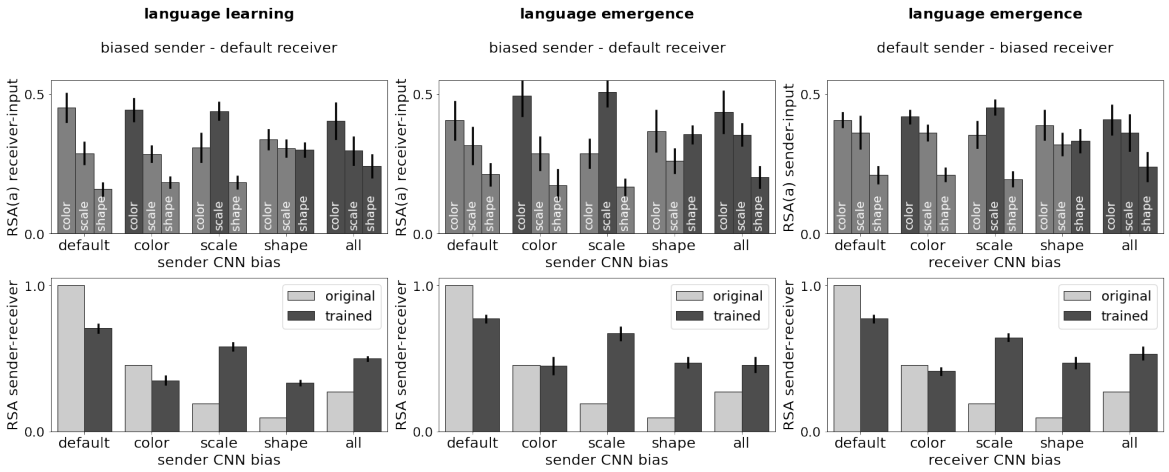


Figure 4.7: Influence of linguistic biases on perception. Shown are the effects of language learning and language emergence on a DEFAULT agent, when paired with agents of different visual bias conditions. The left column covers the language learning scenario with a DEFAULT receiver, the central column the language emergence scenario with a DEFAULT receiver, and the right column the language emergence scenario with a DEFAULT sender. In the language learning scenario, the sender's weights (and therefore also the language) are entirely fixed. In the language emergence scenario, both agents are trained and the language emerges. The visual bias of the communication partner is shown on the x-axis. The top row shows the RSA scores between the DEFAULT agent's visual representations and each object attribute—indicated by the bar label—after training. Attributes that were enforced to create the visual bias of the communication partner are dark gray. The bottom row shows the RSA scores between the visual representations of the DEFAULT agent and those of its communication partner before (light gray) and after (dark gray) training. Reported are means and bootstrapped 95% CIs of ten runs each.

scenario with a DEFAULT sender, with a CI for the two emergence scenarios of $[0.014, 0.033]$. Agents in the language emergence scenarios learn object representations that better reflect the underlying object structure compared to agents in the language learning scenario, with a stronger effect for the sender than the receiver. Thus, it is beneficial, if both agents can adapt their perceptual processes to the game. As the sender dominates the emerging protocol (see above), its visual representations might adapt more strongly to the task. With respect to differences in communication partner bias, we were particularly interested in which communication partners can increase the RSA score compared to a DEFAULT partner ($M = 0.525$ across scenarios). In pairwise comparisons with the DEFAULT partner, a partner with a SHAPE bias leads to the strongest improvement ($M = 0.558, CI = [0.017, 0.047]$), followed by ALL ($M = 0.552, CI = [0.014, 0.040]$), then SCALE ($M = 0.543, CI = [0.005, 0.030]$), and finally COLOR does not seem to yield a significant improvement ($M = 0.535$, $CI = [-0.003, 0.022]$). The DEFAULT agent is good at representing differences in object colors, and bad at representing differences in both scale and shape information, with the largest deficit for shape (see Table 4.1). It seems that talking to SHAPE or ALL agents, which are good at representing shape information, can help overcome the shape deficit, therefore leading to the strongest improvements. Similarly, communication with a COLOR agent does not stimulate the agent to adapt its representations, as the preferred structure based on color values is mutual.

Overall, adapting visual perception for a downstream communication task (while staying true to the original classification objective) improves the visual representation of task-relevant aspects of the environment—in our case the three object-defining attributes. The improvement

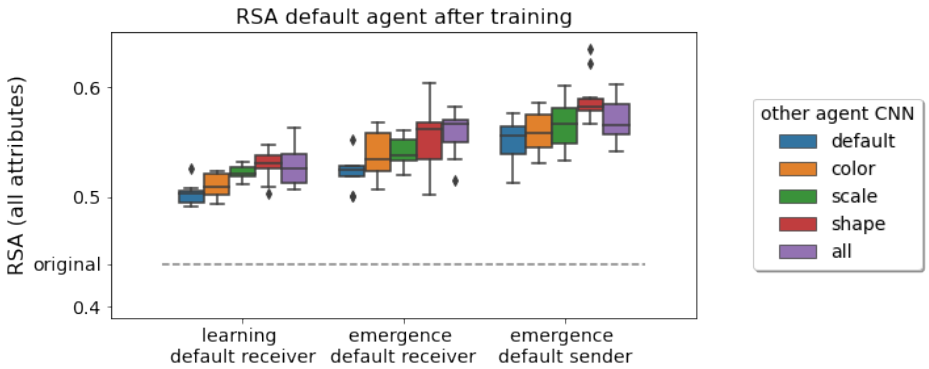


Figure 4.8: RSA scores between symbolic object representations (*k*-hot attribute vectors) and neural object representations in the agent's vision module. Shown are the scores for the DEFAULT agent after training, for different communication partners, and across ten runs each. For the language learning scenario, the DEFAULT receiver is shown (left). For the language emergence scenario, the DEFAULT receiver (left) and the DEFAULT sender (right) are shown. The dashed line indicates the RSA score of the DEFAULT CNN—so the agent's vision module—before training.

is stronger if the communication partner is good at representing aspects for which the agent has a deficit.

The role of classification. The agents' vision modules are trained for classification and communication at the same time. The classification task is used to simulate that the visual representations have other purposes apart from informing communication. We ran additional control simulations without the classification task, to understand its influence on the results above. A detailed description of methods and results can be found in Appendix C.7. The main finding can be confirmed also without classification: If message content is biased towards a specific attribute—because it is predetermined (language learning) or arises through a visual bias of the communication partner (language emergence)—the DEFAULT agent learns to better represent visual differences for this attribute. Still, the classification loss has a moderating effect on the RSA scores as it constrains the visual representations to capture differences between the values of all attributes regardless of linguistic bias. In other words, it keeps the vision module from only representing information that is relevant to the communication game. As the agents discriminate between fewer objects in communication than in classification (communication is less optimal than classification), playing the reference game does not improve the visual representations, i.e. the general RSA score, without the classification loss.

4.6.4 Evolutionary analysis

In the preceding analyses we studied how perceptual biases, or more generally representations, are affected by language use. Here, we take this idea to an extreme by analyzing whether specific perceptual representations (biases) are more likely to result from within- or cross-generational adaptation processes based on their aptitude for communication. For this purpose, we use the static solution concept of evolutionary stability from evolutionary game theory (Maynard Smith, 1974). This solution concept assumes a large, homogeneous population where agents are

randomly paired to play a game of interest. Based on the reward (or payoff) structure between different types of agents, it can be decided whether a population of a certain type can be invaded by an alternative type. In a two-player symmetric game, type t is evolutionary stable, if agents of any mutant type t' achieve less reward playing with an agent of type t than two agents of type t playing with each other, $r(t,t) > r(t',t)$. If there is a competing type t', such that $r(t',t) = r(t,t)$, t is still evolutionary stable if $r(t,t') > r(t',t')$.

While the concept of an ESS has first been introduced in the context of biological evolution, it is useful also for analyzing the stable rest points of non-biological evolutionary optimization processes The latter is made possible by the fact that ESSs are the (locally) asymptotically stable rest points of the replicator dynamic (Hofbauer & Sigmund, 1998; Taylor & Jonker, 1978). The replicator dynamic, in turn, is a rather encompassing high-level formalization of a wide variety of agent-internal optimization processes, be they cross-generational as in cultural evolution or (asexual) reproduction (Sandholm, 2010), or within-generational as in imitation-based dynamics (Franke & Correia, 2018; Sandholm, 2010) or simple forms of reinforcement learning (Börgers & Sarin, 1997).

Enhanced perception of relevant features is evolutionary stable. In our case, the game of interest is the reference game, and the different types are given by different perceptual biases. We assume that agents in the population can act as both sender and receiver. Accordingly, the rewards for two communicating agents with biases t and t' are calculated by averaging the rewards of a t-sender paired with a t'-receiver and a t'-sender paired with a t-receiver. This is also known as *symmetrizing* the game (Cressman, 2003, Section 3.4). Because the training process and the agents' policies are stochastic, the reward for an interaction between two bias types is approximated by averaging across multiple runs. Fig 4.9.A shows the reward matrix for all bias combinations averaged across twenty simulations for each sender-receiver pair. Judging from the average rewards, the DEFAULT and ALL conditions form the only evolutionary stable biases. Pairwise comparisons between the CIs in each matrix column reveal that only the evolutionary stability of the ALL bias is significant. Thus, only the ALL bias prevails in an optimization process for communicative success.

Eliminating potential confounds of task-relevance as evolutionary drive. ALL agents achieve higher rewards than other agents. Intuitively, this is the case because the ALL condition enforces task-relevant attributes. If object color was not relevant to the game, enforcing color similarities should not increase performance, and a color bias should not evolve. However, the advantage of ALL agents could be due to other factors. We noted above that, based on the nature of the reference game, the conceptually relevant (i.e. class-defining) attributes correspond to the attributes that are relevant for successful communication. To achieve perfect performance, all conceptually relevant attributes must be communicated, such that the receiver can identify the target unambiguously against different distractors. ALL agents could therefore achieve higher performance because they are biased towards class-defining attributes rather than task-relevant

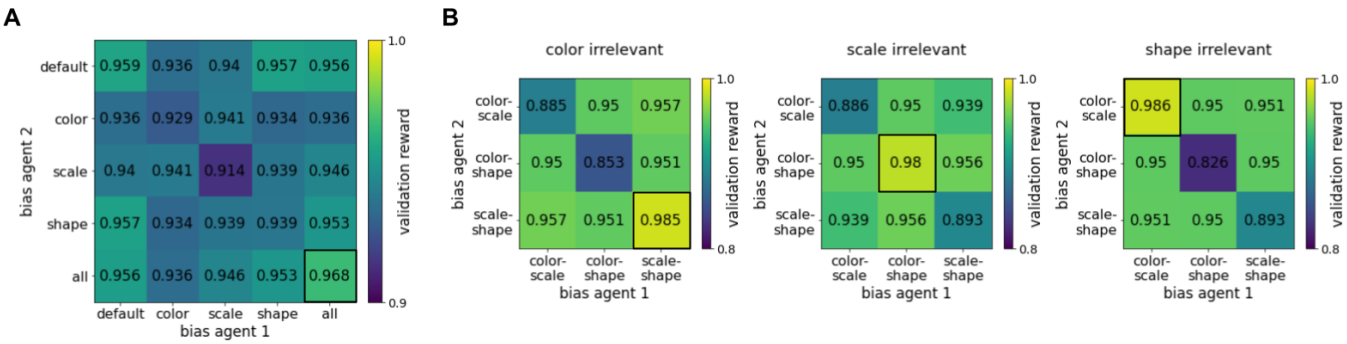


Figure 4.9: Mean reward on the test set for two agents of different bias types communicating with each other. For each sender-receiver combination, we ran twenty simulations. To obtain the average reward for an agent of bias type t' communicating with an agent of bias type t, we average the rewards of the combinations t'-sender/t-receiver and t-sender/t'-receiver, hence the matrices are symmetric. We highlight the results for the combinations where both agents are biased towards all relevant attributes. (A) shows the mean test rewards for agents with $t', t \in \{$DEFAULT, COLOR, SCALE, SHAPE, ALL$\}$ in the basic reference game where all attributes (color, scale, shape) are relevant. (B) shows the mean test rewards for agents with mixed biases $t', t \in \{$COLOR-SCALE, COLOR-SHAPE, SCALE-SHAPE$\}$ for reference games where out of the three attributes either color (left), scale (center), or shape (right) is not relevant.

attributes; or, simply because more attributes are enforced than in the other conditions, which might improve representational structure.

To exclude these alternative explanations, we ran a set of control simulations. We created different mixed-bias conditions, where similarities for two out of three attributes were enforced during perception-pretraining (COLOR-SCALE, COLOR-SHAPE, SCALE-SHAPE). To ensure that the bias strength for enforced attributes is high and approximately equal within and across types, as well as that the bias strength for unenforced attributes is approximately zero, we conducted a grid search across different smoothing factors and weightings between the two enforced biases (for details see Appendix C.8). In addition, we designed reference game variants, where always one of the three object attributes is not relevant (color irrelevant, scale irrelevant, shape irrelevant). E.g., if object color is irrelevant, sender and receiver target may have different colors and still yield maximal reward, while scale and shape must be the same, see Fig 4.10. By training combinations of mixed-bias agents on these games, the set of attributes relevant to pretraining is disentangled from the set of attributes relevant to communication, while the number of enforced biases is constant across agent types.

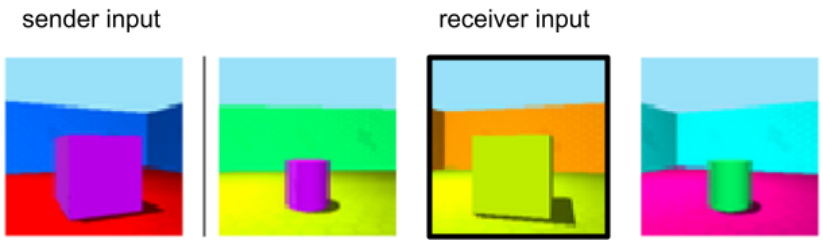


Figure 4.10: Example inputs if object color is irrelevant in the communication game. The receiver target is marked by a black box. Appendix C.9 shows examples of sender and receiver inputs for each game variant (color irrelevant, scale irrelevant, shape irrelevant).

Fig 4.9.B shows the resulting reward matrices (for an analysis of the linguistic biases see Appendix C.10). In each game variant, agent types with a bias for task-relevant attributes form the only

evolutionary stable population. Particularly low performances arise when both agents have the same mismatching bias (low values on the diagonal) because, in that case, the agents' bias does not encourage communication about the respective "missing" attribute. E.g., if both agents have a COLOR-SCALE bias, introducing shape information into the conversation is more difficult than if one agent has a COLOR-SHAPE bias. The matrices further show that representations which are biased towards task-relevant attributes will win against any alternative homogeneous bias. In conclusion, there might be optimization pressure towards representations that accurately capture the relationships between objects, in terms of features that are environmentally relevant.

4.7 Discussion

We proposed that communication games with deep neural network agents can be used to study interactions between perception and emergent communication. Based on systematic manipulations of visual representations and communication protocols, we made the following main observations: 1) biases in either modality are reflected in the other, 2) communication improves the perception of task-relevant attributes, and 3) enforcing accurate representation of task-relevant attributes improves communication—to a degree that specialization of the perceptual system to the linguistic environment could accrue.

Multi-agent communication games account for the interactive and grounded nature of communication. Reinforcement learning (RL) presents a natural framework for modeling learning in these games. Utterances are treated like actions: they are grounded in the environment and driven by objectives. Machine learning models trained on language in isolation—typically under (self-)supervision—have achieved impressive results on various natural language processing tasks by capturing statistical patterns from large corpora (Brown et al., 2020; Devlin et al., 2019; Radford et al., 2019). However, lacking a grounded shared experience, these models cannot address deeper questions about communication and meaning (Bisk et al., 2020).

4.7.1 Influence of perception on language

The first set of analyses investigated the influence of visual perception on emergent communication. We found that semantic category formation was largely shaped by perceptual similarity relationships. In human cognition, the idea that many concepts are characterized by perceptual properties is uncontroversial. For example, objects that are grouped under the same psychological concept often have similar shapes (Rosch et al., 1976). The conceptual structure of the world in our reference game is predetermined: objects are defined by color, scale, and shape, each being equally important. Still, the agents group together several concepts under a single label based on perceptual similarity, which means the emerging protocol is suboptimal. They even do so when the message space and the number of distractors are increased (see Appendix C.4). Recently, it was shown that neural network agents playing a color discrimination game develop *efficient* communication, in the sense that they reach maximum accuracy for a

given language complexity, and that—as in human color-naming systems—low complexity is preferred (Chaabouni et al., 2021). We assume a similar effect in our simulations. The agents develop accurate but simple protocols, and reductions in complexity are achieved by grouping different objects under the same label based on perceptual similarity. We further showed that increasing the perceptual sensitivity for features that are relevant to the communication game debiases communication and improves performance. In line with the above interpretation, it could be that agents with better adapted representational spaces find solutions with higher complexity and accuracy, while still optimizing the trade-off between the two.

These results are also relevant from an engineering perspective. A lot of the existing research in language emergence is focused on developing setups that foster the emergence of communication protocols sharing desirable properties with natural language. The role of how agents perceive and represent the world is mostly ignored (Bouchacourt & Baroni, 2018). However, we not only show that perceptual biases directly influence the emerging protocol but also that they are present in default setups. We find that the organization of pixel inputs into dedicated color channels makes color information more easily accessible than other object information, which leads to a color bias in communication. Neural networks process visual information differently from humans in many ways. For example, they are susceptible to adversarial attacks (Szegedy et al., 2014) and lack useful learning mechanisms observed in children (Gandhi & Lake, 2020). We think that language emergence research can profit from taking into account the effects of differences between human and machine perception. Moreover, we show that agents' performance can be improved by developing representational similarity relationships that are based on task-relevant dimensions, rather than using out-of-the-box pretrained networks.

4.7.2 Influence of language on perception

The second set of analyses studied the influence of (emergent) communication on visual perception. We found that categories established by the communication protocol modulate representational similarities to better reflect this categorical structure, by increasing the similarity between objects that are grouped together under the same expression. It has been shown that learning new color categories (through a perceptual task) induces categorical effects on color discrimination similar to those of natural color categories (Ozgen & Davies, 2002). These results suggest that cross-language differences in perceptual representations may arise as a result of learning linguistic categories, as simulated in our experiments. Besides, we observed that perceptual sensitivity increases for features that are relevant in the communication game and therefore affect the agents' objective. The need to discriminate between features, for communication to be successful, can disentangle their visual representations. This increase in sensitivity occurs even though the exact same features are also relevant in the pretraining classification task. A related effect has been observed in a visual search task. Although there is a baseline effect of conceptual categories on visual processing, this effect increases if the target category is labeled (Lupyan, 2008).

Both these observations have been made in earlier simulations. Harnad et al. (1991) showed that neural networks trained on a supervised classification task show effects of categorical perception, in that a continuous input dimension is warped in the network representations to increase within-category similarity and decrease between-category similarity. Later, Cangelosi and Harnad (2000) compared agents that learned categories from sensorimotor interaction with the world ("sensorimotor toil") to agents that could additionally learn from communication signals ("symbolic theft"). Sensorimotor interaction, comparable to our pretraining classification task, warped the agents' representational similarity space but supervised learning of symbolic object descriptions warped these similarity spaces even further, leading to increasingly categorical perception. Our work extends these computational approaches. We model how a representation space can restructure itself to reflect a categorical partition of a comparatively complex input space, based on communicative interaction rather than supervised learning.

Modeling a communication scenario has the advantage that we can study interactions between communication partners who conceptualize the world differently. Because the emerging language is shaped by the perceptual biases of both agents, and in turn shapes their perceptual biases, the agents' representations become aligned through communication. Comparable effects have been found in empirical studies. Category structure aligns between people who play a reference game (Markman & Makin, 1998), and more generally between people who assign novel labels to stimuli with the goal to coordinate (Suffill et al., 2019).

These analyses, too, have implications for engineering-driven research. Backpropagating the learning signal from the communication game through the vision module of the agents improves their ability to represent and discriminate between relevant features, which might be useful for downstream tasks other than communication. It also provides a way to align perceptual representations of different agents, which can be particularly useful if one agent can thereby correct specific perceptual deficits of the other agent.

4.7.3 Evolutionary analysis

Finally, the evolutionary analysis showed that accurate perception of environmentally relevant aspects constitutes a functional advantage. Related results have been found in experiments with robots playing a color naming game (Bleys et al., 2009). Robots that could adapt their categories to the task performed better than robots starting out with the same, but fixed category structure. Most likely, representational structure in humans is optimized to accommodate environmentally relevant conceptualizations as well (Gärdenfors, 2004; Marstaller et al., 2013). In our simulations, communication was the only task performed by the agents. Representational structure in humans, however, is shaped by various environmental pressures. Our results do not indicate that perception only adapts to optimize communication, but rather that communication (as a means to exchange information about relevant aspects of the environment) may constitute one of these pressures.

