Humans Learn Language from Situated Communicative Interactions. What about Machines?

Katrien Beuls*
Université de Namur, Belgium
Faculté d'informatique
katrien.beuls@unamur.be

Paul Van Eecke*
Vrije Universiteit Brussel, Belgium
Artificial Intelligence Laboratory
paul@ai.vub.ac.be

Humans acquire their native languages by taking part in communicative interactions with their caregivers. These interactions are meaningful, intentional, and situated in their everyday environment. The situated and communicative nature of the interactions is essential to the language acquisition process, as language learners depend on clues provided by the communicative environment to make sense of the utterances they perceive. As such, the linguistic knowledge they build up is rooted in linguistic forms, their meaning, and their communicative function. When it comes to machines, the situated, communicative, and interactional aspects of language learning are often passed over. This applies in particular to today's large language models (LLMs), where the input is predominantly text-based, and where the distribution of character groups or words serves as a basis for modeling the meaning of linguistic expressions. In this article, we argue that this design choice lies at the root of a number of important limitations, in particular regarding the data hungriness of the models, their limited ability to perform human-like logical and pragmatic reasoning, and their susceptibility to biases. At the same time, we make a case for an alternative approach that models how artificial agents can acquire linguistic structures by participating in situated communicative interactions. Through a selection of experiments, we show how the linguistic knowledge that is captured in the resulting models is of a fundamentally different nature than the knowledge captured by LLMs and argue that this change of perspective provides a promising path towards more human-like language processing in machines.

* Both authors contributed equally.

Action Editors: Marianna Apidianaki, Abdellah Fourtassi, and Sebastian Padó. Submission received: 16 December 2023; revised version received: 30 April 2024; accepted for publication: 19 June 2024.

https://doi.org/10.1162/coli_a_00534

1. Humans Learn Language from Situated Communicative Interactions

Human languages are evolutionary systems that continuously adapt to changes in the communicative needs and environment of their users (Schleicher 1869; Darwin 1871; Maynard Smith and Szathmáry 1999). They emerge and evolve as a result of meaningful and intentional communicative interactions between members of a linguistic community. When children acquire their native languages as new members of such a community, they build up their linguistic knowledge by actively taking part in communicative interactions with their caregivers. They thereby face two challenges that can be considered fundamental to the language acquisition process. On the one hand, children need to work out the communicative intents of their interlocutors, effectively making sense of the utterances they perceive. On the other hand, based on these utterances and their "reconstructed" meanings, they need to be able to bootstrap a linguistic system that enables them to understand and produce utterances that they have never observed before.

The cognitive processes involved in the acquisition of language from situated communicative interactions have been extensively studied in the fields of developmental psychology and cognitive linguistics, where they form the basis of constructivist and usage-based theories of language acquisition (Piaget 1923; Boden 1978; Bruner 1983; Langacker 1987; Nelson 1998; Croft 1991; Givón 1995; Clark 1996; Tomasello 2003; Goldberg 2006; Bybee 2010; Lieven 2014; MacWhinney 2014; Diessel 2017; Behrens 2021). The ability of children to acquire language is thereby attributed to two broad sets of skills, for which Tomasello (2003, pages 3–4) coined the terms **intention reading** and **pattern finding**. Intention reading refers to the cognitive abilities that are concerned with the functional, meaningful dimension of linguistic communication. These include in particular the capacity of children to share attention, to follow and direct the attention of others through non-linguistic gestures like pointing, and to recognize the communicative intents of their interlocutors. Pattern finding refers to the cognitive abilities that enable language users to generalize across different communicative interactions. These concern in particular the capacity to recognize similarities and differences in sensory-motor experiences, and to use this capacity for perceptual and conceptual categorization, schema formation, frequency-based distributional reasoning, and analogical thinking.

Intention reading and pattern finding are complementary yet highly interdependent skills. Intention reading enables a language user to reconstruct the intended meaning of an observed utterance. It implements an abductive reasoning process by which the language user constructs a hypothesis about the communicative intents of their interlocutor. This hypothesis is constructed based on clues that are provided by the utterance on the one hand, and its situational embedding in a communicative environment on the other. Pattern finding then provides the ability to generalize over observed utterances and their reconstructed meanings. It implements an inductive reasoning process, which, over time, yields an inventory of productive schemata that constitute the linguistic knowledge of a language user. As small children start out without any prior linguistic knowledge, their intention reading process can only rely on environmental clues. For example, during a communicative interaction with a caregiver, a young child might hypothesize that the observed utterance *bear-gone* refers to the observed disappearance from sight of their favorite stuffed animal. This association can be made based on sensory experience and pre-linguistic reasoning only, without any prior knowledge about the compositional nature of the utterance. An association between an observed form and its hypothesized meaning that is reconstructed based on environmental clues only is referred to as a **holophrastic construction**. When the

same child later hears the utterance *ball-gone* in the context of a certain spherical toy disappearing from sight, their pattern finding ability might enable them to infer the compositional structure of the observed utterance. Indeed, based on the previously acquired holophrastic construction that associates *bear-gone* to their favorite stuffed animal disappearing from sight, and the observation of the utterance *ball-gone* in the context of a certain spherical toy disappearing from sight, the child might infer through syntactico-semantic generalization that the form *bear* is associated with their favorite stuffed animal, that the form *ball* is associated with a certain spherical toy, and that the form *X-gone* is associated with the referent of *X* disappearing from sight. The associations of *bear* and *ball* with their respective meanings are called **holistic constructions**. Constructions that contain abstract slots, such as the association between *X-gone* and its meaning of *X*'s referent disappearing from sight, are referred to as **item-based constructions**. The constructions that emerge from the pattern finding process can in turn provide clues for the process of intention reading during future communicative interactions. For example, if the child later observes the utterance *gimme-bear* in the context of a request by a caregiver to hand over their favorite stuffed animal, they can recognize the form *bear*, relate it to its meaning, and hypothesize that *gimme-X* is associated to a request by their caregiver to hand over the referent of *X*. The interaction between the processes of intention reading and pattern finding is thus bidirectional. Pattern finding relies on intention reading for providing hypotheses about the intended meaning of observed utterances, while intention reading relies on constructions that result from pattern finding to constrain and navigate the space of possible meaning hypotheses. The inter-dependency between the processes of intention reading and pattern finding is schematically depicted in Figure 1, using the examples introduced above.