Whether language could have influenced the brain, and therefore also visual perception, through biological evolution is highly debated. A major problem lies in the fact that it is difficult to estimate the relative change of perception during the evolution of language. The (macaque) monkey visual system is often and successfully taken as a model system for the human visual system. A mainstream view is that the two visual systems share many characteristics but are not identical (Orban et al., 2004; Rapan et al., 2022). Furthermore, it is uncertain when language emerged (Hauser et al., 2014). However, it has been argued that—evolutionarily young and variable—language is rather shaped by the—evolutionarily old and stable—brain than vice versa (Christiansen & Chater, 2008). While we abstain from claims about the time scales of the analyzed optimization process, it seems more likely that language-guided adaptations of visual representations happen within the lifetime of an individual.

Stable state analysis is a static solution approach to evolutionary games. It can identify whether a given population will remain at a certain state but does not explain how a population arrives at that state. The latter question can be answered by dynamic approaches, which apply an explicit model of the optimization process. A prominent example is the *replicator dynamic*, originally defined for a single species by Taylor and Jonker (Taylor & Jonker, 1978) and named by Schuster and Sigmund (Schuster & Sigmund, 1983). Thus, evaluating the probability that a randomly initialized population develops perceptual representations that match communicative needs would require the use of dynamic models.

4.7.4 Flexible-role agents and populations

In the original Lewis game, there are two agents with fixed roles (sender and receiver), two world states, and two actions. In the theoretical analysis of signaling games, it has been of general interest how the agents' behavior changes under variations of this simple case (Skyrms, 2010). Like the original Lewis game, our reference game involves two agents with fixed roles. To make sure that our results do not only pertain to this special case, we ran additional simulations with more agents and flexible-role agents. In particular, we separately tested an extension to flexible-role agents and an extension to a 4-agent game (two senders, two receivers). We repeated the analyses above for the DEFAULT, SCALE, and ALL conditions, as these conditions cover the main manipulations of enforcing no bias, a bias for a single attribute, or a bias for all attributes. Details about methods and results can be found in Appendix C.11 (4 agents) and Appendix C.12 (flexible-role agents). At least for these two extensions, we can establish the same main results as for the fixed-role, 2-agent game. While many more variations are conceivable, our findings seem to reflect general aspects of language-perception interactions in multi-agent communication.

4.7.5 Limitations

Combining communication games with deep learning to study interactions between language and perception (and possibly other areas of cognition) is a novel approach. As a first implemen-

tation, the proposed setup tries to strike a balance between the flexibility of modern DNNs and experimental control. Our images and categories fall clearly short of the visual complexity of the world. However, using objects that are composed of a fixed set of attributes and attribute values has several advantages. We can introduce selective visual biases via relational label smoothing, and we can quantify and compare visual and linguistic biases with respect to these attributes.

Our model also greatly simplifies the functionality of visual perception. Our agents use their vision modules to generate representations that can be used for classification and communication. The visual brain, in contrast, performs a multitude of functions each of which imposes organizational and representational constraints. In particular, visual perception requires an (implicit) understanding of sensorimotor contingencies as it informs and is informed by motor action (Noë, 2004). Hence, unlike our model, the visual system continues to represent information that is irrelevant to categorization or communication. As a consequence, our results likely overestimate the effects of language on perception. In addition, without a significant increase in architectural and functional complexity, an analysis of the penetration depth of language into visual representations (high-level attentional selection mechanisms vs. dynamic re-tuning of receptive fields of primary sensory neurons) does not warrant conclusions about the human visual system. Empirical studies show that the effects of language on vision are dynamic and task-dependent. For example, in color discrimination tasks, categorical effects are observed for naive but not trained observers (Witzel & Gegenfurtner, 2015), and sometimes only in the presence but not in the absence of verbal cues (Forder & Lupyan, 2019). Future work could study these more nuanced effects by using more complex vision modules.

4.7.6 Outlook

Our vision modules are CNNs trained on classification. Thereby, they rely on the same principles—albeit being much simpler—as state of the art models of vision (Lindsay, 2021; Storrs et al., 2021). Still, there are many ideas on how correspondence between artificial and biological neural networks can be further improved by changing architectures, learning algorithms, input statistics, or training objectives (Kietzmann, McClure, et al., 2019; Richards et al., 2019). As a relatively minor change, training on superordinate or both superordinate and basic labels, rather than on subordinate labels as is typically the case, makes visual representations more robust and more human-like (Ahn et al., 2021). Note that information about taxonomic relationships can also be encoded in the training labels directly using the (hierarchical) relational label smoothing method presented here. An example of an architectural change are recurrent CNNs, which include not only bottom-up but also lateral and top-down connections. Including recurrence improves object recognition, especially under challenging conditions (Spoerer et al., 2017), and is required to model the representational dynamics of the visual system (Kietzmann, Spoerer, et al., 2019). As an example of a change in objective, an *embodied* DNN agent has been shown to learn sparse and interpretable representations through interactions with its simulated environment (Clay et al., 2021). In addition to scaling our experiments to more complex input data and deeper networks, future work could draw on these exciting developments to better capture the

functional and architectural constraints on the visual system. The resulting models could be used to investigate how the effect of communication on perceptual representations changes under these additional constraints.

This paper set out to explore mutual influences between language and (visual) perception in multi-agent communication. But language interfaces with other areas of human cognition as well. The embedding of language in general cognition is evident in everyday language use. For instance, in understanding a written text, we are able to recruit from memory the right background assumptions to make the text coherent (Graesser et al., 2001). This can, among others, be observed in bridging inferences. Upon reading "They had a barbecue. The beer was warm.", we can conclude that the beer was part of the barbecue. Another salient example is attention. While we may share a basic attention mechanism for dealing with the non-linguistic world, having a language to "bridge minds" will likely lead to fine-tuning and, in fact, align our attentional mechanisms. Think about saying "Wow!" or adding "surprisingly". These so-called mirative markers convey surprise (Delancey, 1997), thereby telling the audience what we expected, but also what we pay attention to. Essentially, every statement about the world conveys meta-information about what the speaker finds newsworthy in the first place. On a basic level, also the role of attention or memory could be studied with our setup, for example by using neural network agents with attention mechanisms (Chaudhari et al., 2021) or external memory (Graves et al., 2016). In general, due to the versatility of both deep learning architectures and communication games, their combination forms an excellent testbed for various language-related interface problems.

Our experiments go beyond analyzing effects *on* emergent communication. They also account for the reverse direction, i.e. how language shapes other domains. Such Whorfian effects are widespread; apart from visual perception they have, for example, been observed in motion, spatial relations, number, and false belief understanding (Wolff & Holmes, 2011). In fact, it seems likely that all interfaces between cognition and language are mutually adapted towards optimal interaction in the environments we face (Jablonka et al., 2012), such that language can guide the acquisition of cognitive representations from experience, and in turn, can be used to structure and exchange these experiences (Perlovsky, 2009). In a neural network agent, linguistic feedback can be backpropagated into any module that may be considered adaptive to language use. As illustrated by our analyses, language emergence games can address adaptions within and across generations. Future research could use the presented framework to improve our understanding of language in relation to general cognition, from its origins to its cultural and potentially genetic evolution.

Data availability

Materials and code are publicly available at the Open Science Framework (OSF): https://osf.io/qu4xp/.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - GRK 2340. Preliminary results (parts of Section 4.6.1 and Section 4.6.2) were presented at the 43rd Annual Conference of the Cognitive Science Society (Ohmer et al., 2021).

C Appendix

C.1 Entropy analysis between target objects, messages and selections

The schema for a three-way information-theoretic analysis of the relation between target objects O, messages M, and selected objects S, is depicted in Fig 4.11. Quantifying all terms requires generalized definitions of (conditional) mutual information and conditional entropy for three random variables. The mutual information between three variables, also known as interaction information, is defined as

$$I(X, Y, Z) = I(X, Y) - I(X, Y \mid Z)\,,$$

where the conditional mutual information is the expected mutual information between X and Y given Z:

$$I(X, Y \mid Z) = \sum_{z \in Z} \sum_{y \in Y} \sum_{x \in X} p(x, y, z) \log \frac{p(z)p(x, y, z)}{p(x, z)p(y, z)}\,.$$

The conditional entropy of X given Y and Z quantifies the amount of uncertainty that remains about X when knowing Y and Z

$$H(X \mid Y, Z) = -\sum_{z \in Z,\, y \in Y,\, x \in X} p(x, y, z) \log \frac{p(x, y, z)}{p(y, z)}\,.$$

Our analyses show that the mutual information between objects and selections given messages is approximately zero in all experiments, $I(O, S \mid M) \approx 0$. In other words, the shared information between target and selection is fully predicted by the messages. The symmetry between sender (objects-messages) and receiver (messages-selections) analysis can also be identified in this more general framework in terms of the following relationships: $H(O \mid M, S) \approx H(S \mid O, M)$ and $I(O, M \mid S) \approx I(M, S \mid O)$.

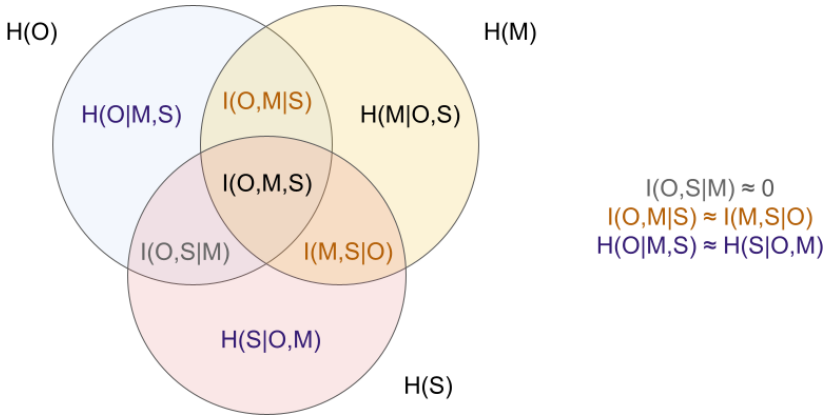


Figure 4.11: Schema of the information in the target objects, O, the corresponding messages, M, and objects selected by the receiver, S. H denotes entropy and I mutual information. Note, the schema is not an actual set-theoretic representation and serves illustrative purposes only.

C.2 T-SNE plots of the visual object representations

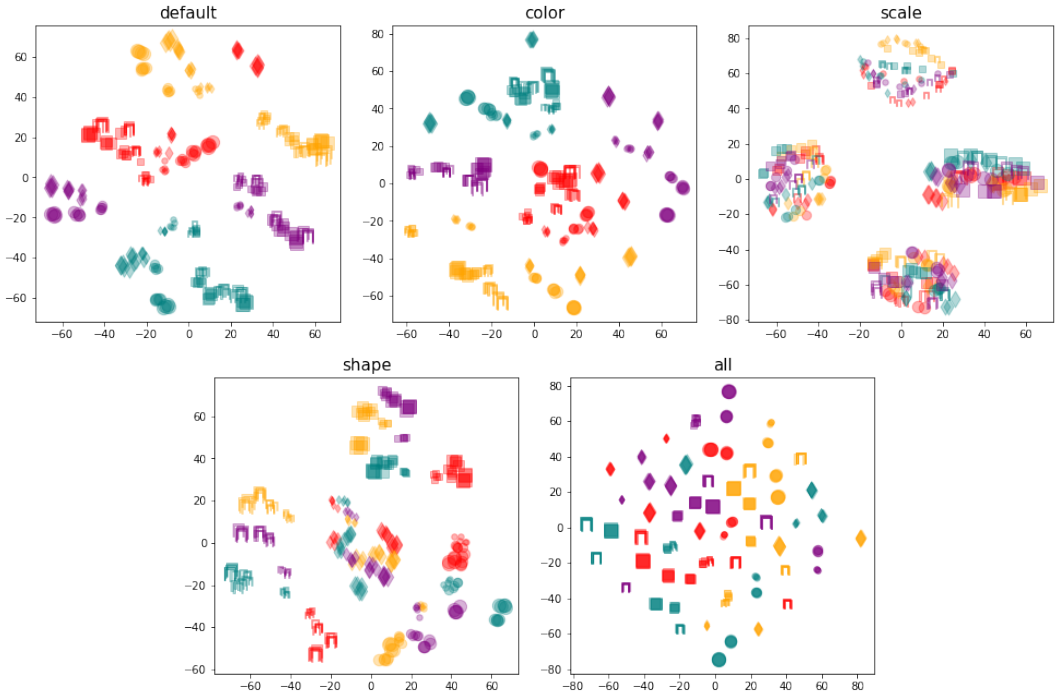


Figure 4.12: Two-dimensional t-SNE plots of the visual object representations in the penultimate CNN layer for each pretraining condition. The four color and scale values are given by the four marker colors and marker sizes, while the following mapping from object shape to marker shape is used: (cube, sphere, cylinder, ellipsoid) → (square, circle, square cap (⊓), rhombus (◊)). t-SNE embeddings were calculated on a data subset of 100 random examples per class (6400 data points) using a perplexity of 100, and 2000 iterations. Plotted are the embeddings for 5 random examples per class. In the DEFAULT and COLOR conditions, clusters form around color values, in the SHAPE condition around shape values, and in the SCALE condition around scale values. The complex similarity relationships in the ALL condition do not fall into clear clusters in two dimensions.

C.3 Representational similarities between object classes

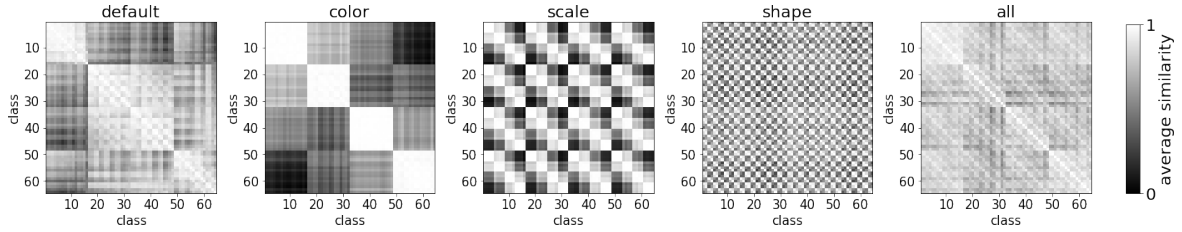


Figure 4.13: Pairwise cosine similarities between object classes in the penultimate CNN layer for each pretraining condition. Average cosine similarities were calculated from 50 random examples per class. Object attributes are structured periodically in the data set. For object class c, color is determined by $(c-1) \mod 16$, scale by $\big((c-1) \mod 16\big)//4$, and shape by $c-1 \mod 4$, where $\mod$ is the modulo operator, and $//$ division without remainder. These periodic patterns are reflected in the similarity matrices. However, the patterns are not perfect as similarities are still influenced by the input topology and not entirely determined by the label distribution.

C.4 Increasing vocabulary size and number of distractors

Fig 4.14 shows the effectiveness scores for different vocabulary sizes and numbers of distractors across ten runs per condition. For $|V| = 4$ (top row) increasing the number of distractors does not increase effectiveness. Given this limited vocabulary size, the communicative content does not improve when more distractors are used. Increasing the vocabulary size to $|V| = 8$ or $|V| = 12$ (center and bottom rows) makes the task easier and allows the agents to find better protocols, which is reflected in higher effectiveness scores (and test rewards, not shown here). Increasing the number of distractors in addition to the vocabulary size (right column) can further increase the average effectiveness for some conditions. Although average effectiveness increases with a larger vocabulary size in the DEFAULT condition, average effectiveness in the ALL condition is still significantly higher for vocab size $|V| = 8$ and $|V| = 12$ and either number of distractors (lower bounds of bootstrapped 95% CIs for differences in means > 0.020); and so are the test rewards (not shown here). So, also when nudged to communicate more information about each attribute, ALL agents develop better protocols than DEFAULT agents.

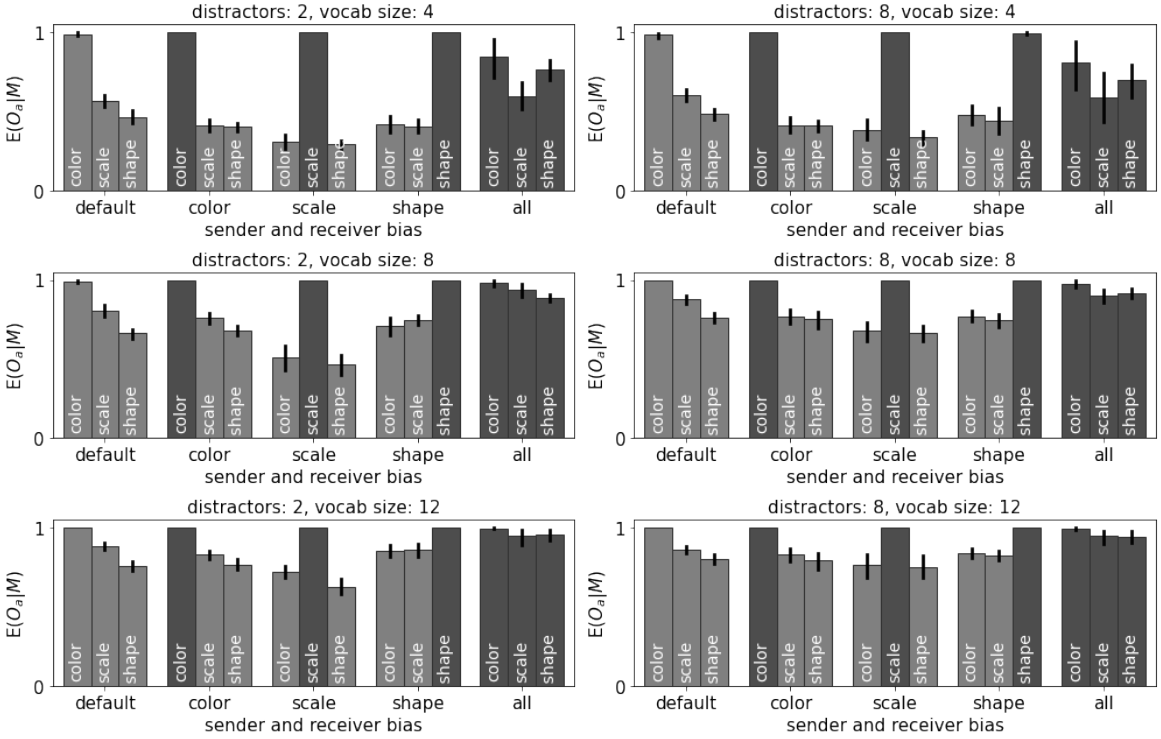


Figure 4.14: Effectiveness per attribute for different vocabulary sizes ($|V| \in \{4, 8, 12\}$), and different numbers of distractors ($k \in \{2, 8\}$). Sender-receiver pairs with the same bias play the reference game, and only the language module weights are trained. The bars are labeled with the attribute a used for calculating $E(O_a|M)$, with attributes enforced via label smoothing in dark gray. We report means and bootstrapped 95% CIs calculated from ten runs each.

C.5 Performance of biased-default agent combinations

Table 4.3: Performance of biased-default agent combinations when only the language modules are trained. Shown are training rewards, test rewards, and average effectiveness across attributes for sender-receiver (S-R) pairs consisting of one biased and one DEFAULT agent. Reported are means and bootstrapped 95% CIs of twenty runs per condition. The best values across conditions are highlighted.

		color	scale	shape	all
S biased, R default	train reward	0.919 ± 0.008	0.914 ± 0.009	0.944 ± 0.006	**0.951** ± 0.005
	test reward	0.922 ± 0.008	0.917 ± 0.009	0.947 ± 0.006	**0.954** ± 0.005
	$\overline{E(O_a, M)}$	0.594 ± 0.015	0.584 ± 0.017	0.656 ± 0.019	**0.688** ± 0.015
R biased, S default	train reward	0.945 ± 0.013	0.959 ± 0.003	**0.965** ± 0.005	0.960 ± 0.003
	test reward	0.947 ± 0.014	0.962 ± 0.003	**0.966** ± 0.005	0.961 ± 0.004
	$\overline{E(O_a, M)}$	0.666 ± 0.020	0.706 ± 0.015	**0.742** ± 0.015	0.689 ± 0.014

C.6 Performance in language learning and language emergence

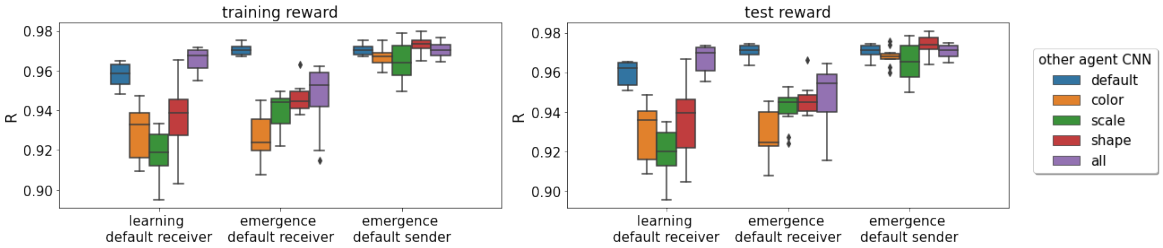


Figure 4.15: Performance on the language learning and language emergence task, when language and vision modules are trained. Shown are boxplots of training and test rewards in the language learning and language emergence scenarios, when studying the influence of differences in language on perception. The plots are generated from the results across ten runs each for communication partners with different perceptual biases (color-coded), always in combination with a DEFAULT agent. In the language learning scenario, the sender (vision and language module) is fixed and we study the effects on the DEFAULT receiver, that is learning the language. In the language emergence scenario, we consider the two cases that a DEFAULT receiver is paired with different senders, and that a DEFAULT sender is paired with different receivers.

C.7 Control simulations without classification loss

In these control simulations, we study the influence of language on perception when the agents are trained on the reference game but not the classification task. We rerun the original simulations without classification loss for the DEFAULT, ALL, and SCALE condition. The latter serves as a representative of the single-attribute bias conditions. The classification loss stabilizes training and allows for a higher learning rate. Without the classification loss, we reduce the learning rate to 0.0001 and increase the number of epochs to 50 in the language learning scenario and 250 in the language emergence scenario. Apart from that, we use the original hyperparameters and training procedure. The average test rewards for the language learning scenario lie between 0.921–0.967 and for the language emergence scenarios between 0.905–0.937.

Fig. 4.16 (top row) shows the visual biases of the DEFAULT agent after training, for communication partners with different visual biases (color-coded). Overall the resulting biases show the same patterns as in the original simulations (see Fig. 4.7). We can confirm the main finding that language influences perception. If the DEFAULT agent communicates with a biased agent (e.g. SCALE), the bias of the communication partner leads to an increase in the RSA score of the corresponding attribute (scale).

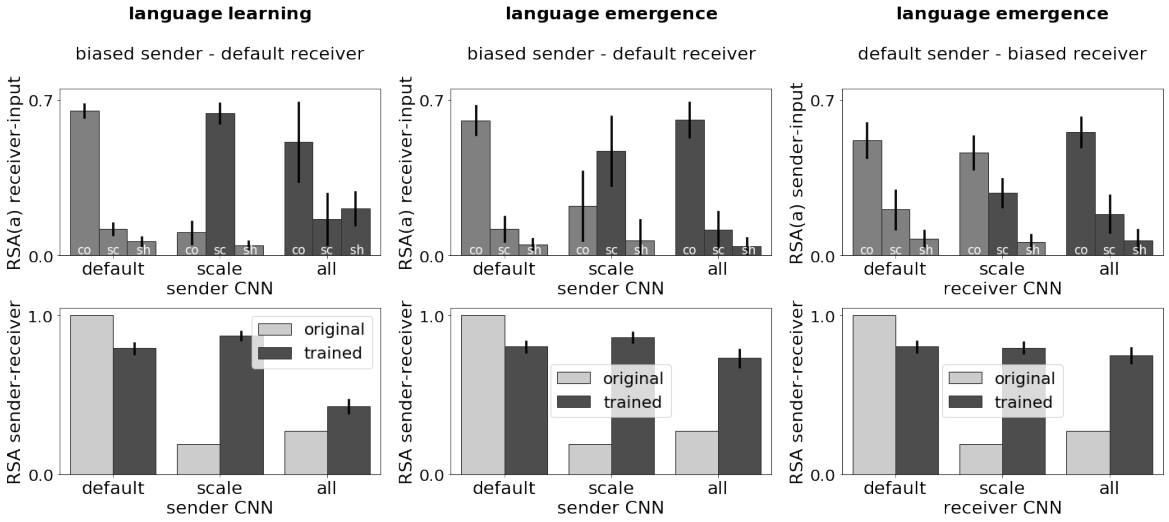


Figure 4.16: Influence of linguistic biases on perception. Shown are the effects of language learning and language emergence on a DEFAULT agent, when paired with agents of different visual bias conditions. The left column covers the language learning scenario with a DEFAULT receiver, the central column the language emergence scenario with a DEFAULT receiver, and the right column the language emergence scenario with a DEFAULT sender. In the language learning scenario, the sender's weights (and therefore also the language) are entirely fixed. In the language emergence scenario, both agents are trained and the language emerges. The visual bias of the communication partner is shown on the x-axis. The top row shows the RSA scores between the DEFAULT agent's visual representations and each object attribute—indicated by the bar label (*co*: color, *sc*: scale, *sh*: shape)—after training. The bottom row shows the RSA scores between the visual representations of the DEFAULT agent and those of its communication partner before (light gray) and after (dark gray) training. Reported are means and bootstrapped 95% CIs of ten runs each.

Comparing the two figures, the classification loss seems to have a moderating effect on the visual representations. Without classification, the linguistic biases are more strongly reflected in the visual biases. In the language learning scenario, this effect can be observed for the DEFAULT sender (color bias) and the SCALE sender. The RSA scores for the biased attribute increase while the other RSA scores decrease. In the language emergence scenario, this effect can be observed from an increase in RSA_{color} in most conditions. One could assume that the agents start out with a strong color bias (DEFAULT agents) and keep that bias because the effect of communication is weaker without classification. However, the language learning scenario shows that this is not the case. Rather, it seems that all agents increasingly focus on color information. The color bias must stem from the input representation or the CNN architecture and not the classification objective. Without classification, the induced biases can revert to a color bias, which then dominates the conversation and as a result also the changes in visual representations. For example, the color bias becomes more prominent in interactions between DEFAULT and ALL agents. At the same time, RSA_{shape} decreases across simulations, as shape information is no longer enforced by the classification task and is not the focus of the DEFAULT or SCALE agent. While shape information does originally play a role for the ALL agent, it is mostly overwritten by color information in the language emergence process. In conclusion, the classification loss constrains the visual representations to also capture differences between the values of attributes that do not conform to the linguistic bias.

As the agents discriminate between fewer objects in communication than in classification (communication is less optimal than classification), the visual representations contain less

information if they only serve communication. Therefore, an increase in the overall RSA scores after training can only be observed if training on the classification task continues.

C.8 Grid search for mixed-bias agents

We conducted a grid search to generate comparable mixed-bias agents. We pretrained CNNs enforcing always two attributes: color and scale, color and shape, or scale and shape. The goal of our search was to identify a network for each condition, such that 1) biases for enforced attributes are strong, 2) biases for enforced attributes are approximately equally strong within and across networks, 3) biases for not enforced attributes are approximately zero, and 4) achieved training accuracies are reasonably high. For the grid search we varied the smoothing factor $\sigma \in \{0.6, 0.7, 0.8\}$, and used different weightings between the two enforced biases $w \in \{[0.05, 0.95], [0.10, 0.90], [0.15, 0.85], \ldots, [0.85, 0.15], [0.90, 0.10], [0.95, 0.05]\}$. We selected a network for each condition (see Table 4.4), by optimizing the first three criteria under the constraint of a minimum training accuracy of 0.97. For each condition, the smoothing factor 0.8 yielded the best network. The weighting parameters show that to obtain these results one must counterbalance the networks' inherent color bias, by using weaker enforcement for color than the other attribute. Biases for enforced attributes lie around 0.45, and biases for other attributes around 0.00.

Table 4.4: Results of the grid search across mixed-bias networks. Each row shows parameters (smoothing factor σ and weighting w), test rewards (test r), and visual biases measured as RSA scores between network representations and attribute templates ($RSA_{attribute}$).

condition	σ	w	test r	RSA_{color}	RSA_{scale}	RSA_{shape}
color-scale	0.8	[0.30, 0.70]	0.996	0.444	0.483	0.000
color-shape	0.8	[0.35, 0.75]	1.000	0.458	-0.002	0.435
scale-shape	0.8	[0.75, 0.25]	0.974	-0.001	0.464	0.430

Simply using a fixed smoothing factor (e.g. $\sigma = 0.6$) and enforcing both relevant traits with equal weight yields the same qualitative but weaker quantitative results in the evolutionary analysis, compared to using the networks obtained from the grid search. Quantitative differences arise due to systematic (inherent color preference) and unsystematic (random seed) imbalances between network biases. For example, in a game where color and shape are relevant, COLOR-SHAPE agents should achieve particularly high rewards. But if a COLOR-SHAPE agent has a very strong color but weak shape bias, and a SCALE-SHAPE agent has a comparatively stronger shape bias, combining the two agents may result in similarly high rewards. The grid search allows us to eliminate such confounding effects.

C.9 Control experiments varying task-relevant attributes

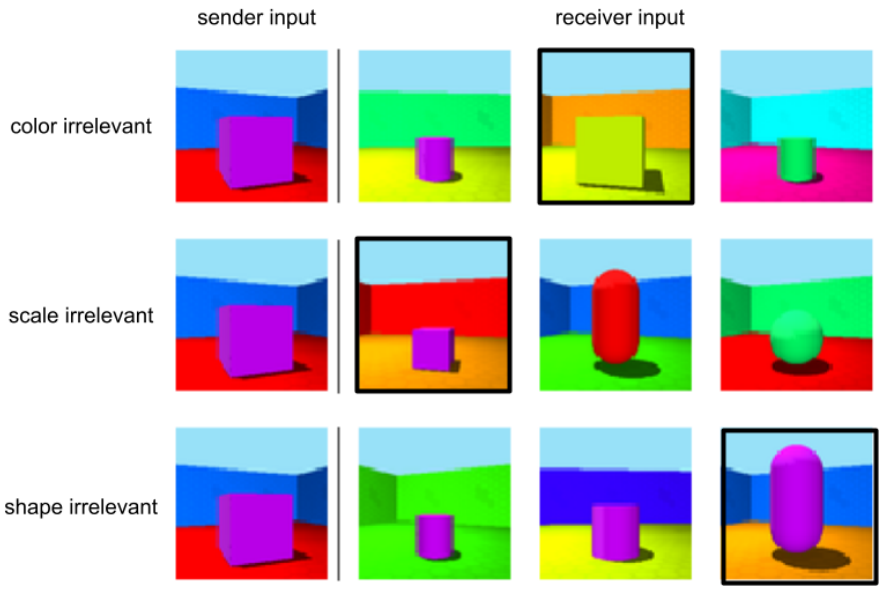


Figure 4.17: Examples of sender and receiver inputs for different relevance conditions. Always two of the attributes color, scale, and shape are relevant, i.e. one attribute is not relevant. For the irrelevant attribute, sender and receiver target may have different values. Shown are example inputs for each variant: color irrelevant (top row), scale irrelevant (middle row), and shape irrelevant (bottom row). The receiver target for each condition is marked by a black box.

C.10 Effectiveness scores for the mixed-bias simulations

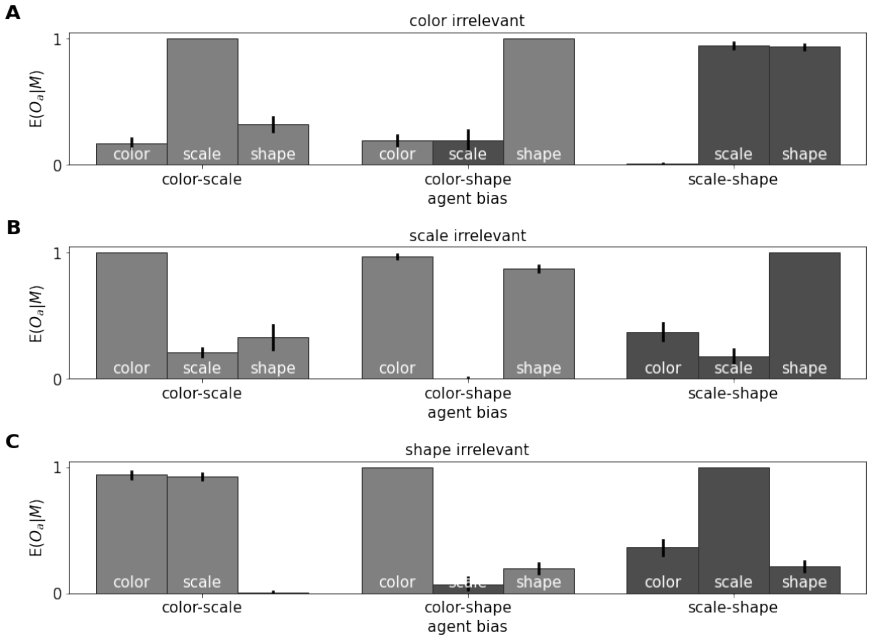


Figure 4.18: Linguistic biases for the mixed-bias control simulations. Shown are the effectiveness scores per attribute when combining a sender and a receiver with the same mixed bias. The agents' bias is given on the x-axis, the score on the y-axis, and the attribute for which the score is calculated is indicated by the bar labels. Bars of enforced attributes are dark gray. Results are shown for the three different relevance conditions: (A) color irrelevant, (B) scale irrelevant, (C) shape irrelevant. We report means and bootstrapped 95% CIs of twenty runs each. Again, the differences in visual perception systematically influence the emerging language. The scores further show that only visual biases for task-relevant attributes are reflected in the language.

C.11 Extension to two senders and two receivers

We test whether our results from the 2-agent setup generalize to a 4-agent setup with two senders and two receivers. We run simulations for the DEFAULT, ALL, and SCALE condition. The latter serves as a representative of the single-attribute bias conditions. The two senders always have the same perceptual bias, and so do the receivers. In general, we use the same architectures, hyperparameters, and training regime as in the original simulations, with the exception that for each batch a sender and a receiver are randomly selected for training. Because convergence speed decreases with the number of agents, we extend the training time to 250 epochs. We rerun each of the three analyses: (i) influence of perception on language, (ii) influence of language on perception, and (iii) evolutionary analysis. The reported values for senders/receivers are obtained by averaging across the two senders/receivers, and the reported values for sender-receiver pairs are obtained by averaging across all sender-receiver pairs. The results for four agents are qualitatively identical to the results with two agents. Hence, we refer the reader to the Results section in the main text for explanations.

(i) For the agents' performance on the test set, please refer to analysis (iii). The effectiveness scores are shown in Fig 4.19, which corresponds to Fig 4.6 of the 2-agent simulations.

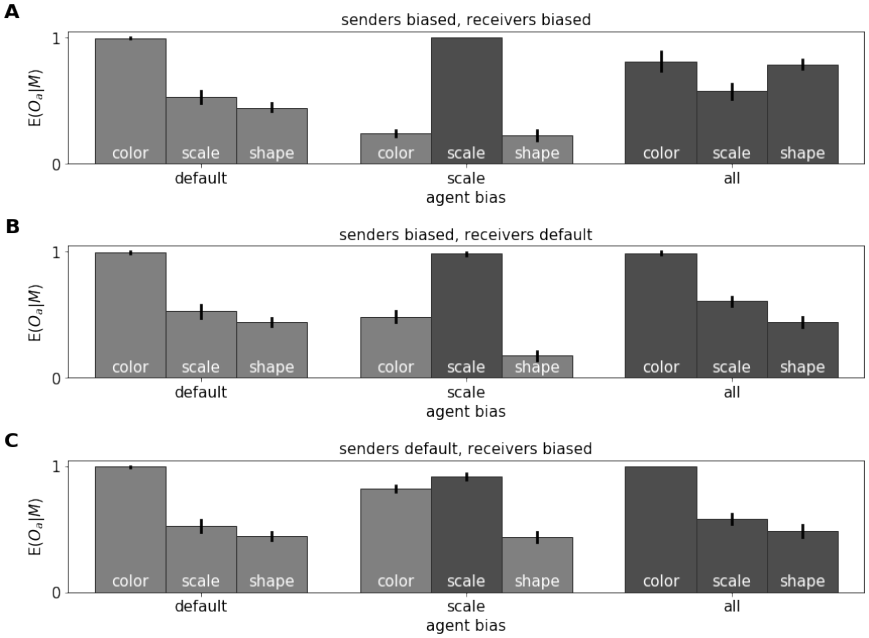


Figure 4.19: Effectiveness per attribute for different combinations of two senders and two receivers. Pairings are (A) senders and receivers with the same perceptual bias, (B) biased senders and DEFAULT receivers, and (C) biased receivers and DEFAULT senders. The x-axis shows the agents' perceptual biases. The bars are labeled with the attribute a used for calculating $E(O_a \mid M)$, with attributes enforced via label smoothing in dark gray. We report means and bootstrapped 95% CIs of ten runs each.

(ii) The language learning scenario does not apply to the 4-agent simulations because it tests the effects of learning a specific language on an individual. Hence, results are reduced to the language emergence scenario. The agents achieve average rewards between 0.927 and 0.966 on the test set. The attribute-wise RSA scores are shown in Fig 4.20, which corresponds to Fig 4.7 of the 2-agent simulations. In analogy to Fig 4.8, we calculate the difference in general RSA scores before and after training. The RSA score of the DEFAULT receiver improves from 0.439 before training to 0.553 (DEFAULT sender), 0.556 (SCALE sender), and 0.595 (ALL sender). The RSA scores of the DEFAULT sender improves from 0.439 before training to 0.574 (DEFAULT receiver), 0.600 (SCALE receiver), and 0.604 (ALL receiver).

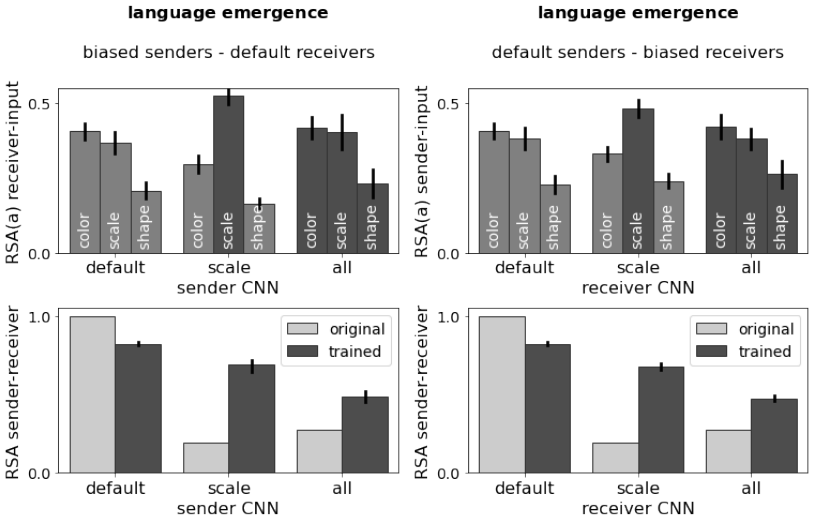


Figure 4.20: Influence of linguistic biases on perception. Shown are the effects of language emergence on the DEFAULT receivers when paired with biased senders (left column) and the effects on the DEFAULT senders when paired with biased receivers (right column). Shown are the effects of language emergence on the DEFAULT receivers when paired with biased senders (left column) and the effects on the DEFAULT senders when paired with biased receivers (right column). The visual bias of the communication partner is shown on the x-axis. The top row shows the RSA scores between the DEFAULT agents' visual representations and each object attribute—indicated by the bar label—after training. The bottom row shows the RSA scores between the visual representations of the DEFAULT agent and those of its communication partner before (light gray) and after (dark gray) training. Reported are means and bootstrapped 95% CIs of ten runs each.

(iii) The payoff matrix for different agent combinations is shown in Fig 4.21, which corresponds to Fig 4.9.A of the 2-agent simulations. Again, the general patterns are comparable. Also in the 4-agent case, pairwise comparisons between the CIs in each matrix column reveal that only the evolutionary stability of the ALL bias is significant.

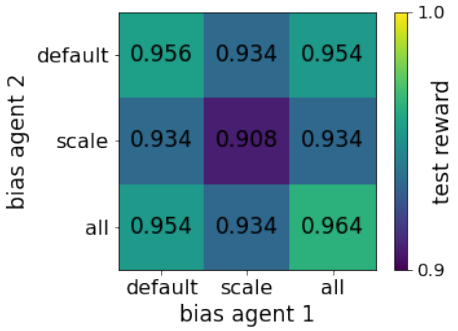


Figure 4.21: Mean reward on the test set for two senders and two receivers of different bias types communicating with each other. For each sender-receiver combination, we ran ten simulations. To obtain the average reward for agents of bias type t' communicating with agents of bias type t, we average the rewards of the combinations t'-sender/t-receiver and t-sender/t'-receiver, hence the matrices are symmetric. Results are shown for the basic reference game where all attributes (color, scale, shape) are relevant.