The abductive nature of the intention reading process entails that the meaning that a language learner attributes to an observed utterance has the epistemological status of a hypothesis. This hypothesis is likely and plausible given the linguistic and non-linguistic knowledge of the language learner on the one hand, and the communicative environment in which the utterance was observed on the other. Hypotheses resulting from intention reading are thereby not free from uncertainty or doubt, and might in some cases not even transfer to any other communicative situation. An example of this uncertainty, rooted in anecdotal evidence, would be the case of a Flemish toddler (2Y2M) occasionally observing the utterance *goed-weekend* (lit. *good-weekend*—Eng. *Enjoy your weekend*) at the end of a conversation between their parents and caregivers when leaving their day-care center. At home, the toddler starts to reuse the phrase as a farewell greeting whenever they leave the room for a few minutes, for example to fetch another toy or garment. The hypothesis of the toddler about the meaning and communicative function of the phrase *goed-weekend* is only partially accurate. While they clearly have managed to reconstruct certain aspects of its meaning, in particular its social function as a farewell greeting, they did not capture the compositionality of the phrase and its meaning of wishing the addressee an enjoyable Saturday and Sunday.

The uncertainty introduced by the intention reading process directly percolates to the constructions that are formed, whether they are holophrastic mappings or the result of a syntactico-semantic generalization process. As a language learner uses these constructions across a multitude of communicative interactions that are situated in a variety of environments, they refine their hypotheses and gradually deepen the entrenchment of constructions that generalize better across situations. At the same time, constructions that are in conflict or competition with more generally applicable constructions, thereby hindering successful communication rather than supporting it, will gradually assume a different function or disappear altogether from the language learner's inventory of

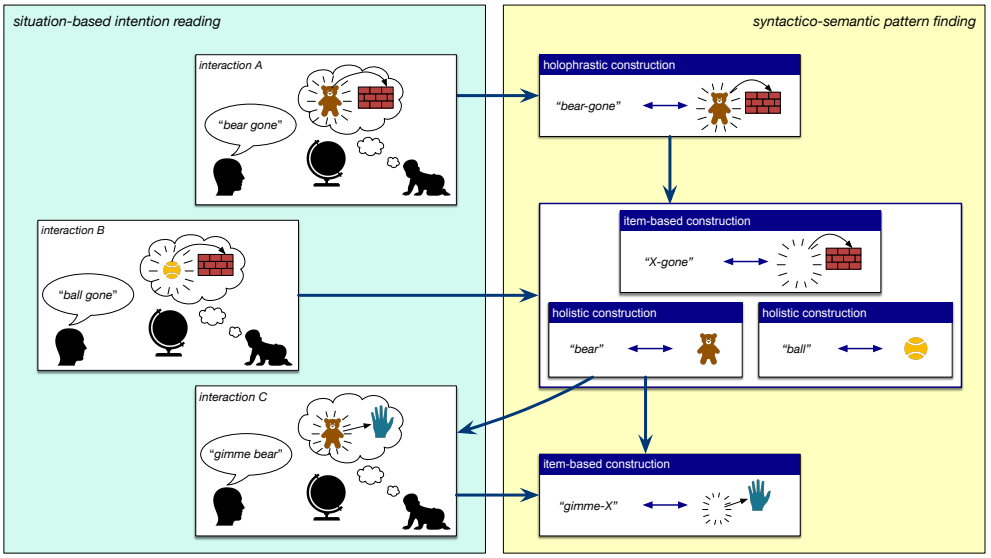

Figure 1
Illustrative example of language acquisition through situation-based intention reading (left) and syntactico-semantic pattern finding (right), featuring a young child taking part in three communicative interactions. On the intention reading side of interactions A and B, the child makes a hypothesis about the meaning of the observed utterances based on environmental clues only. On the pattern finding side, the child creates a holophrastic construction after interaction A and uses this construction to create an item-based construction along with two holistic constructions after interaction B. One of these holistic constructions is then used to help the intention reading process during interaction C, after which another item-based construction is created.

constructions. The evolutionary dynamics of creating, strengthening, and weakening associations between linguistic forms and meanings, on any level of abstraction, based on their successful or unsuccessful use in communication, thereby provides a way to overcome the uncertainty that is inherent to the intention reading process.

As the language acquisition process of a child further unfolds, the linguistic clues that they provide as a speaker and discern as a listener gradually become more fine-grained and reliable. Their language understanding and production processes become less reliant on immediate environmental clues, making displaced communication possible and enabling language users to take part in conversations about abstract topics (Hockett and Hockett 1960). Such conversations rely to a large extent on the interpretation of linguistic clues with respect to the background knowledge and belief system of the individual interlocutors (Van Eecke et al. 2023). Yet, the linguistic and non-linguistic knowledge that is involved remains, on the most fundamental level, grounded in real-world experiences. "Democracy" and "eternity," for example, are often cited as conversational topics at the more abstract end of the spectrum (see, e.g., Löhr 2022). The corresponding concepts have no directly observable referents as they capture social, cultural, and linguistic constructs. However, it does not take much to argue that without the grounding of these concepts in an individual's lived experiences with volition, authority, permanence, perishability, and boredom, for instance, conversations about these topics would be hollow and soulless, if they would ever take place at all.

In sum, humans acquire their native languages by actively taking part in communicative interactions with other language users. These interactions are meaningful,

intentional, and situationally embedded in a communicative environment. The situated and communicative nature of the interactions is essential to the language acquisition process, as language learners crucially depend on clues provided by the communicative environment to make sense of the utterances they perceive. Through the interdependent processes of situation-based intention reading and syntactico-semantic pattern finding, language learners gradually build up an inventory of productive schemata that capture, on any level of abstraction, the relationship between linguistic forms on the one hand, and their meanings and communicative functions on the other. Importantly, the language acquisition process is communicatively motivated, personal, usage-based, and constructivist, in the sense that the linguistic system of each person is individually built up and shaped by their past experiences and interactions. Through the evolutionary dynamics of creating, strengthening, and weakening associations between linguistic forms and meanings based on their successful or unsuccessful use in communication, language learners effectively manage to build up a productive linguistic system that is tied to their own physical and cognitive endowment yet compatible on a communicative level with the linguistic systems of other members of their community.