In sum, across analyses, the findings from simulations with two agents generalize to simulations with four agents.

C.12 Extension to flexible-role agents

We test whether our results generalize from fixed-role agents to flexible-role agents. We run simulations for the DEFAULT, ALL, and SCALE condition. The latter serves as a representative of the single-attribute bias conditions. The sender and the receiver in the original simulations have the same architecture, apart from an additional dense layer in the sender model. Our flexible-role agent therefore uses the same model architecture as the sender. If it is used as a receiver, the additional layer remains unused, and the hidden state of the language module is initialized with a zero vector. The vision module is always used to process the input image(s) and the language module is either used to generate or to interpret a message, depending on the current task of the agent. Note, that this setup does not guarantee that both agents will converge to sending the same messages, and to the best of our knowledge there is no trivial way to enforce such behavior. We use the same hyperparameters and training regime as in the original simulations, with the exception that for each batch one of the agents is randomly assigned the role of sender and the other agent the role of receiver. We rerun each of the three analyses: (i) influence of perception on language, (ii) influence of language on perception, (iii) evolutionary analysis. Across all analyses, simulations with flexible-role agents yield the same results as simulations with fixed-role agents. Hence, we refer the reader to the Results section in the main text for explanations.

(i) For the agents' performance on the test set, please refer to analysis (iii). The effectiveness scores are shown in Fig 4.22, which corresponds to Fig 4.6 of the fixed-role agent simulations. In part A, the effectiveness scores are averaged across both (biased) agents for each run. Parts B and C show the results for the combination of one biased and one DEFAULT agent. As mentioned above, the agents do not necessarily speak the same language, hence we analyze the effectiveness scores for the biased agent (B) and the DEFAULT agent (C) separately. The effectiveness scores for the SCALE-DEFAULT combinations show that the biases of both agents are reflected in the protocol (color bias for the DEFAULT agent and scale bias for the SCALE agent).

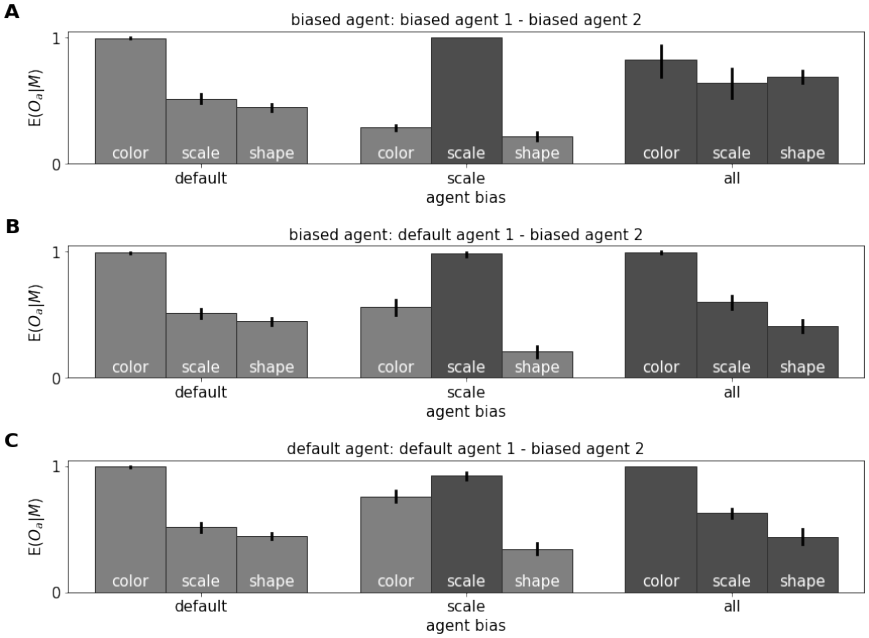


Figure 4.22: Effectiveness per attribute for different pairings of flexible-role agents. Pairings are (A) two agents with the same perceptual bias and (B)+(C) one biased agent and one DEFAULT agent. (B) shows the effectiveness scores for the biased agent and (C) the effectiveness scores for the DEFAULT agent in that mixed combination. The x-axis indicates the bias (of the biased agent or both agents). The bars are labeled with the attribute a used for calculating $E(O_a \mid M)$, with attributes enforced via label smoothing in dark gray. We report means and bootstrapped 95% CIs of ten runs each.

(ii) In the language learning scenario, the flexible-role agent corresponds to the receiver in a fixed-role simulation. Hence, we will only consider the language emergence scenario. The agents achieve average rewards between 0.955 and 0.967 on the test set. The attribute-wise RSA scores are shown in Fig 4.23, which corresponds to Fig 4.7 of the fixed-role agent simulations. In analogy to Fig 4.8, we calculate the difference in general RSA scores before and after training. The RSA score of the DEFAULT agent improves from 0.439 before training to 0.536 (DEFAULT partner), 0.569 (SCALE partner), and 0.572 (ALL partner).

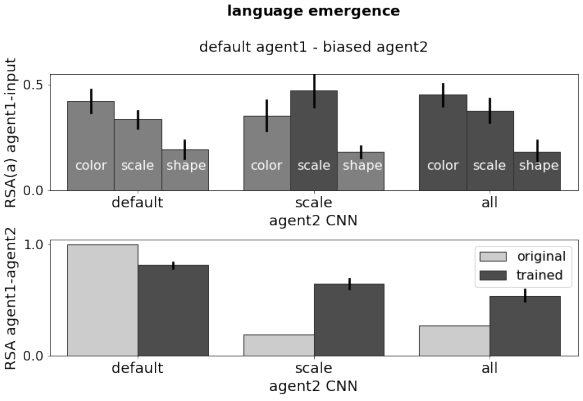


Figure 4.23: Influence of linguistic biases on perception. Shown are the effects of language emergence on a DEFAULT agent, when paired with agents of different visual bias conditions. The visual bias of the communication partner is shown on the x-axis. The top row shows the RSA scores between the DEFAULT agent's visual representations and each object attribute—indicated by the bar label—after training. The bottom row shows the RSA scores between the visual representations of the DEFAULT agent and those of its communication partner before (light gray) and after (dark gray) training. Reported are means and bootstrapped 95% CIs of ten runs each.

(iii) Fig 4.24 shows the test rewards for different combinations of flexible-role agents, which corresponds to Fig 4.9.A of the fixed-role agent simulations. The rewards achieved by the flexible-role agents are slightly lower than those of their fixed-role counterparts. The game is already symmetric, so no additional calculations are necessary to perform a stable state analysis. Only the ALL bias is evolutionary stable, and this stability is significant as determined by pairwise comparisons of the CIs in the third matrix column.

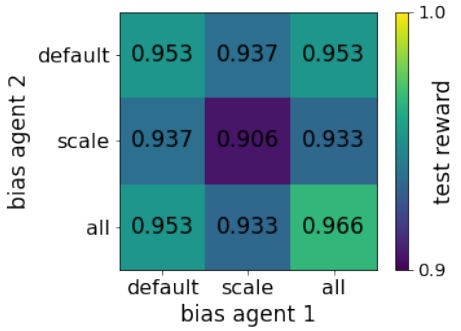


Figure 4.24: Mean reward on the test set for different combinations of two flexible-role agents. We ran 10 simulations for each combination.

In sum, across analyses, the findings from simulations with fixed-role agents generalize to simulations with flexible-role agents.

5 General Discussion

5.1 Summary of the main contributions

The three case studies demonstrate how grounded and functional aspects of language can be studied with (deep) neural networks. An important contribution of this work lies in proposing computational mechanisms for achieving that. In case study 1, we changed the standard architecture of an associative learning model to incorporate pragmatic inference. In case study 2, we designed an alternative to the classical reference game to capture the effects of contextualized reference. In case study 3, we developed a new method to manipulate learned neural network representations to study how differences in representations and communication influence each other.

Case study 1 integrates a pragmatic reasoning mechanism into an associative word learning model, which can be parameterized by lexicon entries or ANN weights. Combining associative learning with pragmatic inference reconciles the two theories of word learning as hypothesis testing and word learning as associative learning. In addition, it provides a natural framework for accommodating different time scales of word learning. The associative component simulates the slow, incremental process by which word-meaning associations come to be stored in the lexicon. The pragmatic component simulates the online process of inferring the meaning of a word in a given context.

Case study 2 demonstrates how agents can learn to refer to an object at different levels of specificity by taking into account other objects present in the context. Our *hierarchical reference game* provides the sender with information about contextual relevance. Our results show that the agents use that information to meet the Maxim of Quantity when referring to an object. They corroborate empirical evidence that the structure of conceptual hierarchies will be reflected in language if it is relevant for communication (Hawkins et al., 2018). In addition, the project highlights the importance of varying the input data in language emergence games, as differences in complexity may lead to differences in communication (cf. Chaabouni et al., 2020a).

Case study 3 establishes the use of DNNs to study the status of labels in cognition. It thereby combines approaches from neuroscience, where DNNs are used to study the visual system, and from evolutionary linguistics and AI, where DNNs are used to study the emergence and evolution of language. The relationship between language and perception has been studied in ANN simulations before (e.g., Lupyan, 2012a), however, deep learning allows us to increase the complexity of the simulations and work with image inputs rather than hand-crafted feature vectors.

Finally, all three case studies make concrete suggestions for improving language learning and communication in ANNs. Case study 1 shows that RSA-like pragmatic reasoning may speed up

(word) learning via the ME bias. Case study 2 shows how artificial agents can learn to choose referential expressions at different levels of specificity. Case study 3 shows that de-biasing visual representations can lead to more successful communication but also that communication can improve visual representations of task-relevant properties. Together, the case studies point out directions for building models of language that take into account the physical and social context of linguistic experience—going beyond supervised learning from text corpora.

5.2 Critical evaluation of our models as models of cognition

The use of ANNs as models of cognition is debated (Cichy & Kaiser, 2019; Griffiths et al., 2010; Kriegeskorte, 2015; Lake et al., 2017; Marcus, 2018; McClelland et al., 2010). There are important differences between how humans and ANNs learn a task. Most prominently, ANNs are notoriously data-hungry whereas humans can learn from a few examples (Lake et al., 2017). The latter is related to the fact that ANNs start out as blank slates whereas humans start out with useful learning mechanisms and constraints that have developed throughout evolution. The use of *deep* neural networks is viewed particularly critically as these are general problem solvers developed for engineering purposes only. DNNs can be fooled in unintuitive ways (Szegedy et al., 2014), and because they comprise a huge number of parameters it is difficult to understand their decisions (Kay, 2017; Marcus, 2018). Below, I will present some main differences between artificial and biological neural networks, argue why ANNs may still be valuable tools for understanding cognition, and discuss what we can learn about cognition from the case studies presented in this thesis.

5.2.1 Differences between artificial and biological neural networks

While originally inspired by biological neuronal networks, ANNs are significantly different in algorithm and architecture, as well as computational ability (for an overview, see Clay, 2022, Chapter 3). In terms of architecture, artificial neurons are an extreme simplification of biological neurons and lack most of their dynamics (Cichy & Kaiser, 2019). Moreover, the neurons in ANNs are highly homogeneous while biological neurons are very heterogeneous in their structure and the way they encode information (Koch & Laurent, 1999). Further simplifications include the lack of bypass connections in feedforward architectures, the lack of feedback and local recurrent connections, the standard types of activation functions (e.g. the rectified linear activation function), the continuous rather than spiking activations, and many more (Kietzmann, McClure, et al., 2019). Concerning the learning algorithm, backpropagation of errors as it is implemented in ANNs does not seem to happen in the brain (Crick, 1989) but it has been proposed that feedback connections induce neural activities that can be used to locally approximate the error signal that is generated in backpropagation (Lillicrap et al., 2020). Much work is being done to increase the similarity between artificial and biological neural networks, for example by adding recurrent connections (e.g., Kietzmann, Spoerer, et al., 2019; Spoerer et al., 2017) or

training on ecologically more valid data sets (Mehrer et al., 2021). At the same time, proponents of deep learning in computational neuroscience point out that abstraction and simplification are desirable and essential to any modeling approach (e.g., Kietzmann, McClure, et al., 2019; Kriegeskorte, 2015).

In terms of computational ability, there are several indicators that ANNs have different learning strategies and learning outcomes than biological neuronal networks. One prominent example is adversarial attacks. Szegedy et al. (2014) identified minimal image perturbations leading to misclassification in DNNs. It turned out that imperceptible perturbations could cause the networks to fail and that these so-called "adversarial examples" generalized to other network instances trained on the same data set as well as networks trained on other data sets. This behavior may be related to the fact that DNNs—unlike humans—rely more on surface texture than object shape when classifying objects (Geirhos et al., 2019). Moreover, ANNs often require large amounts of training data for learning a task (Sun et al., 2017) that humans can learn from a few examples. They also have difficulties learning from data sets that change over time, for example in continuous learning problems or multi-task settings, as the new incoming data can make the network overwrite previously learned weights, leading to catastrophic forgetting (see Section 1.3.3). A lot of these problems seem to arise because ANNs are commonly trained in a supervised manner on a certain type of input data (images, text, etc.) that comes without the richness and context of real-world experience. The networks become experts at pattern recognition but lack a human understanding of the data they are trained on.

As a consequence, Lake et al. (2017) have argued that models of cognition and intelligent behavior must go beyond supervised learning and pattern recognition: First, intelligent machines require causal models of the world, i.e. an understanding of the causes and effects of certain events. Second, they require capacities for intuitive physics and intuitive psychology—which are essentially also causal models. These capacities allow the system to understand the behavior of objects and other agents in the world. Third, to rapidly acquire and generalize knowledge to new tasks, they need to understand the compositional structure underlying their experiences and build up prior knowledge and inductive biases from them.

5.2.2 Predictive, explanatory, and exploratory value of ANN models

Despite the many differences between artificial and biological neural networks, ANNs have been extensively investigated as models of human brain representations and human behavior in cognitive tasks. In some cases, this endeavor has been very successful. Among others, CNNs have been established as state-of-the-art models of neural activity in the visual cortex and human behavior in visual tasks. They can spatially and temporally predict neural activations that are elicited in the visual cortex when viewing images of objects (Cichy et al., 2016; Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014). Further, they have been successful in predicting judgments of object similarity (Jozwik et al., 2017; Peterson et al., 2018), category typicality (Lake et al., 2015), object memorability (Dubey et al., 2015), and shape sensitivity

(Kubilius et al., 2016). Transformers also seem promising as models of human vision. Recent work suggests that, compared to CNNs, their errors might be more consistent with those of humans in visual tasks (Tuli et al., 2021). Besides, DNNs in the form of language models are increasingly being explored as models of human language processing. A systematic comparison between various state-of-the-art language models and brain recordings as well as behavioral signatures from language comprehension tasks demonstrated that the most powerful transformer models can predict human neural and behavioral responses with high accuracy, for some tasks even up to the noise ceiling (Schrimpf et al., 2021). So, even though DNNs may be biologically implausible and lack important principles of human cognition, they can accurately predict neural and behavioral responses for some cognitive tasks.

The question of what we can learn about cognition from building DNNs that can predict brain activations or behavioral responses remains, though. I am sympathetic to the view that DNNs are valuable tools for understanding cognition. To illustrate this point of view, I will rely on the terminology developed by Cichy and Kaiser (2019), who argued for the use of DNNs as scientific models. According to the authors, DNNs have *predictive*, *explanatory*, and *exploratory* power. Predictive power means that DNNs are useful to study cognition because they can predict brain activations or behavioral responses, even without helping us to understand or explain these predictions. For instance, if we know that a DNN can predict human similarity judgments for certain stimuli, we can build new experiments that rely on such perceived similarities without having to collect them. Furthermore, comparing different models in terms of their predictive power may help researchers break down which model components are important to cause certain behaviors. DNNs are often criticized as unsuitable for explanation because the role of their parameters is not determined a priori and it is unclear how these parameters map onto the world. According to the authors, DNNs are for multiple reasons still able to provide scientific explanations. First, their explanation can be considered functional. The behavior of a certain model unit is not explained by identifying a real-world counterpart but rather by understanding its function in a given task. Second, one can consider the model and training setup (training data, architecture, objective, ...) to be meaningful and interpretable parameters. Third, it may be that with more and better techniques for analyzing and visualizing DNN behavior, a mapping between their parameters and the real world can be established. Exploratory power means that analyzing information processing in DNNs may inspire new hypotheses about information processing in the brain. DNNs have exploratory power partly *because* they are different from the brain. Conducting a proof-of-principle demonstration of a certain phenomenon with a DNN can motivate further inquiry into whether aspects that are different in biological neural networks play a role in the explanation of this phenomenon in humans or whether it arises based on general learning principles implemented in both types of networks. In sum, even without increasing the biological plausibility of DNNs—which is an exciting and worthwhile research program—they can be useful models of cognition.

5.2.3 Insights from the case studies

To begin with, some of the criticism passed on ANNs does not apply to the communication game setups of the case studies. On the contrary, the interactive and grounded learning environments directly try to address the issue that ANNs mostly perform pattern recognition without a deeper understanding of objects and other agents in real-world interactions. Of course, there is still a lot of room for improvement, especially in terms of the complexity of the environment and the kinds of interactions afforded by the agents (see Section 5.4). Below, the contribution of each case study will be analyzed critically using the framework proposed by Cichy and Kaiser.

In case study 1, word learning is studied in pragmatic agents. Endowing artificial agents with pragmatic reasoning is an important step toward building intelligent artificial agents. In particular, pragmatic reasoning can be seen as a model of intuitive psychology helping the agents to choose and interpret utterances in context. The DNN implementation is a proof of concept that pragmatic reasoning can lead to an ME bias in DNNs. Its predictions are cognitively implausible, though: The literal listener displays an ME bias as well, and if objects are presented as negative examples the network displays an anti-ME bias. These behaviors arise because some modeling issues—especially how to constrain the embeddings of novel words and objects—have not been solved yet. The lexicon-based models are better suited to model actual word learning than the DNN implementation. We tested different lexicon-based models by simulating the ME bias across time and under various conditions. An evaluation against knowledge about human word learning helped us to narrow down the computational processes of lexicon formation and pragmatic reasoning potentially taking place therein. Hence, the lexicon-based models demonstrate the explanatory potential of (in this case shallow) ANNs.

Case study 2 focuses on the emergence of hierarchical referring expressions in a simple and hand-crafted setup. The cognitive processes of how the input is perceived and how relevant attributes are determined are not modeled but instead hard-coded in the input vectors. Still, the model makes interesting predictions and illustrates that DNNs have exploratory power. First, it predicts that consistent use of the same expression for an abstract concept emerges if many objects can be described by this expression. Second, it predicts that the emergence of hierarchical and (partly) compositional referring expressions go hand in hand. These predictions could be tested in empirical studies.

Case study 3 proposes to combine the two research projects of studying visual perception and the evolution of language with DNNs. The model is highly simplified and constitutes only a first step in this direction. The vision module is implemented as a CNN with two convolutional layers and there are only 64 artificial objects in the world. The model predicts patterns of mutual influence between visual perception and language emergence. Some empirical findings match these predictions but others are unexplained. For example, in our model, changes in visual perception are acquired incrementally over time and remain constant without further learning experiences. In actuality, though, the influence of language on perception is dynamic, online, and task-dependent (e.g., Lupyan et al., 2020; Regier & Xu, 2017). Hence, the model is

mostly exploratory: It suggests that some phenomena, like categorical perception, can arise from general learning principles found in both artificial and biological neural networks while additional (brain-inspired) mechanisms must be implemented to account for the variable nature of Whorfian effects. However, the proposed framework of combining communication games with DNNs to study language-related interface phenomena in cognition has the potential for becoming predictive and explanatory of these phenomena. I envision a cyclic process in future work: changes to the model lead to better predictions of empirical results, which in turn inform changes to the model.

5.3 Communication games and neural networks: mix and match

In part, the presented work aims to establish communication games between DNN agents as a test bed for cognitive language-related phenomena. Communication games and DNN architectures—together with the choice of input and output representations—are highly versatile. As a result, combining the two methodologies creates enormous modeling flexibility. The three case studies nicely illustrate this adaptive "mix-and-match" character.

Communication games can be used to study horizontal (within-generation) and vertical (across-generation) interactions. For example, case study 2 simulates the *emergence* of language in horizontal interactions. Case study 1, in contrast, simulates language *learning*, so vertical transmission from one generation to the next. Similarly, *cultural evolution* can be modeled with repeated language learning simulations spanning multiple generations (e.g., Kirby, 2002b). Case study 3 investigates language emergence, language learning, and language evolution. For the latter, we apply an ESS analysis to the agents' payoff matrices. ESS analysis is compatible with different optimization processes, including cultural and *biological evolution*. Next to ESS analysis, also dynamic models of evolution can be applied. In conclusion, the three case studies illustrate how communication games can be used to zoom in or out on the time scales of language evolution.