2. Large Language Models Learn Language from Texts

Over the last decade, important advances in neural machine learning and infrastructure (Sutskever, Vinyals, and Le 2014; Vaswani et al. 2017), combined with the availability of huge text corpora, have led to previously unimaginable progress in the field of natural language processing. The main catalyst in this process has been the advent of transformer-based large language models (LLMs), such as BERT (Devlin et al. 2019), GPT-3 (Brown et al. 2020), PaLM (Chowdhery et al. 2022), BLOOM (BigScience Workshop 2022), and LLaMA (Touvron et al. 2023). Definitive solutions to a wide variety of NLP subtasks that were previously considered immensely challenging, such as machine translation, speech recognition, conversational question answering, and text summarization, now seem well within reach—if not already considered achieved (for an overview, see, e.g., Lauriola, Lavelli, and Aiolli 2022). The extraordinary performance of these models on tasks that are traditionally considered hallmarks of human intelligence has naturally sparked the interest of researchers in other disciplines, including cognitive science, psychology, and linguistics. The focus of these interdisciplinary investigations is typically on comparing the performance of humans and LLMs on a variety of tasks, with the goal of assessing to what extent LLMs can serve as a model of language understanding and intelligent reasoning in humans. Methodologically, these studies more often than not adopt an approach called **probing** (Hupkes, Veldhoen, and Zuidema 2018), in which textual input prompts are designed to elicit the generation of output text that is taken to unveil the knowledge implicitly captured by the probed LLMs (Vulić et al. 2020). Probing as such does not reveal any explicit reasoning processes or strategies, but leaves it up to the interpretation process of the human experimenter to draw any conclusions about what a model must have "known" or "thought" (Shiffrin and Mitchell 2023). While probing is an effective methodology to systematically investigate the types of problems that LLMs can handle or struggle with—thereby sometimes revealing interesting "curious failures" (Shiffrin and Mitchell 2023)—there is some danger that the output-based interpretation of the inner workings of LLMs becomes a speculative and unfalsifiable endeavor. This risk can only be mitigated by a constant awareness of how these models are constructed, what they do, and how they do it (Shanahan 2024).

A major result obtained through probing studies is that despite the exceptional performance of LLMs on many NLP tasks, they seem to struggle with human-like logical and pragmatic inferencing (Jiang and de Marneffe 2021; Hong et al. 2023). In particular, it is often observed that LLMs fail to reason in a human-like manner about the knowledge they seem to capture (Choudhury, Rogers, and Augenstein 2022; Weissweiler et al. 2022; West et al. 2024). Although the precise nature of this knowledge is still a very active field of investigation, sometimes referred to as "BERTology" (Rogers, Kovaleva, and Rumshisky 2020), it has been convincingly argued that the reasons why large language models struggle so much with logical and pragmatic inferencing are to a large extent ascribable to the fundamental differences between how large language models are constructed and how human languages are acquired (Bender and Koller 2020; Trott et al. 2023).

Large language models are rooted in the tradition of distributional linguistics, as pioneered by Joos (1950), Harris (1954), and Firth (1957). Distributional linguistics offers a framework for analyzing language in terms of "the occurrence of parts [...] relative to other parts" (Harris 1954). These parts can be of any nature and can be chosen to correspond, for example, to phonemes, words, or dependency relations. Let us consider the example of words, given the focus of today's NLP techniques on the word (or subword/character group) level. A word-level distributional analysis of a language describes that language solely in terms of the occurrence of words relative to each other. Words are thereby represented in terms of their frequency of collocation with all other words of the language as observed in corpora of language use. A word's frequency of collocation is captured in the form of a high-dimensional vector of which each dimension corresponds to one other word of the language. These other words are referred to as the **context words** of a given **target word**. Frequency of collocation is defined in terms of the number of times a context word appears in the **context window** of a target word, by which the context window spans a chosen number of words to the left and to the right of the target word. The vectors representing words are thus the result of counting word co-occurrences in corpora of language use. Interestingly, words that exhibit similar distributions, indicated by a small angle between their vector representations, tend to have similar meanings. For example, the words *salmon* and *cod* would both frequently appear in the neighborhood of *swim*, *water*, and *dinner*, and only rarely co-occur with *democracy* and *eternity*. On the other hand, *cod* would be collocated more often than *salmon* with *chips* and less often with *river*. Their overall similar collocational behavior thereby indicates a high degree of semantic similarity, whereas their collocational differences are taken to indicate a semantic difference. The assumption that collocational behavior can serve as a proxy to meaning has become known as the **distributional hypothesis**, or perhaps more famously through the quote "You shall know a word by the company it keeps!" (Firth 1957). The elegance of distributional linguistics lies in the fact that linguistic forms can be analyzed independently from any external factors (Harris 1954). Word-level distributional analyses can indeed be constructed based on raw text only, without the need to include etymology, history, semantics, pragmatics, or grounding in (knowledge of) the world, for example. At the same time, aspects of those other factors can only be captured as a side-effect of their influence on the distribution of words with respect to each other (Hill, Reichart, and Korhonen 2015).

While the foundations of the field of distributional linguistics were laid out in the 1950s, and its routine application to specific NLP tasks such as word sense disambiguation (Schütze 1992, 1998) and essay grading (Rehder et al. 1998) date back to the 1990s, the wide-spread adoption of distributional semantics as the standard meaning

representation in the field of NLP is much more recent. Turney and Pantel (2010) are often credited with drawing the attention of the NLP community at large on the potential of using distributional meaning representations. A major turning point in this direction was the introduction of the famous CBOW and Skip-gram architectures for efficiently estimating high-quality word vectors, along with their implementation in the *word2vec* software package and the public availability of pre-trained word vectors (Mikolov et al. 2013a, b). In contrast to the high-dimensional, "sparse" vectors traditionally obtained through counting word co-occurrences, low-dimensional, "dense" vectors, called **embeddings**, were now estimated by a neural network in a word prediction task. While the method is somewhat different, the underlying idea and intuition remain the same: Words are represented as vectors that reflect their observed collocational behavior in corpora of language use. Interestingly, the dense vectors that are obtained through machine learning methods turn out to outperform traditional sparse vectors on virtually every NLP task they are used for (Baroni, Dinu, and Kruszewski 2014), for reasons that are still not fully understood today (Jurafsky and Martin 2024). The next major milestone concerns the use of transformer networks (Vaswani et al. 2017) to learn word representations via a masked language modeling task (Devlin et al. 2019). Such a task consists in filling in randomly selected words that are blanked in a given stretch of text. The major difference with the Skip-gram and CBOW-based methods is that the resulting word embeddings are "contextual" rather than "static." As such, a target word is no longer represented by a single vector, but by a different vector for each different context in which it appears. The fact that words are now represented on the level of their textual occurrence rather than their dictionary entry makes for a more fine-grained modeling of their collocational behavior, and consequently leads to a drastically improved performance on downstream NLP tasks. For a more extensive overview of vector-based word representation methods, we refer the reader to the recent survey by Apidianaki (2022).

The scale on which today's large language models capture the collocational behavior of words has led to a situation in which the texts that are generated do not only exhibit human-like lexical and morpho-syntactic correctness, fluency, and eloquence, but in which the resulting texts are also characterized by a remarkable discourse coherence and semantic adequacy. The contextual embeddings learned by these models mirror the cognitive semantic spaces of humans to a remarkable extent (Vulić et al. 2020) and even seem to capture non-linguistic knowledge to an extent that they can be used to solve natural language understanding tasks that were always thought to require vast knowledge of the world or human-like intelligent reasoning capabilities (cf. Hu et al. 2023; Webb, Holyoak, and Lu 2023). The reasoning errors made by LLMs under certain circumstances even resemble those observed in humans (Dasgupta et al. 2022).[1] Yet, these models still fully subscribe to the distributional linguistics paradigm,

1 In particular, Dasgupta et al. (2022) show that the tendency of humans to commit logical fallacies when faced with reasoning problems that are not consistent with real-world knowledge and beliefs are mirrored by LLMs. For example, syllogistic fallacies tend to go unnoticed when the conclusion is intuitive. Likewise, logic puzzles are more likely to be solved correctly when instantiated as common real-world situations. Taking into account the word prediction task through which LLMs are trained, the fact that they are more likely to consider "probable utterances" to be valid conclusions of syllogisms and that they achieve better results on logic problems that are expressed in more realistic language is not entirely surprising (cf. McCoy et al. 2023). While interesting to see that human reasoning exhibits similar effects, more research would be needed in humans and machines to investigate whether this similar behavior cannot be ascribed to familiarity with utterances in machines vs. familiarity with situations in humans. Adversarial robustness experiments in which utterances are varied but describe the same situation could provide an interesting methodological framework here.

where languages are modeled solely in terms of the observed "co-occurrence of parts [...] relative to other parts [...] without intrusion of other features such as history or meaning" (Harris 1954). As such, all "knowledge" and "reasoning capabilities" they might hold are rooted in contextual word prediction based on collocations that were observed in huge amounts of raw textual data. While it is truly inconceivable that such results are obtained based on word distributions only, admittedly distilled from equally inconceivable amounts of texts, the fact that these LLMs are learned in complete absence of situationally grounded and intentional communicative interactions lies at the origin of a number of inherent limitations:

- **Hallucinations**. Large language models seamlessly mix fact and fiction in the output they produce, a phenomenon referred to by the term **hallucination**. Hallucination is a direct consequence of the generative nature of the models, by which they use collocation-based patterns and structures observed in input texts to generate output texts that exhibit similar patterns and structures. The epistemological status of the generated texts is thereby uniform and may consequently correspond to facts, beliefs, opinions, or fantasies in human terms (McKenna et al. 2023). Simply put, very probable word combinations might very well not correspond to factual truths or non-factual information contained in the training data. The use of the term "hallucination" to refer to generated fantasies is somewhat misleading, as hallucinations result from exactly the same generative process as any other output text. The use of the label is thus necessarily the result of a post-hoc interpretation by the human interpreter. One could as well argue that all output of LLMs is hallucinated, but that their hallucinations are remarkably often semantically adequate.
- **Human-like logic and pragmatic reasoning**. Large language models suffer from an apparent dissonance between the vast knowledge they seem to capture and their difficulty to use this knowledge to perform human-like logical and pragmatic reasoning (Choudhury, Rogers, and Augenstein 2022; Weissweiler et al. 2022; Mitchell and Krakauer 2023; Hong et al. 2023; West et al. 2024; Mitchell, Palmarini, and Moskvichev 2024). This limitation is ascribable to two main reasons. The first reason concerns the lack of situational grounding during the training process.[2] LLMs are trained to generate output texts that exhibit similar (sub)word distributions as those observed in input texts. This task, at which LLMs clearly excel, perfectly aligns with the standard assumption of generative machine learning that models should generate output that follows the same distributions as the input they are trained on. However, when the task shifts from generating texts that reflect the distribution of words in texts written by humans to human-like logic and pragmatic reasoning, the standard machine learning assumption is no longer upheld. Indeed, when the distribution of words in texts is used to make predictions about

2 We are using the term **situational grounding** to refer to "the communicative, perceptual, or goal-oriented contexts in which language occurs" (Pavlick 2023), thereby spanning at least Coelho Mollo and Millière's (2023) notions of referential grounding, sensory-motor grounding, and communicative grounding.

the distribution of objects and events in the world, it is no longer guaranteed that the input and target distributions coincide. The distribution of objects and events in the world might well be unlearnable from the distribution of words in texts. Moreover, it is extremely hard to gauge where these distributions coincide sufficiently to justify accurate predictions and where they are too different (Shichman et al. 2023). The second reason why LLMs struggle so much with human-like logic and pragmatic inferencing concerns the absence of communicative intent during the training process (cf. Bender and Koller 2020). Human linguistic communication is intentional, in the sense that speakers seek to achieve an effect in their interlocutors (Austin 1962). At the same time, language is an inferential coding system, by which is meant that a speaker's intentions are not losslessly encoded in the utterances they produce, but need to be inferentially reconstructed by their interlocutors (Sperber and Wilson 1986). Pragmatic inferences are drawn based on these communicative intents, whereby the redundancy of information is particularly meaningful (Grice 1967; Kravtchenko and Demberg 2022). For example, if your co-worker mentions that the door was locked when they entered the lab in the morning, this presupposes (at least) that the door is usually not locked when they arrive. Or if you explicitly mention that your friend paid for their coffee, this presupposes that they normally don't pay for their coffee. These presuppositions do not result from the information encoded in the utterances, or from general knowledge of the world, but from a reconstruction of the intention that motivated the speaker to convey this information that would otherwise have been redundant or irrelevant. While pragmatic inferences play a central role in human communication, they are notoriously difficult to draw for LLMs (Chang and Bergen 2024; Ruis et al. 2023). This should not be surprising given the fact that LLMs are trained on raw textual utterances in the absence of clues that might reveal the communicative intents that motivated their production. At the end of the day, the distributional hypothesis states that the distribution of words reflects their semantic similarities and differences, but remains silent about why particular utterances are produced in the first place.[3]

- **Data-hungriness**. Large language models are learned from raw textual data through a word prediction task. Consequently, all aspects of meaning, reasoning, and world knowledge that we expect LLMs to capture need to be learned indirectly through the intermediary of their effect on the distribution of words. In practice, this requires access to text corpora that consist of hundreds of billions of tokens (Chang and Bergen 2024). For example, the GPT-3 family of models was trained on 300 billion tokens sampled from a base corpus of more than 500 billion tokens (Brown et al. 2020) and the Chinchilla model was even trained on 1.4