Besides, communication games and ANN architectures can be flexibly adapted to study specific phenomena. While case study 1 and case study 3 rely on the prototypical reference game, case study 2 develops the hierarchical reference game which requires communication of compositional and hierarchical concepts. Case study 2, in turn, relies on standard ANN architectures, whereas case study 1 integrates a pragmatic reasoning mechanism, and case study 3 manipulates visual processing. Combinations of these changes are also conceivable. One could, for example, simulate the emergence of hierarchical reference systems (case study 2) in pragmatic agents (case study 1) that reason about the optimal level of specificity.

In general, various game setups and ANN architectures are being employed in current research. The reference game is widely used but has been instantiated very differently. Among others, variations arise from the use of different input data. For example, Chaabouni et al. (2021) use color patches as inputs to study the emergence of color-naming systems. Alternatively, the

reconstruction game is also very popular. In that game, the receiver has to reconstruct the target object instead of selecting it among a set of distractors (e.g., Chaabouni et al., 2020a). But also more specialized games have been developed. The *fruit and tools game* (Bouchacourt & Baroni, 2019), for instance, was developed to study communication about affordances. One agent has a fruit, the other agent has different tools, and they have to communicate to identify the right tool for eating the fruit. In terms of ANNs, different standard architectures are employed, including CNNs (for image processing) (Lazaridou et al., 2017, e.g.,), MLPs (e.g., Lazaridou et al., 2017), different RNN variants (e.g., Chaabouni et al., 2020a), and attention-based networks (e.g., Inala et al., 2020). But also specialized networks have been developed, for example *SARNet* (e.g., Rangwala & Williams, 2020), which includes memory and attention mechanisms to selectively reason about the value of information received from other agents while considering past experience. Taken together, different neural network modules can be combined to construct different architectures which can in turn be combined with different data sets and communication games, resulting in considerable flexibility.

5.4 Outlook

The agents, the actions they can perform, and the environments in which they learn are very simple in our experiments. These simple setups offer control and interpretability. In addition, all case studies modify standard architectures or communication games and the effect of such modifications is best evaluated without introducing additional, potentially confounding complexity. Still, it seems plausible that ANNs become more powerful models of cognition when the discrepancy between real-world and simulated learning environments is reduced.

5.4.1 From toy data sets to natural images and natural language

Working with more realistic input data is one way to reduce this discrepancy. Future work should extend our experiments to natural images, natural language (where appropriate), and consequently deeper networks. These experiments are important to investigate whether some of our results are dependent on the simplified input structure, and how they change under increasing realism. In both language learning and language emergence simulations, models have been shown to behave differently when trained on natural images versus symbolic inputs (e.g., Gulordava et al., 2020; Lazaridou et al., 2018). For example, emergent protocols have less compositional structure when agents communicate about images instead of disentangled feature vectors (Lazaridou et al., 2018). A main feature of DNNs is that they can learn from large and complex data sets. Thus, in principle, nothing stands in the way of using more realistic data.

Case study 1 started out with simple associative models and then transitioned to a proof-of-concept DNN implementation. The latter showed the implementation of pragmatic reasoning on a joint embedding space is not trivial even with idealized input data. Future work needs to

find plausible and systematic ways of constraining the embedding space before transitioning to natural images and natural language. In case study 2, the objects in the game were attribute vectors with well-defined compositional structure and the agents' knowledge about relevant aspects of the environment was hard-coded. Proving that hierarchical reference systems emerge in this very artificial setting has paved the way for future work a) endowing the agents with pragmatic reasoning abilities and b) studying whether compositional and hierarchical reference systems also emerge in the case of natural images, which have less obvious structure. In case study 3, using well-defined and abstract image data had the advantage that we could introduce systematic variations in perceived object similarity and easily quantify perceptual bias. Using larger DNNs and data sets will likely not change the main principles of interaction between language and perception that we discovered in our toy setup. However, such an upgrade is essential to reach a point where the model can exploit the ability of powerful CNNs and transformers models to predict human visual and linguistic processing.

5.4.2 From perceiving to embodied agents

A natural next step toward more human-like learning scenarios lies in the use of agents whose interactions with the environment go beyond communication. This thesis promotes the use of communication games between DNN agents as a framework for studying interactions between language and other cognitive processes. While the case studies consider the role of perception, theory-of-mind reasoning, and reasoning about the context, many other aspects of language use could be studied in this framework. In particular, the agents in the three case studies are not embodied and cannot act in their environment apart from producing or interpreting messages. However, as discussed in Section 1.2, language is grounded in our sensori-*motor* interaction with the world. To account for the interactions between language and motor cognition, future work should study representation learning and emergent communication in situated agents.

In learning about the world through embodied interaction, situated agents can construct rich pre-linguistic representations from which to generalize. In simulations by Clay et al. (2021), a situated agent learned stable representations of meaningful concepts without supervision. The agent encoded concepts relative to their affordances, i.e. relative to the actions that it could perform. For example, doors in the left visual field were encoded differently from doors in the right visual field, as different movements are required to pass through these doors. In addition, the authors discovered that, in contrast to non-embodied models such as classifiers and auto-encoders, the agent learned sparse representations of its environment. These simulations capture the relation between concept formation and sensorimotor interaction with the environment and show how embodied learning simulations can lead to the emergence of more human-like (here: action-oriented and sparse) representations. Working with situated multi-agent communication games will be essential for building agents that understand, interact with, and communicate about objects and more abstract concepts in terms of their affordances.

Simulations with situated agents and suitable training environments are not accessible in the same off-the-shelf manner as computer vision models and image data sets. Still, the field of deep reinforcement learning is developing rapidly and by now various platforms for embodied AI research exist (Beattie et al., 2016; Chevalier-Boisvert et al., 2019; Johnson et al., 2016; Kempka et al., 2016; Savva et al., 2019; Wu et al., 2018; Zhu et al., 2017), some of which enable training of embodied agents in highly efficient (photo-)realistic 3D simulation (e.g., Savva et al., 2019; Wu et al., 2018; Zhu et al., 2017). At the same time, there is increasing interest in language learning and language emergence simulations with situated agents (e.g., Chevalier-Boisvert et al., 2019; Das et al., 2018; Das et al., 2019; Fried, Hu, et al., 2018; Hermann et al., 2017; Hill et al., 2017; Hill, Clark, et al., 2020; Hill, Lampinen, et al., 2020; Jaques et al., 2018; Kajic et al., 2020; Mordatch & Abbeel, 2018). Even though most of this research is application-driven, simulations with embodied agents in these ever more realistic environments could serve as a basis for modeling human language learning and language use. Aside from maximizing performance, future work should strive to analyze representation learning, concept formation, and language understanding in dependence on the agents' environment, tasks, and interaction possibilities.

5.4.3 From reference to more uses of language

Our simulations never go beyond referential communication. Reference is a good starting point as it is arguably a core function of language around which more complex functions are organized (Jackendoff, 1999). Still, there is so much more we can do with language and eventually models of language learning and language emergence will have to deal with these other use cases.

In modern language emergence research, there has been considerable progress in going beyond the classical sender-receiver reference game. A large variety of games have been tested involving for example negotiation (Cao et al., 2018), navigation (Das et al., 2019; Kajic et al., 2020), communication about affordances (Bouchacourt & Baroni, 2019), and coordination in complex social dilemma environments (Jaques et al., 2018). In addition, the reference game has been extended to populations of agents to study community size as a driver of systematicity (Kim & Oh, 2021; Rita et al., 2022) or contact linguistic phenomena (Harding Graesser et al., 2019); and it has been combined with cultural transmission to study the emergence of compositionality (Chaabouni et al., 2020a; Li & Bowling, 2019; Ren et al., 2020). Still, the languages emerging in these experiments are highly task-specific and far away from the ultimate goal of flexible, goal-directed language use.

Rich, interactive 3D environments for multi-agent communication simulations have the potential to create more versatile communicative needs and as a result more flexible language use. To exploit this potential, asymmetries in information have to arise for multiple ways of interacting with the world. Maybe one agent has already explored the environment when a new agent joins, maybe agents move around in different locations with limited visual fields, maybe they need to coordinate to solve a task quickly, and so on. Referring to objects will only be one component of

the protocols these agents need to develop. Building environments and tasks that foster these different types of communicative exchanges is one of the main challenges for future work.

5.5 Conclusion

This thesis embraces a holistic perspective on language. It set out to study grounded and functional language phenomena with ANNs. Three case studies were presented that illustrate how communication games and ANNs can be combined and, importantly, modified to investigate word learning in pragmatic agents, reference at different levels of specificity based on the context, and interactions between language and perception. They also demonstrate the value of ANNs as models of cognition by making predictions that either (qualitatively) match human behavior or by generating new hypotheses that can be evaluated by empirical studies. Although ANNs are arguably useful scientific models, there are significant differences between how ANNs and humans learn and solve tasks. The case studies in this work are in line with a general effort in cognitive science and AI research to bridge this gap. Using communication games to study grounded and interactive language use is an important step toward artificial agents that can "understand" the meanings of words. Similarly, taking into account the role of intentions and context is an important step toward artificial agents that can use language flexibly in different situations. However, the human language faculty is tightly intertwined with many other cognitive functions, generating a web of interactions. Even if DNNs are excellent tools for modeling learning from complex large-scale data sets, it is difficult to imagine how a *human-like* understanding and use of language can be realized with out-of-the-box architectures. I think that significant progress will be made by following current trends toward brain-inspired DNNs and training in simulated or even real interactive environments.

General Appendix

D Self-organizing models of word learning

Self-organizing models of word learning rely on the framework of self-organizing maps (SOMs) in combination with activity-dependent learning algorithms such as Hebbian learning. SOMs are an unsupervised machine learning technique for learning low-dimensional representations from high-dimensional input spaces while preserving the original topology (Kohonen, 1997). In Hebbian learning, correlated activation of connected neurons strengthens their connection weight (Choe, 2013). Single-layer SOMs have been used to simulate specific aspects of language, but to account for multiple sources of information (e.g. vision and language) several SOMs are combined (Li & Zhao, 2013).

One of the earliest implementations of an integrated SOM architecture was *DisLex* by Miikkulainen (1997a). The model combined SOMs for different modalities (orthographic input and phonological input) with an SOM of semantic concepts, and connections were trained through Hebbian learning. By applying selective lesions, the model could simulate different language impairments. Inspired by this work, Li and colleagues proposed two variants of an SOM-based model of lexical development, *DevLex-I* (Li et al., 2004) and *DevLex-II* (Li et al., 2007). *DevLex-I* connected a phonological input map and a semantic map to simulate comprehension and in *DevLex-II* a phonological output map was added to simulate production. *DevLex-II* could account for various empirical word learning patterns, among others the vocabulary spurt and effects of word frequency and length on the age of acquisition. Later, Mayor and Plunkett (2010) introduced a self-organizing model with one SOM for visual input (the object) and one SOM for acoustic input (the word) (see Figure D.1). Here, too, weights between the layers were updated through Hebbian learning. The model displayed a vocabulary spurt, a taxonomic bias, and fast mapping. Taken together, these works establish SOMs as successful models of several important word learning phenomena. Advantages of SOMs are the unsupervised learning process and arguably biological plausibility (e.g., Li & Zhao, 2013; Li et al., 2007; Miikkulainen, 1997b).

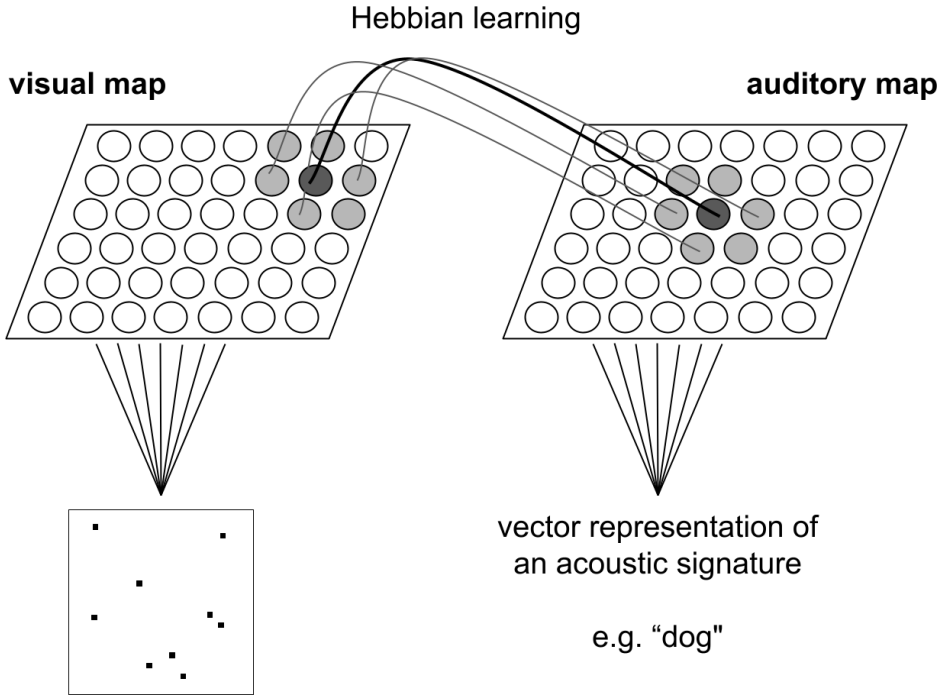


Figure D.1: Illustration of the self-organizing model by Mayor and Plunkett (2010) (inspired by Figure 1 in the paper). The visual input consists of a set of random dot images and the linguistic input consists of vector encodings of acoustic tokens (words). A visual and an auditory SOM are trained separately on the respective type of input. Then the two input types are presented simultaneously and associative weights between the two maps are learned according to Hebb's rule.

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., ... Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. http://tensorflow.org/

Ahn, S., Zelinsky, G. J., & Lupyan, G. (2021). Use of superordinate labels yields more robust an human-like visual representations in convolutional neural networks. *Journal of Vision*, *21*(13), 1–19. https://doi.org/10.1167/jov.21.13.13

Andreas, J., & Klein, D. (2016). Reasoning about pragmatics with neural listeners and speakers. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1173–1182. https://doi.org/10.18653/v1/D16-1125

Appelt, D. (1985). Planning English referring expressions. *Artificial Intelligence*, *26*(1), 1–33. https://doi.org/10.1016/0004-3702(85)90011-6

Appelt, D., & Kronfeld, A. (1987). A computational model of referring. *Proceedings of the 10th International Joint Conference on Artificial Intelligence (IJCAI)*, 640–647.

Aslin, R. N., Woodward, J. Z., LaMendola, N. P., & Bever, T. G. (1996). Models of word segmentation in fluent maternal speech to infants. In J. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 117–134). https://doi.org/10.4324/9781315806822

Austin, J. L. (1962). *How to do things with words*. Harvard University Press.

Axelsson, E. L., Churchley, K., & Horst, J. S. (2012). The right thing at the right time: Why ostensive naming facilitates word learning. *Frontiers in Psychology*, *3*(88), 1–8. https://doi.org/10.3389/fpsyg.2012.00088

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349. https://doi.org/10.1016/j.cognition.2009.07.005

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition*, *118*(1), 84–93. https://doi.org/10.1016/j.cognition.2010.10.010

Baronchelli, A., Loreto, V., & Steels, L. (2008). In-depth analysis of the naming game dynamics: The homogeneous mixing case. *International Journal of Modern Physics C*, *19*(5), 785–812. https://doi.org/10.1142/S0129183108012522

Barrett, L., Lindquist, K. A., & Gendron, M. (2007). Language as context for the perception of emotion. *Trends in cognitive science*, *11*(8), 327–332. https://doi.org/10.1016/j.tics.2007.06.003

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–609. https://doi.org/10.1017/s0140525x99002149

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Barsalou, L. W. (2010). Grounded cognition: Past, present, and future. *Topics in Cognitive Science*, *2*, 716–724. https://doi.org/10.1111/j.1756-8765.2010.01115.x

Barsalou, L. W., Simmons, W., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences*, *7*(2), 84–91. https://doi.org/10.1016/S1364-6613(02)00029-3

Batali, J. (1998). Computational simulations of the emergence of grammar. In J. R. Hurford, M. Studdert-Kennedy, & C. Knight (Eds.), *Approaches to the evolution of language: Social and cognitive biases*. Cambridge University Press. https://doi.org/10.1525/jlin.2000.10.2.291

Beattie, C., Leibo, J. Z., Teplyashin, D., Ward, T., Wainwright, M., Küttler, H., Lefrancq, A., Green, S., Valdés, V., Sadik, A., Schrittwieser, J., Anderson, K., York, S., Cant, M., Cain, A., Bolton, A., Gaffney, S., King, H., Hassabis, D., . . . Petersen, S. (2016). Deepmind lab. *arXiv preprint*, *arXiv:1612.03801*. https://doi.org/10.48550/arXiv.1612.03801

Bechtel, W. (1988). *Philosophy of science: An overview for cognitive science*. Lawrence Erlbaum Associates. https://doi.org/10.4324/9781315802084

Behrend, D., Scofield, J., & Kleinknecht, E. (2001). Beyond fast mapping: Young children's extensions of novel words and novel facts. *Developmental Psychology*, *37*(5), 698–705. https://doi.org/10.1037//0012-1649.37.5.698

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166. https://doi.org/10.1109/72.279181

Bernardi, R., Boleda, G., Fernández, R., & Paperno, D. (2015). Distributional semantics in use. *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, 95–101. https://doi.org/10.18653/v1/W15-2712

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, *15*(11), 527–536. https://doi.org/10.1016/j.tics.2011.10.001

Bion, R., Borovsky, A., & Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel word-object mappings in relation to vocabulary learning at 18, 24, and 30 months. *Cognition*, *126*(1), 39–53. https://doi.org/10.1016/j.cognition.2012.08.008

Bishop, C. M. (2006). *Pattern recognition and machine learning (information science and statistics)*. Springer-Verlag.

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N., & Turian, J. (2020). Experience grounds language. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8718–8735. https://doi.org/10.18653/v1/2020.emnlp-main.703

Bleys, J., Loetzsch, M., Spranger, M., & Steels, L. (2009). The grounded colour naming game. *Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication (Ro-Man)*.

Bloom, P. (2000). *How children learn the meanings of words*. MIT Press. https://doi.org/10.7551/mitpress/3577.001.0001

Blume, A., DeJong, D. V., Kim, Y.-G., & Sprinkle, G. B. (1998). Experimental evidence on the evolution of meaning of messages in sender-receiver games. *The American Economic Review*, *88*(5), 1323–1340.

Bohn, M., & Frank, M. C. (2019). The pervasive role of pragmatics in early language. *Annual Review of Developmental Psychology*, *1*(1), 223–249. https://doi.org/10.1146/annurev-devpsych-121318-085037

Börgers, T., & Sarin, R. (1997). Learning through reinforcement and replicator dynamics. *Journal of Economic Theory*, *77*(1), 1–14. https://doi.org/10.1006/jeth.1997.2319

Bouchacourt, D., & Baroni, M. (2018). How agents see things: On visual representations in an emergent language game. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 981–985. https://doi.org/10.18653/v1/D18-1119

Bouchacourt, D., & Baroni, M. (2019). Miss Tools and Mr Fruit: Emergent communication in agents learning about object affordances. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 3909–3918. https://doi.org/10.18653/v1/P19-1380

Brighton, H., & Kirby, S. (2006). Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artificial Life*, *12*(2), 229–242. https://doi.org/10.1162/106454606776073323

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., . . . Amodei, D. (2020). Language models are few-shot learners. *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, 1877–1901.

Buckner, C., & Garson, J. (2019). Connectionism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2019). Metaphysics Research Lab, Stanford University.

Burgess, C., & Kim, H. (2018). 3D shapes dataset.

Cangelosi, A. (1999). Modeling the evolution of communication: From stimulus associations to grounded symbolic associations. In D. Floreano, J.-D. Nicoud, & F. Mondada (Eds.), *Advances in artificial life* (pp. 654–663). Springer. https://doi.org/10.1007/3-540-48304-7_86

Cangelosi, A. (2005). The emergence of language: Neural and adaptive agent models. *Connection Science*, *17*(3–4), 185–190. https://doi.org/10.1080/09540090500177471

Cangelosi, A. (2010). Grounding language in action and perception: From cognitive agents to humanoid robots. *Physics of life reviews*, *7*(2), 139–151. https://doi.org/10.1016/j.plrev.2010.02.001

Cangelosi, A., & Harnad, S. (2000). The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in perceptual categories. *Evolution of Communication*, *4*, 117–142. https://doi.org/10.1075/eoc.4.1.07can

Cangelosi, A., & Parisi, D. (1998). The emergence of a 'language' in an evolving population of neural networks. *Connection Science*, *10*(2), 83–97. https://doi.org/10.1080/095400998116512

Cangelosi, A., & Parisi, D. (2001). How nouns and verbs differentially affect the behavior of artificial organisms. *Proceedings of the 23rd Annual Meeting of the Cognitive Science Society (CogSci)*.

Cangelosi, A., & Parisi, D. (2002). *Simulating the evolution of language*. Springer. https://doi.org/10.1007/978-1-4471-0663-0

Cao, K., Lazaridou, A., Lanctot, M., Leibo, J. Z., Tuyls, K., & Clark, S. (2018). Emergent communication through negotiation. *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

Carey, S. (1978). The child as a word learner. In M. Halle, J. Bresnan, & G. A. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 264–293). MIT Press.

Chaabouni, R., Kharitonov, E., Bouchacourt, D., Dupoux, E., & Baroni, M. (2020a). Compositionality and generalization in emergent languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 4427–4442. https://doi.org/10.18653/v1/2020.acl-main.407

Chaabouni, R., Kharitonov, E., Bouchacourt, D., Dupoux, E., & Baroni, M. (2020b). Compositionality and generalization in emergent languages. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 4427–4442. https://doi.org/10.18653/v1/2020.acl-main.407

Chaabouni, R., Kharitonov, E., Dupoux, E., & Baroni, M. (2019). Anti-efficient encoding in emergent communication. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Proceedings of the 33rd conference on neural information processing systems (NeurIPS)*.

Chaabouni, R., Kharitonov, E., Dupoux, E., & Baroni, M. (2021). Communicating artificial neural networks develop efficient color-naming systems. *Proceedings of the National Academy of Sciences (PNAS)*, *118*(12). https://doi.org/10.1073/pnas.2016569118

Chaabouni, R., Strub, F., Altché, F., Tarassov, E., Tallec, C., Davoodi, E., Mathewson, K. W., Tieleman, O., Lazaridou, A., & Piot, B. (2022). Emergent communication at scale. *Proceedings of the 10th International Conference on Learning Representations (ICLR)*.

Chaudhari, S., Mithal, V., Polatkan, G., & Ramanath, R. (2021). An attentive survey of attention models. *ACM Transactions on Intelligent Systems and Technology*, *12*(5), 1–32. https://doi.org/10.1145/3465055

Chevalier-Boisvert, M., Bahdanau, D., Lahlou, S., Willems, L., Saharia, C., Nguyen, T. H., & Bengio, Y. (2019). BabyAI: First steps towards grounded language learning with a human in the loop. *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. https://doi.org/10.3115/v1/D14-1179

Choe, Y. (2013). Hebbian learning. In D. Jaeger & R. Jung (Eds.), *Encyclopedia of computational neuroscience* (pp. 1–5). Springer. https://doi.org/10.1007/978-1-4614-7320-6_672-1

Choi, E., Lazaridou, A., & de Freitas, N. (2018). Compositional obverter communication learning from raw visual input. *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

Chomsky, N. (1965). *Language and mind*. Harcourt, Brace, & Jovanovich. https://doi.org/10.1017/CBO9780511791222

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, *13*(2-3), 221–268. https://doi.org/10.1080/016909698386528

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *The behavioral and brain sciences*, *31*(5), 489–558. https://doi.org/10.1017/S0140525X08004998

Cibelli, E., Xu, Y., Austerweil, J. L., Griffiths, T. L., & Regier, T. (2016). The sapir-whorf hypothesis and probabilistic inference: Evidence from the domain of color. *PLOS ONE*, *11*(7), 1–28. https://doi.org/10.1371/journal.pone.0158725

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences*, *23*(4), 305–317. https://doi.org/10.1016/j.tics.2019.01.009

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*, 1–13. https://doi.org/10.1038/srep27755

Clark, A. (1993). *Associative engines: Connectionism, concepts, and representational change*. MIT Press. https://doi.org/10.5555/164767

Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 1–33). Lawrence Erlbaum Associates.

Clark, E. V. (1988). On the logic of contrast. *Journal of Child Language*, *15*(2), 317–335. https://doi.org/10.1017/s0305000900012393

Clark, E. V. (2009). *First language acquisition*. Cambridge University Press. https://doi.org/10.1017/CBO9781316534175

Clark, E. V., & Amaral, P. M. (2010). Children build on pragmatic information in language acquisition. *Language and Linguistics Compass*, *4*(7), 445–457. https://doi.org/10.1111/j.1749-818X.2010.00214.x

Clark, H. H. (1992). *Arenas of language use*. University of Chicago Press. https://doi.org/10.1177/002383099403700209

Clark, H. H. (1996). *Using language*. Cambridge University Press. https://doi.org/10.1017/CBO9780511620539

Clay, V. (2022). *The role of task and environment in biologically inspired artificial intelligence: Learning as an active, sensorimotor process* (Dissertation). Universität Osnabrück. https://osnadocs.ub.uni-osnabrueck.de/handle/ds-202204226768

Clay, V., König, P., Kühnberger, K.-U., & Pipa, G. (2021). Learning sparse and meaningful representations through embodiment. *Neural Networks*, *134*, 23–41. https://doi.org/10.1016/j.neunet.2020.11.004

Cohn-Gordon, R., & Goodman, N. (2019). Lost in machine translation: A method to reduce meaning loss. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 437–441. https://doi.org/10.18653/v1/N19-1042

Cohn-Gordon, R., Goodman, N., & Potts, C. (2018). Pragmatically informative image captioning with character-level inference. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 439–443. https://doi.org/10.18653/v1/N18-2070

Correia, J. (2019). *Meaningful play: Signaling games in light of later wittgenstein* (Doctoral dissertation). Universidade do Porto.

Crawford, V. P. (1998). A survey of experiments on communication via cheap talk. *Journal of Economic Theory*, *78*(2), 286–298. https://doi.org/10.1006/jeth.1997.2359

Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. *Econometrica*, *50*(6), 1431–1451. https://doi.org/10.2307/1913390

Cressman, R. (2003). *Evolutionary dynamics and extensive form games*. MIT Press. https://doi.org/10.7551/mitpress/2884.003.0001

Crick, F. (1989). The recent excitement about neural networks. *Nature*, *337*(6203), 129–132. https://doi.org/10.1038/337129a0

Dagan, G., Hupkes, D., & Bruni, E. (2021). Co-evolution of language and agents in referential games. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2993–3004. https://doi.org/10.18653/v1/2021.eacl-main.260

Dale, R., & Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, *19*(2), 233–263. https://doi.org/10.1016/0364-0213(95)90018-7

Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., & Batra, D. (2018). Embodied question answering. *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–10. https://doi.org/10.1109/CVPR.2018.00008

Das, A., Gervet, T., Romoff, J., Batra, D., Parikh, D., Rabbat, M., & Pineau, J. (2019). TarMAC: Targeted multi-agent communication. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 1538–1546.

Delancey, S. (1997). Mirativity: The grammatical marking of unexpected information. *Linguistic Typology*, *1*, 33–52. https://doi.org/10.1515/lity.1997.1.1.33

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Dessi, R., Kharitonov, E., & Baroni, M. (2021). Interpretable agent communication from scratch (with a generic visual processor emerging on the side). *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 26937–26949.

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

DeVries, T., & Taylor, G. W. (2018). Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint*, *arXiv:1802.04865*. https://doi.org/10.48550/arXiv.1802.04865

Diesendruck, G., & Bloom, P. (2003). How specific is the shape bias? *Child Development*, *74*(1), 168–178. https://doi.org/10.1111/1467-8624.00528

Dils, A. T., & Boroditsky, L. (2010). Processing unrelated language can change what you see. *Psychonomic Bulletin & Review*, *17*(6), 882–888. https://doi.org/10.3758/PBR.17.6.882

Dubey, R., Peterson, J., Khosla, A., Yang, M.-H., & Ghanem, B. (2015). What makes an object memorable? *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, 1089–1097. https://doi.org/10.1109/ICCV.2015.130

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211. https://doi.org/10.1016/0364-0213(90)90002-E

Foerster, J. N., Assael, Y. M., de Freitas, N., & Whiteson, S. (2016). Learning to communicate with deep multi-agent reinforcement learning. *Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS)*, 2145–2153.

Foglia, L., & Wilson, R. A. (2013). Embodied cognition. *WIREs Cognitive Science*, *4*(3), 319–325. https://doi.org/10.1002/wcs.1226

Forder, L., & Lupyan, G. (2019). Hearing words changes color perception: Facilitation of color discrimination by verbal and visual cues. *Journal of Experimental Psychology: General*, *148*(7), 1105–1123. https://doi.org/10.1037/xge0000560

Frank, M. C. (2016). Rational speech act models of pragmatic reasoning in reference games. *PsyArXiv*, *osf.io/x9mre/*. https://doi.org/10.31234/osf.io/f9y6b

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998. https://doi.org/10.1126/science.1218633

Frank, M. C., & Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, *75*, 80–96. https://doi.org/10.1016/j.cogpsych.2014.08.002

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–585. https://doi.org/10.1111/j.1467-9280.2009.02335.x

Frank, M. C., Sugarman, E., Horowitz, A. C., Lewis, M. L., & Yurovsky, D. (2016). Using tablets to collect data from young children. *Journal of Cognition and Development*, *17*(1), 1–17. https://doi.org/10.1080/15248372.2015.1061528

Franke, M., & Correia, J. (2018). Vagueness and imprecise imitation in signaling games. *The British Journal for the Philosophy of Science*, *69*(4), 1037–1067. https://doi.org/10.1093/bjps/axx002

Franke, M., & Wagner, E. O. (2014). Game theory and the evolution of meaning. *Language and Linguistics Compass*, *8*(9), 359–372. https://doi.org/10.1111/lnc3.12086

Fried, D., Andreas, J., & Klein, D. (2018). Unified pragmatic models for generating and following instructions. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 1951–1963. https://doi.org/10.18653/v1/N18-1177

Fried, D., Hu, R., Cirik, V., Rohrbach, A., Andreas, J., Morency, L.-P., Berg-Kirkpatrick, T., Saenko, K., Klein, D., & Darrell, T. (2018). Speaker-follower models for vision-and-language navigation. *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*.

Gandhi, K., & Lake, B. M. (2020). Mutual exclusivity as a challenge for neural networks. *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, 14182–14192.

Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought*. MIT press. https://doi.org/10.7551/mitpress/2076.001.0001

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

Gendron, M., Lindquist, K. A., Barsalou, L. W., & Barrett, L. (2012). Emotion words shape emotion percepts. *Emotion*, *12*(2), 314–325. https://doi.org/10.1037/a0026007

Gibbs, R. W., Bogdanovich, J. M., Sykes, J. R., & Barr, D. J. (1997). Metaphor in idiom comprehension. *Journal of Memory and Language*, *37*(2), 141–154. https://doi.org/10.1006/jmla.1996.2506

Glenberg, A. M., Witt, J. K., & Metcalfe, J. (2013). From the revolution to embodiment: 25 years of cognitive psychology. *Perspectives on Psychological Science*, *8*(5), 573–585. https://doi.org/10.1177/1745691613498098

Gliozzi, V., Mayor, J., Hu, J.-F., & Plunkett, K. (2009). Labels as features (not names) for infant categorization: A neurocomputational approach. *Cognitive Science*, *33*(4), 709–738. https://doi.org/10.1111/j.1551-6709.2009.01026.x

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *WIREs Cognitive Science*, *1*, 69–78. https://doi.org/10.1002/wcs.26

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. https://doi.org/10.5555/3086952

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*(1), 173–184. https://doi.org/10.1111/tops.12007

Graesser, A. C., Wiemer-Hastings, P., & Wiemer-Hastings, K. (2001). Constructing inferences and relations during text comprehension. In T. Sanders, J. Schilperoord, & W. Spooren (Eds.), *Text representation: Linguistic and psycholinguistic aspects* (pp. 249–271). Benjamins. https://doi.org/10.1075/hcp.8.14gra

Graf, C., Degen, J., Hawkins, R. X., & Goodman, N. D. (2016). Animal, dog, or dalmatian? level of abstraction in nominal referring expressions. *Proceedings of the 38th Annual Meeting of the Cognitive Science Society (CogSci)*, 2261–2266.

Grandison, A., Sowden, P. T., Drivonikou, V. G., Notman, L. A., Alexander, I., & Davies, I. R. L. (2016). Chromatic perceptual learning but no category effects without linguistic input. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00731

Grassmann, S., Schulze, C., & Tomasello, M. (2015). Children's level of word knowledge predicts their exclusion of familiar objects as referents of novel words. *Frontiers in Psychology*, *6*, 1–8. https://doi.org/10.3389/fpsyg.2015.01200

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S. G., Grefenstette, E., Ramalho, T., Agapiou, J., Badia, A. P., Hermann, K. M., Zwols, Y., Ostrovski, G., Cain, A., King, H., Summerfield, C., Blunsom, P., Kavukcuoglu, K., & Hassabis, D. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, *538*(7626), 471–476. https://doi.org/10.1038/nature20101

Grice, H. P. (1975). Logic and conversation. In *Syntax and semantics 3: Speech acts* (pp. 41–58). Academic Press.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*, 357–364. https://doi.org/10.1016/j.tics.2010.05.004

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., & Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, *77*, 354–377. https://doi.org/10.1016/j.patcog.2017.10.013

Güçlü, U., & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015

Gulordava, K., Brochhagen, T., & Boleda, G. (2020). Deep daxes: Mutual exclusivity arises through both learning biases and pragmatic strategies in neural networks. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society (CogSci)*, 2089–2095.

Halberda, J. (2003). The development of a word-learning strategy. *Cognition*, *87*(1), B23–B34. https://doi.org/10.1016/S0010-0277(02)00186-5

Harding Graesser, L., Cho, K., & Kiela, D. (2019). Emergent linguistic phenomena in multi-agent communication games. *Proceedings of the 2019 Conference on Empirical Methods in Natural*

Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3700–3710. https://doi.org/10.18653/v1/D19-1384

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1), 335–346. https://doi.org/10.1016/0167-2789(90)90087-6

Harnad, S., Hanson, S. J., & Lubin, J. (1991). Categorical perception and the evolution of supervised learning in neural nets. In D. W. Powers & L. Reeker (Eds.), *Working papers of the AAAI spring symposium on machine learning of natural language and ontology* (pp. 65–74).

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, *41*(2), 301–307. https://doi.org/10.1016/S0896-6273(03)00838-9

Hauser, M. D., Yang, C., Berwick, R. C., Tattersall, I., Ryan, M. J., Watumull, J., Chomsky, N., & Lewontin, R. C. (2014). The mystery of language evolution. *Frontiers in Psychology*, *5*(401), 1–12. https://doi.org/10.3389/fpsyg.2014.00401

Havas, D. A., Glenberg, A. M., Gutowski, K. A., Lucarelli, M. J., & Davidson, R. J. (2010). Cosmetic use of botulinum toxin-a affects processing of emotional language. *21*(7), 895–900. https://doi.org/0.1177/0956797610374742

Havrylov, S., & Titov, I. (2017). Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*, 2149–2159.

Hawkins, R. D., Franke, M., Smith, K., & Goodman, N. D. (2018). Emerging abstractions: Lexical conventions are shaped by communicative context. *Proceedings of the 40th Annual Meeting of the Cognitive Science Society (CogSci)*, 463–468.

Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.

Hermann, K. M., Hill, F., Green, S., Wang, F., Faulkner, R., Soyer, H., Szepesvari, D., Czarnecki, W. M., Jaderberg, M., Teplyashin, D., Wainwright, M., Apps, C., Hassabis, D., & Blunsom, P. (2017). Grounded language learning in a simulated 3d world. *arXiv preprint*, *arXiv:1706.06551*. https://doi.org/10.48550/arXiv.1706.06551

Higgins, I., Sonnerat, N., Matthey, L., Pal, A., Burgess, C. P., Bošnjak, M., Shanahan, M., Botvinick, M., Hassabis, D., & Lerchner, A. (2018). SCAN: Learning hierarchical compositional visual concepts. *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

Hill, F., Clark, S., Hermann, K. M., & Blunsom, P. (2017). Understanding early word learning in situated artificial agents. *arXiv preprint*, *arXiv:1710.09867*. https://doi.org/10.48550/arXiv.1710.09867

Hill, F., Clark, S., Hermann, K. M., & Blunsom, P. (2020). Simulating early word learning in situated connectionist agents. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society (CogSci)*, 875–881.

Hill, F., Lampinen, A., Schneider, R., Clark, S., Botvinick, M., McClelland, J. L., & Santoro, A. (2020). Environmental drivers of systematicity and generalization in a situated agent. *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Hinton, G., & LeCun, Y. (2019). The deep learning revolution. *Turing Lecture at FCRC*. https://rse.org.uk/wp-content/uploads/2021/08/Hintons-Presentation_20190718.pdf

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *6*(2), 107–116. https://doi.org/10.1142/S0218488598000094

Hochreiter, S. (2022). *Untersuchungen zu dynamischen neuronalen netzen* (Dissertation). Technische Universitat München. https://people.idsia.ch/~juergen/SeppHochreiter1991ThesisAdvisorSchmidhuber.pdf

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hofbauer, J., & Sigmund, K. (1998). *Evolutionary games and population dynamics*. Cambridge University Press. https://doi.org/10.1017/CBO9781139173179

Hollich, G. J., Hirsh-Pasek, K., & Golinkoff, R. M. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development*, *65*(3), i–vi, 1–123. https://doi.org/10.1111/1540-5834.00091

Hollis, G. (2019). Learning about things that never happened: A critique and refinement of the Rescorla-Wagner update rule when many outcomes are possible. *Memory and Cognition*, *47*, 1415–1430. https://doi.org/10.3758/s13421-019-00942-4

Holtgraves, T. M. (2008). *Language as social action: Social psychology and language use*. Lawrence Erlbaum Associates.

Holtgraves, T. M., & Kashima, Y. (2008). Language, meaning, and social cognition. *Personality and Social Psychology Review*, *12*(1), 73–94. https://doi.org/10.1177/1088868307309605

Horst, J. S., & Samuelson, L. K. (2008). Fast mapping but poor retention by 24-month-old infants. *Infancy*, *13*(2), 128–157. https://doi.org/10.1080/15250000701795598

Huang, Y. T., & Snedeker, J. (2009). Semantic meaning and pragmatic interpretation in 5-year-olds: Evidence from real-time spoken language comprehension. *Developmental Psychology*, *45*(6), 1723–1739. https://doi.org/10.1037/a0016704

Hutchins, E., & Hazlehurst, B. (1995). How to invent a lexicon: The development of shared symbols in interaction. In N. Gilbert & R. Conte (Eds.), *Artificial societies: The computer simulation of social life* (pp. 157–189). UCL Press. https://doi.org/10.4324/9780203993699

Inala, J. P., Yang, Y., Paulos, J., Pu, Y., Bastani, O., Kumar, V., Rinard, M., & Solar-Lezama, A. (2020). Neurosymbolic transformers for multi-agent communication. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Proceedings of the 34th conference on neural information processing systems (NeurIPS)* (pp. 13597–13608). Curran Associates, Inc.

Jablonka, E., Ginsburg, S., & Dor, D. (2012). The co-evolution of language and emotions. *Philosophical Transactions of the Royal Society B*, *367*, 2152–2159. https://doi.org/10.1098/rstb.2012.0117

Jackendoff, R. (1999). Possible stages in the evolution of the language capacity. *Trends in Cognitive Sciences*, *3*(7), 272–279. https://doi.org/10.1016/S1364-6613(99)01333-9

Jang, E., Gu, S., & Poole, B. (2017). Categorical reparameterization with gumbel-softmax. *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.

Jaques, N., Lazaridou, A., Hughes, E., Gülçehre, Ç., Ortega, P. A., Strouse, D., Leibo, J. Z., & de Freitas, N. (2018). Intrinsic social motivation via causal influence in multi-agent RL. *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

Johnson, M., Hofmann, K., Hutton, T., & Bignell, D. (2016). The malmo platform for artificial intelligence experimentation. *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, 4246–4247. https://doi.org/10.5555/3061053.3061259

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology*, *8*, 1726. https://doi.org/10.3389/fpsyg.2017.01726

Kachergis, G., Yu, C., & Shiffrin, R. M. (2012). An associative model of adaptive inference for learning word–referent mappings. *Psychonomic Bulletin & Review*, *19*(2), 317–324. https://doi.org/10.3758/s13423-011-0194-6

Kågebäck, M., Carlsson, E., Dubhashi, D., & Sayeed, A. (2020). A reinforcement-learning approach to efficient communication. *PLoS ONE*, *15*(7), 1–26. https://doi.org/10.1371/journal.pone.0234894

Kajic, I., Aygün, E., & Precup, D. (2020). Learning to cooperate: Emergent communication in multi-agent navigation. *arXiv preprint*, *arXiv:2004.01097*. https://doi.org/10.48550/arXiv.2004.01097

Kamath, U., Graham, K., & Emara, W. (2022). *Transformers for machine learning: A deep dive*. CRC Press. https://doi.org/10.1201/9781003170082

Kang, Y., Wang, T., & de Melo, G. (2020). Incorporating pragmatic reasoning communication into emergent language. *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, 10348–10359.