3 These arguments relate to a broader philosophical discussion within the community about whether LLMs are capable of symbol grounding, reference, and capturing meaning. In this debate, it has been argued that language models cannot capture meaning as they lack reference and communicative intent (Bender and Koller 2020), but also that reference (Piantadosi and Hill 2022) and communicative intent (Pavlick 2023) are not prerequisites for meaning, and that beliefs and experiences are not prerequisites for reference (Mandelkern and Linzen 2024).

trillion tokens (Hoffmann et al. 2022). These amounts of data are necessary to stretch the distributional hypothesis from making predictions about semantic similarity on the word level to generating output that exhibits aspects of human-like intelligence. The number of tokens that an LLM is exposed to is several orders of magnitude larger than the number of tokens a human is exposed to, estimated at about 60 million at the age of 5 and 200 million at the age of 20 (Frank 2023). It should be noted, however, that such a comparison is virtually meaningless, as the nature of their input, namely, texts vs. situated communicative interactions, is fundamentally different.

- **Sensitivity to biases**. Large language models mirror the biases that are present in the data they were trained on (Nissim, van Noord, and van der Goot 2020; Vallor 2024), concerning, for example, age, ethnicity, gender, religion, and sexual orientation (Navigli, Conia, and Ross 2023). These biases do not even originate from the unjustifiable extrapolation of correlations observed in the world, but from texts, written by humans, that at times literally contain stereotypes, hate speech, or fringe opinions. While a laudable idea in principle, the curation of training data is not a scalable solution given the difficulty of this task and the sheer amount of training data that is needed.

In sum, LLMs learn language from texts. They meticulously capture the collocational behavior of (sub)words in a fine-grained manner and on an inconceivable scale. All knowledge they seem to capture and reasoning abilities they seem to possess originate from observing the distribution of (sub)words with respect to each other in corpora of human-written texts. While their capabilities are astonishing, improving on almost any NLP/NLU task they are applied to and stretching the distributional hypothesis further than anyone would have imagined, the fact that they are learned from textual data and pass over the situated, communicative, and intentional aspects of human language use lies at the root of a number of inherent limitations. These concern in particular the uniform epistemological status of their output, their limited ability to perform logical and pragmatic reasoning, their data-hungriness, and their susceptibility to biases.

3. Towards Human-like Language Learning in Machines

The text-internal prediction task through which LLMs learn their linguistic capabilities sharply contrasts with the situated communicative interactions through which humans acquire their native languages. As we have argued above, the fundamental differences between the learning processes of humans and LLMs in turn lead to fundamental differences in the kind of linguistic and non-linguistic knowledge that is built up, as well as in how this knowledge is used to engage in communicative and non-communicative actions. In no way does this argument refute the role of (distributional) statistical learning in human language acquisition (cf. Saffran, Aslin, and Newport 1996; Aslin 2017), which we consider, like Tomasello (2003) and Kuhl, Tsao, and Liu (2003) among others, to be an integral part of a human's broader pattern finding ability (cf. Section 1). It does, however, imply that text-based statistical learning is insufficient on its own, and that the key towards more human-like language learning in machines lies in more

faithfully modeling the circumstances under which human languages are acquired and used to communicate.

The goal of the present section is to zoom in on approaches that incorporate aspects of the meaningful, intentional, situated, communicative, and interactional nature of linguistic communication with the aim of operationalizing more human-like language learning in machines. We start by briefly considering recent work that remains firmly rooted within the core of the LLM paradigm, but extends it beyond a purely text-based prediction task (Section 3.1). In particular, we touch upon the integration of input from other modalities and the use of reinforcement learning to better align the output of LLMs with externally defined criteria such as human preferences. We then move on to discuss a line of work that takes a very different perspective on the matter, as it explicitly aims to study how artificial agents can acquire languages through the processes of situation-based intention reading and syntactico-semantic pattern finding during situated communicative interactions (Section 3.2). We discuss two concrete experiments that operationalize this paradigm and show how the linguistic systems that are acquired by the agents in such a setting are of a fundamentally different nature than those acquired by LLMs.

3.1 Large Language Models beyond (Sub)word Distributions

The argument that the key towards more human-like language learning in machines lies in more faithfully modeling the circumstances under which human languages are acquired and used to communicate minimally calls for letting go of a strict interpretation of the distributional hypothesis. In other terms, the idea that the "co-occurrence of parts [...] relative to other parts [...] without intrusion of other features such as history or meaning" (Harris 1954) suffices to capture all there is to human linguistic communication can no longer be upheld. When it comes to LLMs, this implies that the paradigm needs to be extended to integrate information beyond the collocational behavior of (sub)words as learned from raw text corpora through a text-internal prediction task. Two main approaches to extending LLMs beyond observed (sub)word distributions have already proven to be valuable additions to the paradigm. While the first approach focuses on broadening the input to the learning process to other modalities than text, the second approach concentrates on better aligning the output of LLMs with externally defined criteria such as human preferences or task-specific ground-truth accuracies.

Multi-modal Language Models. The high-level idea behind multi-modal language models is that information originating from multiple modalities can be integrated into an LLM's embedding space (see, e.g., Radford et al. 2021; Baevski et al. 2022; Lyu et al. 2023).[4] The theoretical promise of this approach is that models capable of cross-modal inferencing break free from the limitation that knowledge of the world, or any knowledge other than word distributions per se, can only be captured as a side-effect of the influence of external factors on the observed distribution of words (see Section 2). For example, the graphical structures that are commonly used to represent spatial and temporal relations through the visual modality, for example through maps or timelines, concisely hold information about distances, durations, and temporal orderings that are hard, inefficient,

4 Note that approaches that first map input from other modalities to texts, for example through image captioning, and then feed the resulting texts to an LLM, are also sometimes referred to as multi-modal LLMs. However, we do not consider these approaches in this section, as the LLMs themselves remain purely text-based.

or practically impossible to capture using the textual modality (see, e.g., Li et al. 2020). After all, there are good reasons why humans make extensive use of graphical representations in their atlases, road maps, and handbooks on subjects such as anatomy, botany, and chemistry. Apart from the theoretical advantages that come with enriching the representational power of LLMs beyond observed (sub)word distributions, the integration of non-textual modalities has also facilitated the development of a variety of real-world applications, such as multi-modal chatbots (OpenAI 2023; Yamazaki et al. 2023), robotic manipulation planning systems (Driess et al. 2023), and navigation aids (Fan et al. 2023).

Language Model Alignment. It has often been argued that the text-internal training objective of LLMs yields models that might well generate texts that are in some sense indistinguishable from human-written texts, but at the same time fail to satisfy user expectations in terms of helpfulness, relevance, factual correctness, safety, and adherence to human values, among others (Weidinger et al. 2021; Liu et al. 2022; Tan et al. 2023). A particularly successful approach to better satisfying user expectations concerns the incorporation of text-external objectives through a two-stage language model alignment process. Concretely, a standard LLM is first pre-trained on a text-internal prediction task. Then, the learned model is fine-tuned to optimize a text-external criterion such as human preference or ground-truth accuracy on a specific task. The language model alignment approach has perhaps become best known through its use in combination with **reinforcement from human feedback** to align the output of open-domain chatbots to user expectations. Typically, a reward model is first trained in interaction with humans to capture how humans would rank a set of possible output texts. Then, this model is used as a reward function to fine-tune the pre-trained LLM. For a more detailed introduction into language model alignment, we refer the reader to Ziegler et al. (2019), Ouyang et al. (2022), and Bai et al. (2022).

The integration of input data from modalities other than text as well as the alignment of LLMs to text-external criteria cautiously lift the limitation that all knowledge, reasoning capabilities, and linguistic behavior of LLMs needs to be learned through the intermediary of observed (sub)word distributions. While the integration of input from multiple modalities provides a way to incorporate richer representations of the situational embedding of observed utterances, the alignment of LLMs to text-external criteria facilitates the integration of any optimization criterion for which an effective reward function can be designed or learned. Naturally, this sparks the question of whether the limitations ascribed in Section 2 to the absence of situationally grounded and intentional communicative interactions during the training process of LLMs still persist. Any answer to this question crucially hinges on two variables that were left open in the argument.

The first variable concerns the nature of the multi-modal input to the learning process, in particular the extent to which the input can reflect the sensory-motor interactions in which humans engage every day. There seems to be no reason to believe that it would be fundamentally impossible to collect training data in real-world situations using sensors modeled after the sensory apparatus of humans.[5] At the same time, no

5 For an exploratory operationalization of this idea, see, for example, Vong et al. (2024), who trained a language model based on a toddler's auditory and visual experiences captured using head-mounted sensors for 61 hours over a period of 19 months (Sullivan et al. 2021). See Gabrieli et al. (2022) and Park et al. (2022), respectively, for examples of devices specifically designed to emulate human taste and tactile sensing.

comprehensive data sets of this kind exist to date and their construction would require overcoming extensive technical, ethical, and financial challenges. For now at least, the input to multi-modal LLMs typically remains restricted to a combination of texts, speech recordings, images, and/or videos.

The second variable concerns the reward function that is used in the language model alignment process. The central question here is to what extent such a function can capture human linguistic and non-linguistic behavior. The situation is much thornier than in the case of the first variable. The design of robust and reliable reward functions is notoriously difficult even in simple environments (Skalse et al. 2022; Ngo, Chan, and Mindermann 2024). The underlying reason is closely related to Goodhart's law (Goodhart 1975; Manheim and Garrabrant 2018), which states that measures that become targets cease to be useful as measures (Ngo, Chan, and Mindermann 2024). For instance, the number of times the entrance door to a shop opens and closes might well be an excellent proxy to assess how busy that shop is. However, rewarding employees for maximizing this number is unlikely to be effective at making the shop any busier, as employees might spend all their time flipping switches and keeping clients out so that the door can be opened and closed more efficiently. If the goal is to capture something as multifaceted as human linguistic communication, it seems highly unlikely that an entirely non-gameable proxy can even exist. At the same time, it is difficult to give up on proxies altogether, as the alternative of replacing every invocation of the reward function by an authentic real-world interaction is definitely not feasible in such a data-hungry setting. Reward functions that rely on sufficiently realistic simulations of the situated communicative interactions that humans engage in might offer a way out, but they will somewhere need to strike a balance between robustness and versatility on the one hand, and computational efficiency on the other.

Do the inherent limitations of LLMs then persist when extending the paradigm beyond modeling (sub)word distributions? The argument developed in Section 2 that the tendency of LLMs to hallucinate, their limited ability to perform human-like logical and pragmatic reasoning, their data-hungriness, and their sensitivity to biases are limitations that are inherent to the paradigm was based on the premise that LLMs are learned from texts, optimizing a text-internal prediction objective. If the input shifts from raw texts to human-like sensory-motor observations and if the text-internal prediction objective is complemented with a human-like linguistic communication objective, the original argument no longer holds. The reality is, however, less clear-cut. While current LLM extensions beyond (sub)word distributions already exhibit more desirable behavior than text-internal LLMs, they are still far removed from the ideal, both on the input side and on the task side. The most fundamental question here seems whether a non-gameable proxy that is both practically usable for language model alignment and sufficiently close to capturing human linguistic behavior could in principle be designed or learned. In the end, the extent to which the limitations of LLMs can be lifted when extended beyond (sub)word distributions is bounded by the extent to which the meaningful, intentional, situated, communicative, and interactional aspects of human linguistic communication can be integrated into the training process.

3.2 Language Learning from Situated Communicative Interactions

The goal of the present section is to discuss an alternative approach that leaves the text-based prediction paradigm altogether and takes situated communicative interactions as the starting point for modeling language acquisition in machines. In particular, we

examine a line of work that studies how artificial agents can acquire languages through the processes of situation-based intention reading and syntactico-semantic pattern finding during simulated communicative interactions. It is not our aim to provide a comprehensive overview of computational models of human language acquisition or emergent communication, for which we refer the reader to Doumen et al. (2025), Steels (2011), and Lazaridou and Baroni (2020), nor to provide an application-level alternative to current LLMs. Instead, we discuss two concrete experiments that show how the linguistic systems that are acquired by the agents in such a setting are of a fundamentally different nature than those acquired by LLMs. Through these experiments, we aim to further convince the reader that more faithfully modeling the circumstances under which human languages are acquired is likely to constitute a major leap towards more human-like language processing in machines.