Kaur, M., & Mohta, A. (2019). A review of deep learning with recurrent neural network. *Proceedings of the 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 460–465. https://doi.org/10.1109/ICSSIT46314.2019.8987837

Kay, D. A., & Anglin, J. M. (1982). Overextension and underextension in the child's expressive and receptive speech. *Journal of Child Language*, *9*(1), 83–98. https://doi.org/10.1017/S0305000900003639

Kay, K. N. (2017). Principles for models of neural information processing. *Neuroimage*, *180*(Pt A), 101–109. https://doi.org/10.1016/j.neuroimage.2017.08.016

Kazemzadeh, S., Ordonez, V., Matten, M., & Berg, T. (2014). ReferItGame: Referring to objects in photographs of natural scenes. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 787–798. https://doi.org/10.3115/v1/D14-1086

Kemmerer, D. (2015). Are the motor features of verb meanings represented in the precentral motor cortices? yes, but within the context of a flexible, multilevel architecture for conceptual knowledge. *Psychonomic Bulletin & Review*, *22*(4), 1068–1075. https://doi.org/10.3758/s13423-014-0784-1

Kempka, M., Wydmuch, M., Runc, G., Toczek, J., & Jaśkowski, W. (2016). ViZDoom: A Doom-based AI research platform for visual reinforcement learning. *Proceedings of the 2016 IEEE Conference on Computational Intelligence and Games*, 341–348. https://doi.org/https://doi.org/10.1109/CIG.2016.7860433

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLOS Computational Biology*, *10*(11), 1–29. https://doi.org/10.1371/journal.pcbi.1003915

Kharitonov, E., & Baroni, M. (2020). Emergent language generalization and acquisition speed are not tied to compositionality. *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 11–15. https://doi.org/10.18653/v1/2020.blackboxnlp-1.2

Kharitonov, E., Chaabouni, R., Bouchacourt, D., & Baroni, M. (2019). EGG: A toolkit for research on emergence of lanGuage in games. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, 55–60. https://doi.org/10.18653/v1/D19-3010

Kiefer, M., & Pulvermüller, F. (2012). Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions [Language and the Motor System]. *Cortex*, *48*(7), 805–825. https://doi.org/10.1016/j.cortex.2011.04.006

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2019). Deep neural networks in computational neuroscience. https://oxfordre.com/neuroscience/view/10.1093/acrefore/9780190264086.001.0001/acrefore-9780190264086-e-46

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences (PNAS)*, *116*(43), 21854–21863. https://doi.org/10.1073/pnas.1905544116

Kim, J., & Oh, A. (2021). Emergent communication under varying sizes and connectivities. In A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Proceedings of the 35th conference on neural information processing systems (NeurIPS)*.

Kirby, S. (1999). *Function, selection and innateness: The emergence of language universals*. Oxford University Press. https://doi.org/10.1515/ling.2000.009

Kirby, S. (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. https://doi.org/10.1109/4235.918430

Kirby, S. (2002a). Learning, bottlenecks and the evolution of recursive syntax. In T. Briscoe (Ed.), *Linguistic evolution through language acquisition* (pp. 173–204). Cambridge University Press. https://doi.org/10.1017/CBO9780511486524.006

Kirby, S. (2002b). Natural language from artificial life. *Artificial life*, *8*(2), 185–215. https://doi.org/10.1162/106454602320184248

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language [SI: Communication and language]. *Current Opinion in Neurobiology*, *28*, 108–114. https://doi.org/10.1016/j.conb.2014.07.014

Kirby, S., & Hurford, J. R. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi & D. Parisi (Eds.), *Simulating the evolution of language* (pp. 121–147). Springer. https://doi.org/10.1007/978-1-4471-0663-0_6

Koch, C., & Laurent, G. (1999). Complexity and the nervous system. *Science*, *284*(5411), 96–98. https://doi.org/10.1126/science.284.5411.96

Kohonen, T. (1997). *Self-organizing maps* [2nd edition]. Springer. https://doi.org/10.1007/978-3-642-97966-8

Korta, K., & Perry, J. (2020). Pragmatics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2020). Metaphysics Research Lab, Stanford University.

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

Krahmer, E., & van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, *38*(1), 173–218. https://doi.org/10.1162/COLI_a_00088

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446. https://doi.org/10.1146/annurev-vision-082114-035447

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis — connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(4), 1–28. https://doi.org/10.3389/neuro.06.004.2008

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.), *Proceedings of the 36th conference on neural information processing systems (NeurIPS)*.

Kruschke, J. K. (1991). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44. https://doi.org/10.1037/h0028558

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLOS Computational Biology*, *12*, e1004896. https://doi.org/10.1371/journal.pcbi.1004896

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253. https://doi.org/10.1017/S0140525X16001837

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). Deep neural networks predict category typicality ratings for images. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society (CogSci)*, 1243–1248.

Lakoff, G. (1987). Women, fire, and dangerous things: What categories reveal about the mind.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226470993.001.0001

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought*. Basic Books. https://doi.org/10.1590/S0102-44502001000100008

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*(3), 299–321. https://doi.org/10.1016/0885-2014(88)90014-7

Lazaridou, A., & Baroni, M. (2020). Emergent multi-agent communication in the deep learning era. *arXiv preprint*, *arXiv:2006.02419*. https://doi.org/10.48550/arXiv.2006.02419

Lazaridou, A., Hermann, K. M., Tuyls, K., & Clark, S. (2018). Emergence of linguistic communication from referential games with symbolic and pixel input. *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

Lazaridou, A., Peysakhovich, A., & Baroni, M. (2017). Multi-agent cooperation and the emergence of (natural) language. *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 1–11.

Lazaridou, A., Potapenko, A., & Tieleman, O. (2020). Multi-agent communication meets natural language: Synergies between functional and structural language learning. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 7663–7674. https://doi.org/10.18653/v1/2020.acl-main.685

Le, H., Daryanto, T., Zhafransyah, F., Wijaya, D., Coppock, E., & Chin, S. (2022). Referring expressions with rational speech act framework: A probabilistic approach. *arXiv preprint*, *arXiv:2205.07795*. https://doi.org/10.48550/arxiv.2205.07795

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551. https://doi.org/10.1162/neco.1989.1.4.541

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

LeCun, Y., Cortes, C., & Burges, C. (2010). MNIST handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, *2*.

Lewis, D. (1969). *Convention: A philosophical study*. Harvard University Press. https://doi.org/10.2307/2218418

Lewis, M., Cristiano, V., Lake, B. M., Kwan, T., & Frank, M. C. (2020). The role of developmental change and linguistic experience in the mutual exclusivity effect. *Cognition*, *198*, 104191. https://doi.org/10.1016/j.cognition.2020.104191

Lewis, M., & Frank, M. C. (2013). Modeling disambiguation in word learning via multiple probabilistic constraints. *Proceedings of the 35th Annual Meeting of the Cognitive Science Society (CogSci)*, 876–881.

Li, F., & Bowling, M. (2019). Ease-of-teaching and language structure from emergent communication. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Proceedings of the 33rd conference on neural information processing systems (NeurIPS)* (pp. 1–11).

Li, P., Farkas, I., & MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural Networks*, *17*(8), 1345–1362. https://doi.org/10.1016/j.neunet.2004.07.004

Li, P., & Zhao, X. (2013). Self-organizing map models of language acquisition. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00828

Li, P., Zhao, X., & Mac Whinney, B. (2007). Dynamic self-organization and early lexical development in children. *Cognitive Science*, (4), 581–612. https://doi.org/10.1080/15326900701399905

Li, Y. (2017). Deep reinforcement learning: An overview. *arXiv preprint*, *arXiv:1701.07274*. https://doi.org/10.48550/arXiv.1701.07274

Liang, S., Li, Y., & Srikant, R. (2018). Enhancing the reliability of out-of-distribution image detection in neural networks. *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, *21*(6), 335–346.

Lindquist, K. A., & Gendron, M. (2013). What's in a word? language constructs emotion perception. *Emotion Review*, *5*(1), 66–71. https://doi.org/10.1177/1754073912451351

Lindquist, K. A., Gendron, M., Barrett, L., & Dickerson, B. C. (2014). Emotion perception, but not affect perception, is impaired with semantic memory loss. *Emotion*, *14*(2), 375–387. https://doi.org/10.1037/a0035293

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, *33*(10), 2017–2031. https://doi.org/10.1162/jocn_a_01544

Lock, A. (1980). *The guided reinvention of language*. Academic Press.

Lowe, R., Foerster, J., Boureau, Y.-L., Pineau, J., & Dauphin, Y. (2019). On the pitfalls of measuring emergent communication. *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 693–701.

Lucy, L., & Gauthier, J. (2017). Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning. *Proceedings of the First*

Workshop on Language Grounding for Robotics, 76–85. https://doi.org/10.18653/v1/W17-2810

Luo, G., Zhou, Y., Sun, X., Cao, L., Wu, C., Deng, C., & Ji, R. (2020). Multi-task collaborative network for joint referring expression comprehension and segmentation. *Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.48550/arXiv.2003.08813

Luo, R., & Shakhnarovich, G. (2017). Comprehension-guided referring expressions. *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3125–3134. https://doi.org/10.1109/CVPR.2017.333

Lupyan, G. (2008). The conceptual grouping effect: Categories matter (and named categories matter more). *Cognition*, *108*(2), 566–577. https://doi.org/10.1016/j.cognition.2008.03.009

Lupyan, G. (2012a). Chapter seven - what do words do? toward a theory of language-augmented thought. In B. H. Ross (Ed.), *The psychology of learning and motivation* (pp. 255–297). Academic Press. https://doi.org/10.1016/B978-0-12-394293-7.00007-8

Lupyan, G. (2012b). Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00054

Lupyan, G., & Lewis, M. (2019). From words-as-mappings to words-as-cues: The role of language in semantic knowledge. *Language, Cognition and Neuroscience*, *34*(10), 1319–1337. https://doi.org/10.1080/23273798.2017.1404114

Lupyan, G., Rahman, R. A., Boroditsky, L., & Clark, A. (2020). Effects of language on visual perception. *Trends in Cognitive Science*, *24*(11), 930–944. https://doi.org/10.1016/j.tics.2020.08.005

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological Science*, *18*(12), 1077–1083. https://doi.org/10.1111/j.1467-9280.2007.02028.x

Lyon, C., Nehaniv, C., & Cangelosi, A. (Eds.). (2007). *Emergence of communication and language*. Springer. https://doi.org/10.1007/978-1-84628-779-4

MacWhinney, B. (1998). Models of the emergence of language. *Annual Review of Psychology*, *49*(1), 199–227. https://doi.org/10.1146/annurev.psych.49.1.199

MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk (third edition): Volume ii: The database. *Computational Linguistics*, *26*(4), 657–657. https://doi.org/10.1162/coli.2000.26.4.657

Maddison, C., Mnih, A., & Teh, Y. (2017). The concrete distribution: A continuous relaxation of discrete random variables. *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.

Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A., & Murphy, K. (2016). Generation and comprehension of unambiguous object descriptions. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 11–20. https://doi.org/10.1109/CVPR.2016.9

Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint*, *arXiv:1801.00631*. https://doi.org/10.48550/arXiv.1801.00631

Markman, A. B., & Makin, V. S. (1998). Referential communication and category acquisition. *Journal of Experimental Psychology: General*, *127*(4), 331–354. https://doi.org/10.1037/0096-3445.127.4.331

Markman, E. M. (1989). *Categorization and naming in children*. MIT Press. https://doi.org/10.7551/mitpress/1750.001.0001

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science*, *14*(1), 57–77. https://doi.org/10.1207/s15516709cog1401_4

Markman, E. M. (1991). The whole-object, taxonomic, and mutual exclusivity assumptions as initial constraints on word meanings. In S. A. Gelman & J. P. Byrnes (Eds.), *Perspectives on language and thought: Interrelations in development* (pp. 72–106). Cambridge University Press. https://doi.org/10.1017/CBO9780511983689.004

Markman, E. M. (1992). Constraints on word learning: Speculations about their nature, origins, and domain specificity. In M. R. Gunnar & M. Maratsos (Eds.), *Modularity and constraints in language and cognition: The minnesota symposia on child psychology* (pp. 59–101). Lawrence Erlbaum Associates.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121–157. https://doi.org/10.1016/0010-0285(88)90017-5

Markman, E. M., Wasow, J. L., & Hansen, M. B. (2003). Use of the mutual exclusivity assumption by young word learners. *Cognitive Psychology*, *47*(3), 241–275. https://doi.org/10.1016/S0010-0285(03)00034-3

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Henry Holt; Co., Inc. https://doi.org/10.7551/mitpress/9780262514620.001.0001

Marstaller, L., Hintze, A., & Adami, C. (2013). The Evolution of Representation in Simple Cognitive Networks. *Neural Computation*, *25*(8), 2079–2107. https://doi.org/10.1162/NECO_a_00475

Martin, A., Onishi, K. H., & Vouloumanos, A. (2012). Understanding the abstract role of speech in communication at 12 months. *Cognition*, *123*(1), 50–60. https://doi.org/10.1016/j.cognition.2011.12.003

Maynard Smith, J. (1974). The theory of games and the evolution of animal conflicts. *Journal of Theoretical Biology*, *47*(1), 209–221. https://doi.org/10.1016/0022-5193(74)90110-6

Mayor, J., & Plunkett, K. (2010). A neuro-computational account of taxonomic responding and fast mapping in early word learning. *Psychological Review*, *117*(1), 1–31. https://doi.org/10.1037/a0018130

McClelland, J. L. (2010). Emergence in cognitive science. *Topics in Cognitive Science*, *2*(4), 751–770. https://doi.org/10.1111/j.1756-8765.2010.01116.x

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition (2010/07/02). *Trends in Cognitive Sciences*, *14*(8), 348–356. https://doi.org/10.1016/j.tics.2010.06.002

McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., & Schütze, H. (2020). Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences*, *117*(42), 25966–25974. https://doi.org/10.1073/pnas.1910416117

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.). Academic Press. https://doi.org/10.1016/S0079-7421(08)60536-8

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*(4), 831–877. https://doi.org/10.1037/a0029872

McShane, J. (1979). The development of naming. *Linguistics*, 879–905.

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., & Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, *118*(8), e2011417118. https://doi.org/10.1038/s41583-020-0277-3

Mervis, C. B., Golinkoff, R. M., & Bertrand, J. (1994). Two-year-olds readily learn multiple labels for the same basic-level category. *Child Development*, *65*(4), 1163–1177. https://doi.org/10.2307/1131312

Miikkulainen, R. (1997a). Dyslexic and category-specific aphasic impairments in a self-organizing feature map model of the lexicon. *Brain and Language*, *59*(2), 334–366. https://doi.org/10.1006/brln.1997.1820

Miikkulainen, R. (1997b). Natural language processing with subsymbolic neural networks. In A. Browne (Ed.), *Neural network perspectives on cognition and adaptive robotics* (pp. 120–139). Institute of Physics Publishing.

Mikolov, T., Joulin, A., & Baroni, M. (2015). A roadmap towards machine intelligence. *arXiv preprint*, *arXiv:1511.08130*. https://doi.org/10.48550/arXiv.1511.08130

Mirolli, M., & Parisi, D. (2005). How can we explain the emergence of a language that benefits the hearer but not the speaker? *Connection Science*, *17*(3–4), 307–324. https://doi.org/10.1080/09540090500177539

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., & Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. *Proceedings of The 33rd International Conference on Machine Learning (ICML)*, *48*, 1928–1937.

Monroe, W., Hawkins, R. X. D., Goodman, N. D., & Potts, C. (2017). Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, *5*, 325–338. https://doi.org/10.1162/tacl_a_00064

Monroe, W., & Potts, C. (2015). Learning in the Rational Speech Acts model. *arXiv preprint*, *arXiv:1510.06807*. https://doi.org/10.48550/arXiv.1510.06807

Moradi, R., Berangi, R., & Minaei, B. A. (2020). A survey of regularization strategies for deep models. *Artificial Intelligence Review*, *53*, 3947–2986. https://doi.org/10.1007/s10462-019-09784-7

Mordatch, I., & Abbeel, P. (2018). Emergence of grounded compositional language in multi-agent populations. *Proceedings of the 32nd AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and 8th AAAI Symposium on Educational Advances in Artificial Intelligence*, 1495–1502. https://doi.org/10.5555/3504035.3504218

Najnin, S., & Banerjee, B. (2018). Pragmatically framed cross-situational noun learning using computational reinforcement models. *Frontiers in Psychology*, *9*(5), 1–18. https://doi.org/10.3389/fpsyg.2018.00005

Newell, A. (1980). Physical symbol systems. *Cognitive Science*, *4*(2), 135–183. https://doi.org/10.1016/S0364-0213(80)80015-2

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, *19*(3), 113–126. https://doi.org/10.1145/360018.360022

Nie, A., Cohn-Gordon, R., & Potts, C. (2020). Pragmatic issue-sensitive image captioning. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1924–1938. https://doi.org/10.18653/v1/2020.findings-emnlp.173

Niedenthal, P. M. (2007). Embodying emotion. *Science*, *316*(5827), 1002–1005. https://doi.org/10.1126/science.1136930

Noë, A. (2004). *Action in perception*. MIT Press.

Nowak, M. A., & Krakauer, D. C. (1999). The evolution of language. *Proceedings of the National Academy of Sciences*, *96*(14), 8028–8033. https://doi.org/10.1073/pnas.96.14.8028

Nowak, M. A., Plotkin, J. B., & Jansen, V. A. A. (2000). The evolution of syntactic communication. *Nature*, *404*(6777), 495–498. https://doi.org/10.1038/35006635

Ohmer, X., Franke, M., & König, P. (2022). Mutual exclusivity in pragmatic agents. *Cognitive Science*, *46*(1), e13069. https://doi.org/10.1111/cogs.13069

Ohmer, X., König, P., & Franke, M. (2020). Reinforcement of semantic representations in pragmatic agents leads to the emergence of a mutual exclusivity bias. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society (CogSci)*, 1779–1785.

Ohmer, X., Marino, M., König, P., & Franke, M. (2021). Why and how to study the impact of perception on language emergence in artificial agents. *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society (CogSci)*, 1139–1145.

Orban, G. A., van Essen, D., & Vanduffel, W. (2004). Comparative mapping of higher visual areas in monkeys and humans. *Trends in Cognitive Sciences*, *8*(7), 315–324. https://doi.org/10.1016/j.tics.2004.05.009

Otter, D. W., Medina, J. R., & Kalita, J. K. (2021). A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(2), 604–624. https://doi.org/10.1109/TNNLS.2020.2979670

Ozgen, E., & Davies, I. R. L. (2002). Acquisition of categorical color perception: A perceptual learning approach to the linguistic relativity hypothesis. *Journal of Experimental Psychology: General*, *131*(4), 477–493. https://doi.org/10.1037//0096-3445.131.4.477

Pecher, D., & Zwann, R. (2005). *Grounding cognition: The role of perception and action in memory, language, and thinking*. Cambridge University Press. https://doi.org/10.1017/CBO9780511499968

Peng, H., Khashabi, D., & Roth, D. (2015). Solving hard coreference problems. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 809–819. https://doi.org/10.3115/v1/N15-1082

Perlovsky, L. (2009). Language and cognition. *Neural Networks*, *22*(3), 247–257. https://doi.org/10.1016/j.neunet.2009.03.007

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science*, *42*(8), 2648–2669. https://doi.org/10.1111/cogs.12670

Plunkett, K., Sinha, C., Møller, M. F., & Strandsby, O. (1992). Symbol grounding or the emergence of symbols? Vocabulary growth in children and a connectionist net. *Connection Science*, *4*(3 & 4), 293–312. https://doi.org/10.1080/09540099208946620

Portelance, E., Frank, M. C., Jurafsky, D., Sordoni, A., & Laroche, R. (2021). The emergence of the shape bias results from communicative efficiency. *Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL)*, 607–623. https://doi.org/10.18653/v1/2021.conll-1.48

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3), 353–363. https://doi.org/10.1037/h0025953

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology*, *83*(2, Pt.1), 304–308. https://doi.org/10.1037/h0028558

Pulvermüller, F. (2013). How neurons make meaning: Brain mechanisms for embodied and abstract-symbolic semantics. *Trends in Cognitive Sciences*, *17*(9), 458–470. https://doi.org/10.1016/j.tics.2013.06.004

Pulvermüller, F. (2018). Neural reuse of action perception circuits for language, concepts, and communication. *Progress in Neurobiology*, *160*, 1–44. https://doi.org/10.1016/j.pneurobio.2017.07.001

Qing, C., & Franke, M. (2015). Variations on a bayesian theme: Comparing bayesian models of referential reasoning. In H. Zeevat & H.-C. Schmitz (Eds.), *Bayesian natural language semantics and pragmatics* (pp. 201–220). Springer. https://doi.org/10.1007/978-3-319-17064-0_9

Quine, W. V. O. (1960). *Word and object: An inquiry into the linguistic mechanisms of objective reference*. MIT Press.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*.