The two experiments that we discuss below highlight different aspects of the approach. The first experiment, which is presented in Section 3.2.1, demonstrates how a population of autonomous agents can learn to communicate about entities that they observe in their environment. Concretely, by taking part in meaningful and intentional situated communicative interactions, the population converges on a communicatively adequate linguistic convention. Through the processes of intention reading and pattern finding, each individual agent builds up an inventory of holistic constructions that associate atomic forms to concept representations that are grounded in their own sensory endowment and experiences. The second experiment, presented in Section 3.2.2, focuses on the acquisition of constructions that capture compositional patterns of higher morpho-syntactic and semantic complexity. We consider a tutor-learner scenario in which the learner agent has previously acquired an inventory of conceptual distinctions that are grounded in the environment in which the agents operate, for example through the mechanisms presented in the first experiment. Through the processes of intention reading and pattern finding, the learner agent now bootstraps an inventory of item-based and holistic constructions that they can use to ask and answer English questions about their environment. The meaning side of these constructions is again grounded in the agent's personal endowment and experiences.

3.2.1 Experiment 1: Acquisition of Grounded Concepts. The first experiment illustrates how a population of artificial agents that are equipped with sensors can establish a linguistic convention that is adequate to refer to entities that they perceive in their surroundings. The emergent convention consists of holistic constructions that associate atomic forms to concept representations that are grounded in the agents' environment. Not only does this experiment show how the acquisition of situationally grounded concept representations through task-oriented communicative interactions can concretely be operationalized in artificial agents but it also shows how these concept representations can emerge and evolve to satisfy the communicative needs of a community of language users. The foundations of the methodological framework that is adopted date back to pioneering work by Steels (1995), Batali (1998), and Oliphant (1999), among others. The concrete experiment that we discuss was originally introduced by Botoko Ekila et al. (2024a, b). A skeleton version of the methodology is presented below to better fit the illustrative role of the experiment in this article.

In order to set up the grounded concept learning experiment, we first define a population, a world, and an interaction script. The population consists of a number of autonomous agents, say 10, which are each endowed with a sensory apparatus. The world consists of a number of entities that are represented through feature vectors.

Each dimension of these vectors corresponds to a particular sensor that the agents are endowed with. Imagine, for example, that the world consists of geometrical objects and that the agents are endowed with 20 visual sensors. As such, the agents would perceive individual objects along 20 dimensions, say the number of corners of an object, its width-height ratio, its color channel values, its area, and its position on the horizontal and vertical axes, among others. At the beginning of the experiment, agents do not know any words or concepts. Any objects they might encounter in their environment are perceived as continuously valued feature vectors resulting from sensor read-outs.

The agents in the population engage in a series of situated communicative interactions that follow a structured interaction script. The script that is adopted in this experiment, and which will be formally defined below, proceeds as follows. First, a scene is created as a random subset of entities from the world. We will assume for now that a scene consists of a minimum of three and a maximum of ten entities. Then, two agents are drawn from the population and are randomly assigned the roles of speaker and listener for the purposes of the current interaction. One entity from the scene is randomly selected to be the topic of the interaction. As such, it will be the task of the speaker to draw the attention of the listener to this specific entity in the scene. Obviously, the identity of the topic is disclosed to the speaker only. The speaker then retrieves the construction in its construction inventory that best combines entrenchment and discriminative power for the topic with respect to other entities in the scene. If no adequate construction exists, as will often be the case in the initial stages of the experiment, the speaker invents a new construction that couples a concept representation, created based on the observed feature vector, to a new word form, say *demoxu*. The form of the selected construction is then uttered to the listener. If the listener already knows a construction of which the form side matches the utterance, they identify the entity in the scene that most closely matches the meaning side of the construction and point towards it. If the listener does not know a construction of which the form side matches the utterance, they indicate that they did not understand. In both cases, the speaker then provides feedback to the listener by signaling communicative success or failure and by pointing towards the topic entity. In case of success, both agents boost the entrenchment score of the constructions that they have used and inhibit the score of any competing constructions, that is, constructions that would also have discriminated the topic in the scene. At the same time, the concept representations of the used constructions are shifted towards the perceived feature vector. In case of failure, the speaker will inhibit the entrenchment score of their used construction. If the listener knew a construction of which the form side matches the observed utterance, they will inhibit its entrenchment score and shift its meaning side towards the observed feature vector. If they did not know such a construction, they will learn a new construction that couples the observed utterance with a concept representation that is created based on the perceived feature vector. A schematic overview of a situated communicative interaction of the grounded concept learning experiment is presented in Figure 2. The figure shows an interaction in which the speaker could successfully use the utterance *walibu* to draw the attention of the listener to the topic entity in the scene.

Formally, an experiment $E = (W, P, G)$ is defined as a tuple that groups a world $W = \{e_1, \ldots, e_m\}$ comprising m entities, a population $P = \{a_1, \ldots, a_k\}$ of k agents, and a sequence $G = (g_j)_{j=1}^{i}$ of i situated communicative interactions. Each agent is equipped with a set of l sensors $S = \{s_1, \ldots, s_l\}$ and is initialized with its own empty construction inventory $I_a = \{\}$. Each entity in the world is represented by means of a continuously valued l-dimensional feature vector X, of which the dimensions correspond to