Ramscar, M., Dye, M., & Klein, J. (2013). Children value informativity over logic in word learning. *Psychological Science*, *24*(6), 1017–1023. https://doi.org/10.1177/0956797612460691

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, *34*(6), 909–957. https://doi.org/10.1111/j.1551-6709.2009.01092.x

Rangwala, M., & Williams, R. (2020). Learning multi-agent communication through structured attentive reasoning. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Proceedings of the 34th conference on neural information processing systems (NeurIPS)* (pp. 10088–10098).

Rapan, L., Niu, M., Zhao, L., Funck, T., Amunts, K., Zilles, K., & Palomero-Gallagher, N. (2022). Receptor architecture of macaque and human early visual areas: Not equal, but comparable. *Brain Structure and Function*, *227*, 1247–1263. https://doi.org/10.1007/s00429-021-02437-y

Reggia, J. A., Schulz, R., Wilkinson, G. S., & Uriagereka, J. (2001). Conditions enabling the evolution of inter-agent signaling in an artificial world. *Artificial Life*, *7*(1), 3–32. https://doi.org/10.1162/106454601300328007

Regier, T. (2003). Emergent constraints on word-learning: A computational perspective. *Trends in Cognitive Sciences*, *7*, 263–268. https://doi.org/10.1016/S1364-6613(03)00108-6

Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive Science*, *29*(6), 819–865. https://doi.org/10.1207/s15516709cog0000_31

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences (PNAS)*, *104*(4), 1436–1441. https://doi.org/10.1073/pnas.0610341104

Regier, T., & Xu, Y. (2017). The Sapir-Whorf hypothesis and inference under uncertainty. *WIREs Cognitive Science*, *8*(6), e1440. https://doi.org/10.1002/wcs.1440

Reiter, E., & Dale, R. (1992). A fast algorithm for the generation of referring expressions. *Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, 232–238. https://doi.org/10.3115/992066.992105

Ren, Y., Guo, S., Labeau, M., Cohen, S. B., & Kirby, S. (2020). Compositional languages emerge in a neural iterated learning model. *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Rescorla, M. (2019). Convention. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (summer edition). https://plato.stanford.edu/archives/sum2019/entries/convention/

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., ... Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770. https://doi.org/10.1038/s41593-019-0520-2

Rita, M., Strub, F., Grill, J.-B., Pietquin, O., & Dupoux, E. (2022). On the role of population heterogeneity in emergent communication. *Proceedings of the 10th International Conference on Learning Representations (ICLR)*.

Roberson, D., Pak, H., & Hanley, J. (2008). Categorical perception of colour in the left and right visual field is verbally mediated: Evidence from Korean. *Cognition*, *107*, 752–762. https://doi.org/10.1016/j.cognition.2007.09.001

Roberts, C. (2012). Information structure in discourse: Towards an integrated theory of pragmatics. *Semantics & Pragmatics*, *5*(6), 1–69. https://doi.org/10.3765/sp.5.6

Robins, A. (1993). Neural networks and models of cognition: A review. *Proceedings of the 1st New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems*, 63–64. https://doi.org/10.1109/ANNES.1993.323081

Rodríguez Luna, D., Ponti, E. M., Hupkes, D., & Bruni, E. (2020). Internal and external pressures on language emergence: Least effort, object constancy and frequency. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 4428–4437. https://doi.org/10.18653/v1/2020.findings-emnlp.397

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605. https://doi.org/10.1016/0010-0285(75)90024-9

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439. https://doi.org/10.1016/0010-0285(76)90013-X

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint*, *arXiv:1609.04747*. https://doi.org/10.48550/arXiv.1609.04747

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition, volume 1: Foundations & volume 2: Psychological and biological models*. MIT Press. https://doi.org/mitpress/5236.003.0001

Samaha, J., Boutonnet, B., Postle, B. R., & Lupyan, G. (2018). Effects of meaningfulness on perception: Alpha-band oscillations carry perceptual expectations and influence early visual responses. *Scientific Reports*, *8*(6606), 1–14. https://doi.org/10.1038/s41598-018-25093-5

Sandholm, W. H. (2010). *Population games and evolutionary dynamics*. MIT Press.

Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., & Batra, D. (2019). Habitat: A platform for embodied ai research. *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/ICCV.2019.00943

Schelling, T. C. (1960). *The strategy of conflict*. Harvard University Press. https://doi.org/10.2307/2126712

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

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

Schuster, P., & Sigmund, K. (1983). Replicator dynamics. *Journal of Theoretical Biology*, *100*(3), 533–538. https://doi.org/10.1016/0022-5193(83)90445-9

Scontras, G., Tessler, M. H., & Franke, M. (2018). *Probabilistic language understanding: An introduction to the Rational Speech Act framework*. http://www.problang.org

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University Press. https://doi.org/10.1017/CBO9781139173438

Searle, J. R. (1979). *Expression and meaning: Studies in the theory of speech acts*. Cambridge University Press. https://doi.org/10.1017/CBO9780511609213

Shafto, P., Goodman, N. D., & Frank, M. C. (2012). Learning from others: The consequences of psychological reasoning for human learning. *Perspectives on Psychological Science*, *7*(4), 341–351. https://doi.org/10.1177/1745691612448481

Shapiro, L., & Spaulding, S. (2021). Embodied cognition. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2021). Metaphysics Research Lab, Stanford University.

Shapiro, L., & Stockman, G. (2002). *Computer vision*. Prentice-Hall. https://doi.org/10.5555/558008

Shen, S., Fried, D., Andreas, J., & Klein, D. (2019). Pragmatically informative text generation. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 4060–4067. https://doi.org/10.18653/v1/N19-1410

Singh, A., Jain, T., & Sukhbaatar, S. (2019). Individualized controlled continuous communication model for multiagent cooperative and competitive tasks. *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings [Compositional Language Acquisition]. *Cognition*, *61*(1), 39–91. https://doi.org/10.1016/S0010-0277(96)00728-7

Skyrms, B. (1996). *Evolution of the social contract*. Cambridge University Press. https://doi.org/10.1515/jso-2019-0041

Skyrms, B. (2010). *Signals: Evolution, learning, and information*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199580828.001.0001

Sloutsky, V., & Deng, W. (2019). Categories, concepts, and conceptual development. *Language, Cognition and Neuroscience*, *34*(10), 1284–1297. https://doi.org/10.1080/23273798.2017.1391398

Sloutsky, V. M. (2003). The role of similarity in the development of categorization. *Trends in Cognitive Sciences*, *7*(6), 246–251. https://doi.org/10.1016/s1364-6613(03)00109-8

Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology: General*, *133*(2), 166–188. https://doi.org/10.1037/0096-3445.133.2.166

Sloutsky, V. M., & Fisher, A. V. (2012). Linguistic labels: Conceptual markers or object features? *Journal of Experimental Child Psychology*, *111*(1), 65–86. https://doi.org/10.1016/j.jecp.2011.07.007

Sloutsky, V. M., Lo, Y.-F., & Fisher, A. V. (2001). How much does a shared name make things similar? linguistic labels, similarity, and the development of inductive inference. *Child Development*, *72*(6), 1695–1709. https://doi.org/10.1111/1467-8624.00373

Smith, K., Brighton, H., & Kirby, S. (2003). Complex systems in language evolution: The cultural emergence of compositional structure. *Advances in Complex Systems*, *6*(4), 537–558. https://doi.org/10.1142/S0219525903001055

Smith, K., Smith, A. D. M., & Blythe, R. A. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, *35*(3), 480–498. https://doi.org/10.1111/j.1551-6709.2010.01158.x

Smith, L. B. (1989). From global similarities to kinds of similarities: The construction of dimensions in development. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 146–178). Cambridge University Press. https://doi.org/10.1017/CBO9780511529863.008

Smith, L. B., & Samuelson, L. (2006). An attentional learning account of the shape bias: Reply to cimpian and markman (2005) and booth, waxman, and huang (2005). *Developmental Psychology*, *42*(6), 1339–1343. https://doi.org/10.1037/0012-1649.42.6.1339

Smith, N. J., Goodman, N. D., & Frank, M. C. (2013). Learning and using language via recursive pragmatic reasoning about other agents. *Proceedings of the 27th Conference on Neural Information Processing Systems (NeurIPS)*, 3039–3047.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition*. Blackwell Publishers.

Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent convolutional neural networks: A better model of biological object recognition. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01551

Steels, L. (1995). A Self-Organizing Spatial Vocabulary. *Artificial Life*, *2*(3), 319–332. https://doi.org/10.1162/artl.1995.2.3.319

Steels, L. (1997). The synthetic modeling of language origins. *Evolution of Communication*, *1*(1), 1–34. https://doi.org/10.1075/eoc.1.1.02ste

Steels, L. (1998). Synthesising the origins of language and meaning using co-evolution, self-organisation and level formation. In J. R. Hurford, M. Studdert-Kennedy, & C. Knight (Eds.), *Approaches to the evolution of language* (pp. 384–404). Cambridge University Press.

Steels, L. (2001). Language games for autonomous robots. *IEEE Intelligent Systems*, *16*(5), 16–22. https://doi.org/10.1109/MIS.2001.956077

Steels, L. (2003). Evolving grounded communication for robots. *Trends in cognitive sciences*, *7*(7), 308–312. https://doi.org/10.1016/s1364-6613(03)00129-3

Steels, L. (2005). The emergence and evolution of linguistic structure: From lexical to grammatical communication systems. *Connection Science*, *17*(3–4), 213–230. https://doi.org/10.1080/09540090500269088

Steels, L. (2012). Grounding language through evolutionary language games. In L. Steels & M. Hild (Eds.), *Language grounding in robots* (pp. 1–22). Springer. https://doi.org/10.1007/978-1-4614-3064-3_1

Steels, L., & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, *28*(4), 469–489. https://doi.org/10.1017/S0140525X05000087

Steels, L., & Kaplan, F. (1999). Situated grounded word semantics. *Proceedings of the 16th nternational Joint Conference on Artificial Intelligence (IJCAI)*, 862–867. https://doi.org/10.5555/1624312.1624342

Steels, L., & Vogt, P. (1997). Grounding adaptive language games in robotic agents. In P. Husbands & I. Harvey (Eds.), *Proceedings of the 4th European conference on artificial life (ECAL)* (pp. 474–482). MIT Press. https://doi.org/https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.588.100&rep=rep1&type=pdf

Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2021). Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of Cognitive Neuroscience*, *33*(10), 2044–2064. https://doi.org/10.1162/jocn_a_01755

Suffill, E., Branigan, H., & Pickering, M. (2019). Novel labels increase category coherence, but only when people have the goal to coordinate. *Cognitive Science*, *43*(11), e12796. https://doi.org/10.1111/cogs.12796

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, 843–852. https://doi.org/10.1109/ICCV.2017.97

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). The MIT Press. https://doi.org/10.5555/3312046

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 1–10.

Taylor, P. D., & Jonker, L. B. (1978). Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, *40*(1), 145–156. https://doi.org/10.1016/0025-5564(78)90077-9

Tomasello, M. (2001). Perceiving intentions and learning words in the second year of life. In M. Bowerman & S. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 132–158). Cambridge University Press. https://doi.org/10.1017/CBO9780511620669.007

Tuli, S., Dasgupta, I., Grant, E., & Griffiths, T. L. (2021). Are convolutional neural networks or transformers more like human vision? *arXiv preprint*, *arXiv:2105.07197*. https://doi.org/10.48550/arXiv.2105.07197

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research (JMLR)*, *9*, 2579–2605.

van der Wal, O., de Boer, S., Bruni, E., & Hupkes, D. (2020). The grammar of emergent languages. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3339–3359. https://doi.org/10.18653/v1/2020.emnlp-main.270

Vong, W. K., & Lake, B. M. (2020). Learning word-referent mappings and concepts from raw inputs. *arXiv preprint*, *arXiv:2003.05573*. https://doi.org/10.48550/arXiv.2003.05573

Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational Intelligence and Neuroscience*, *2018*, 1–13. https://doi.org/10.1155/2018/7068349

Vouloumanos, A., Onishi, K. H., & Pogue, A. (2012). Twelve-month-old infants recognize that speech can communicate unobservable intentions. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, *109*(32), 12933–12937. https://doi.org/10.1073/pnas.1121057109

Wagner, K. (2000). Cooperative strategies and the evolution of communication. *Artificial Life*, *6*(2), 149–179. https://doi.org/10.1162/106454600568384

Wagner, K., Reggia, J. A., Uriagereka, J., & Wilkinson, G. S. (2003). Progress in the simulation of emergent communication and language. *Adaptive Behavior*, *11*(1), 37–69. https://doi.org/10.1177/10597123030111003

Warlaumont, A. S., Westermann, G., Buder, E. H., & Oller, D. K. (2013). Prespeech motor learning in a neural network using reinforcement. *Neural Networks*, *38*, 64–75. https://doi.org/10.1016/j.neunet.2012.11.012

Waxman, S. R., & Gelman, S. A. (2009). Early word-learning entails reference, not merely associations. *Trends in Cognitive Sciences*, *13*(6), 258–263. https://doi.org/10.1016/j.tics.2009.03.006

Weijters, A. J. M. M., & Hoppenbrouwers, G. A. J. (1995). Backpropagation networks for grapheme-phoneme conversion: A non-technical introduction. In P. J. Braspenning, F. Thuijsman, & A. J. M. M. Weijters (Eds.), *Artificial neural networks: An introduction to ANN theory and practice*. https://doi.org/10.1007/BFb0027021

Westermann, G., & Mareschal, D. (2012). Mechanisms of developmental change in infant categorization [The Potential Contribution of Computational Modeling to the Study of Cognitive Development: When, and for What Topics?]. *Cognitive Development*, *27*(4), 367–382. https://doi.org/10.1016/j.cogdev.2012.08.004

Westermann, G., & Mareschal, D. (2014). From perceptual to language-mediated categorization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 20120391. https://doi.org/10.1098/rstb.2012.0391

Westermann, G., & Miranda, E. R. (2004). A new model of sensorimotor coupling in the development of speech. *Brain & Language*, *89*(2), 393–400. https://doi.org/10.1016/S0093-934X(03)00345-6

Westermann, G., & Twomey, K. (2017). Computational models of word learning. In G. Westermann & N. Mani (Eds.), *Early word learning* (pp. 138–154). Routledge. https://doi.org/10.1093/oxfordhb/9780195376746.013.0031

Willems, R. M., Labruna, L., D'Esposito, M., Ivry, R., & Casasanto, D. (2011). A functional role for the motor system in language understanding: Evidence from theta-burst transcranial magnetic stimulation. *Psychological Science*, *22*(7), 849–854. https://doi.org/10.1177/0956797611412387

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, *8*, 229–256. https://doi.org/10.1007/BF00992696

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, *104*(19), 7780–7785. https://doi.org/10.1073/pnas.0701644104

Winograd, T. (1972). Understanding natural language. *Cognitive Psychology*, *3*(1), 1–191. https://doi.org/10.1016/0010-0285(72)90002-3

Wittgenstein, L. (1953). *Philosophical investigations*. MacMillan. https://doi.org/10.2307/2217461

Witzel, C., & Gegenfurtner, K. R. (2015). Categorical facilitation with equally discriminable colors. *Journal of Vision*, *15*(8), 22. https://doi.org/10.1167/15.8.22

Wolff, P., & Holmes, K. J. (2011). Linguistic relativity. *WIREs Cognitive Science*, *2*(3), 253–265. https://doi.org/10.1002/wcs.104

Wu, Y., Wu, Y., Gkioxari, G., & Tian, Y. (2018). Building generalizable agents with a realistic and rich 3d environment. *arXiv preprint*, *arXiv:1801.02209*. https://doi.org/10.48550/arXiv.1801.02209

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological Review*, *114*(2), 245–272. https://doi.org/10.1037/0033-295X.114.2.245

Yoshikawa, Y., Asada, M., Hosoda, K., & Koga, J. (2003). A constructivist approach to infants' vowel acquisition through mother–infant interaction. *Connection Science*, *15*(4), 245–258. https://doi.org/10.1080/09540090310001655075

Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., & Berg, T. L. (2018). Mattnet: Modular attention network for referring expression comprehension. *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1307–1315. https://doi.org/10.1109/CVPR.2018.00142

Yu, L., Poirson, P., Yang, S., Berg, A. C., & Berg, T. L. (2016). Modeling context in referring expressions. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Proceedings of the 14th european conference on computer vision (ECCV)* (pp. 69–85). https://doi.org/10.1007/978-3-319-46475-6_5

Yuan, A., Monroe, W., Bai, Y., & Kushman, N. (2018). Understanding the Rational Speech Act model. *Proceedings of the 40th Annual Meeting of the Cognitive Science Society (CogSci)*, 2759–2764.

Yuan, L., Fu, Z., Shen, J., Xu, L., Shen, J., & Zhu, S.-C. (2020). Emergence of pragmatics from referential game between theory of mind agents. *arXiv preprint*, *arXiv:2001.07752*. https://doi.org/10.48550/arXiv.2001.07752

Yurovsky, D., Yu, C., & Smith, L. B. (2013). Competitive processes in cross-situational word learning. *Cognitive Science*, *37*(5), 891–921. https://doi.org/10.1111/cogs.12035

Zarr, N., Ferguson, R., & Glenberg, A. M. (2013). Language comprehension warps the mirror neuron system. *Frontiers in Human Neuroscience*, *7*, 1–5. https://doi.org/10.3389/fnhum.2013.00870

Zarrieß, S., & Schlangen, D. (2019). Know what you don't know: Modeling a pragmatic speaker that refers to objects of unknown categories. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 654–659. https://doi.org/10.18653/v1/P19-1063

Zettersten, M., & Lupyan, G. (2020). Finding categories through words: More nameable features improve category learning. *Cognition*, *196*, 104135. https://doi.org/10.1016/j.cognition.2019.104135

Zhou, K., Mo, L., Kay, P., Kwok, V. P. Y., Ip, T. N. M., & Tan, L. H. (2010). Newly trained lexical categories produce lateralized categorical perception of color. *Proceedings of the National Academy of Sciences*, *107*(22), 9974–9978. https://doi.org/10.1073/pnas.1005669107

Zhu, Y., Mottaghi, R., Kolve, E., Lim, J. J., Gupta, A., Fei-Fei, L., & Farhadi, A. (2017). Target-driven visual navigation in indoor scenes using deep reinforcement learning. *Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA)*, 3357–3364.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley. https://doi.org/10.1002/1097-4679(195007)6:3<306::AID-JCLP2270060331>3.0.CO;2-7

Erklärung über die Eigenständigkeit der erbrachten wissenschaftlichen Leistung

Ich erkläre hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet.

Bei der Auswahl und Auswertung folgenden Materials haben mir die nachstehend aufgeführten Personen in der jeweils beschriebenen Weise unentgeltlich geholfen.

- Die Veröffentlichung in Kapitel 2 wurde von Michael Franke und Peter König in regelmäßigen Treffen betreut, mitkonzipiert, diskutiert und schriftlich überarbeitet.
- Die Veröffentlichung in Kapitel 3 wurde von Elia Bruni und Marko Duda in regelmäßigen Treffen mitkonzipiert und diskutiert. Elia Bruni hat das Projekt betreut und bei der schriftlichen Überarbeitung geholfen. Marko Duda hat bei der Implementation des Modells sowie bei der Entwicklung und Implementation der Analysen mitgewirkt.
- Die Veröffentlichung in Kapitel 4 wurde von Michael Franke und Peter König betreut. Sie wurde von Michael Franke, Peter König und Michael Marino in regelmäßigen Treffen mitkonzipiert, diskutiert und schriftlich überarbeitet. Michael Marino hat das *relational label smoothing* entwickelt sowie die CNNs implementiert und trainiert, die später in den Sprachspielen als Module der visuellen Wahrnehmung genutzt wurden.
- Alle restlichen Teile der Dissertation (Abstract, General introduction, Lay summaries, General discussion, General appendix) wurden von mir allein verfasst. Der Inhalt wurde mit meinen Betreuern besprochen und ihr Feedback wurde von mir integriert. Außerdem haben Sören Becker, Britta Grusdt, Henrik Löfberg, und Charlotte Ohmer Feedback zur Verständlichkeit des Texts und zur sprachlichen Richtigkeit gegeben, das ich ebenfalls berücksichtigt habe.

Weitere Personen waren an der inhaltlichen materiellen Erstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich hierfür nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten (Promotionsberater oder andere Personen) in Anspruch genommen. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

....

(Ort, Datum) (Unterschrift)