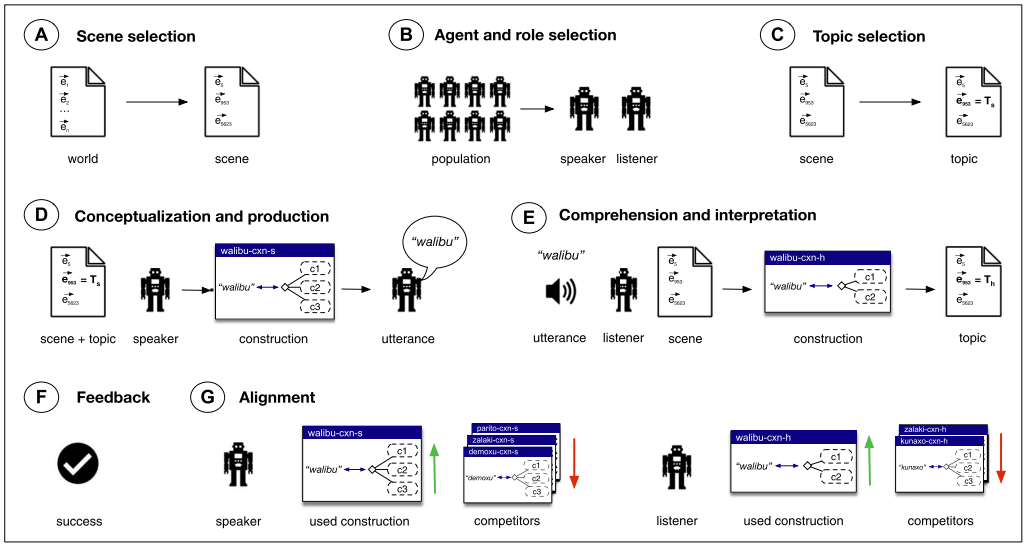

Figure 2
Schematic overview of a successful situated communicative interaction during the grounded concept learning experiment. A scene comprising three entities is drawn from the world (A). Two agents are drawn from the population and assigned the roles of speaker and listener (B). A topic entity is drawn from the scene and disclosed to the speaker only (C). The speaker selects the construction from its inventory with the highest communicative adequacy for the topic in the scene. In this case, the speaker identifies the WALIBU-CXN-S and utters its form *walibu* (D). The listener observes the produced utterance. In this case, the listener has previously acquired a construction with the form *walibu* and retrieves the entity from the scene that matches its meaning side most closely. The listener points to this entity (E). The speaker signals success (F). Both speaker and listener boost the entrenchment score of the constructions they used, shift their meaning side towards the observation, and inhibit the score of competing constructions (G).

properties of the entities that can be recorded by the agents' corresponding sensors.[6] A holistic construction $w = (f, c, s)$ is defined as a pairing between a form $f \in F$, a concept representation c, and an entrenchment score s. Entrenchment scores are bounded between a minimum of 0 and a maximum of 1. F is an infinite set of possible forms. Concept representations are each defined as a sequence of l tuples $c = ((\omega_1, \mu_1, \sigma_1), \ldots, (\omega_l, \mu_l, \sigma_l))$ that group three numerical values: ω, μ, and σ. Each tuple relates to a feature channel of one sensor with which the agents are endowed. The weight value ω_i represents the importance of the feature channel for the concept, the mean value μ_i holds the prototypical value for the concept on this channel, and σ_i holds the standard deviation for the concept on this channel. Concepts are thus represented as a sequence of normal distributions. Each distribution relates to a feature channel and the relevance of the channel for the concept is indicated by a weight value.

6 In order to fit the illustrative purposes of this article, we present a simplified version of the original methodology. We assume here that all agents are equipped with the same set of calibrated sensors. We kindly refer the reader to Botoko Ekila et al. (2024a, 2024b) for a more elaborate version of the methodology that successfully accommodates heteromorphic populations, perceptual differences, and even sudden sensor defects.

The interaction script according to which each communicative interaction $g \in G$ takes place is formalized as follows:

(A) Scene selection. n entities are drawn from the world to constitute a scene: $C = \{e_1, \ldots, e_n\} \in W$.

(B) Agent and role selection. A speaker $S \in P$ and a listener $L \in P$ are drawn from the population P.

(C) Topic selection. The topic entity $T \in C$ is drawn from the scene C and disclosed to the speaker S.

(D) Conceptualization and production. The speaker considers all constructions in its inventory that discriminate the topic in the scene, that is all $w = (f, c, s) \in I_S$ for which holds that the similarity between their concept representation c and the feature vector X for the topic entity T is higher than the similarity between c and the feature vector for any other entity in the scene $e \in C$. The similarity between a construction's concept representation c and a feature vector X is computed according to Equation (1), taking into account the similarity on each feature channel as well as the weight value of each feature channel in the construction's concept representation.

$$sim(c, X) = \underbrace{\sum_{i=1}^{l}}_{\text{sum over channels}} \underbrace{\frac{\omega_i}{\sum_{j=1}^{l} \omega_j}}_{\text{weight}} \underbrace{\exp(-|\frac{x_i - \mu_i}{\sigma_i}|)}_{\text{similarity}} \quad (1)$$

If multiple discriminating constructions exist in the construction inventory of the speaker, the construction with the highest communicative adequacy is selected. The communicative adequacy reflects both the degree of entrenchment of the construction and the extent to which it discriminates the topic entity in the scene. The idea is that constructions that are more entrenched and better discriminate the topic are more likely to lead to communicative success. The communicative adequacy is concretely computed by multiplying the entrenchment score s of the construction with the difference in similarity between the construction's concept representation c and the feature vector for the topic on the one hand, and the similarity between c and the feature vector for the closest other entity in the scene on the other. The form f of the construction with the highest communicative adequacy is then uttered by S as U and shared with L. If no discriminating constructions exist in the construction inventory of the speaker I_S, a new construction $w = (f, c, s)$ is added, with f being randomly selected from the infinite set of forms F, s being assigned a default initial value, and $c = ((\omega_1, \mu_1, \sigma_1), \ldots, (\omega_l, \mu_l, \sigma_l))$ being initialized with the exact values of the feature vector X for $\mu_1 \ldots \mu_l$, with a default initial value for $\sigma_1 \ldots \sigma_l$, and with a default initial value for $\omega_1 \ldots \omega_l$. The form f of this construction is then uttered as U and shared with L.

(E) Comprehension and interpretation. If a construction with the form U exists in the construction inventory of the listener, that is, $w = (U, c, s) \in I_L$, L points to the entity in the scene $e \in C$ of which the feature vector X is most similar to c, according to the similarity metric defined in Equation (1). If no such construction exists in I_L, L signals that they could not understand U.

(F) Feedback. S signals success if L correctly identified T. Otherwise, S signals failure, and provides feedback to L by pointing to T.

(G) Alignment. If the communicative interaction g was successful, both the speaker S and the listener L will increase the entrenchment scores s_S and s_L of the constructions they have used, namely, $w = (U, c_S, s_S) \in I_S$ and $w = (U, c_L, s_L) \in I_L$, by a fixed value. S and L also shift the concept representations c_S and c_L of these constructions towards the feature vector X_T for topic entity T given the context C. They do that by updating μ_i and σ_i on each feature channel using Welford's online algorithm (Welford 1962) to incorporate X_T. As for the channel weights $\omega_i \ldots \omega_l$, the subset of channels with the highest discriminative power for T with relation to C is selected from all channels with positive discriminative power. The weights on the channels in this set are increased by a fixed step on a sigmoid function. The weights on all other channels are decreased by a fixed step on the same function. Both S and L also decrease the entrenchment score of any competing constructions, defined as other constructions in their inventories that have a positive discriminative power for T in C. The decrease in score for these constructions is proportional to how similar their concept representations are to the concept representation of the used construction. In other terms, constructions with more similar concept representations are punished harder as they are considered stronger competitors. The similarity of two concept representations is computed according to Equation (2) and takes into account the similarity of the distributions on each channel, the similarity of the weights on each channel, and the average of the weights on each channel. This last component is included to reflect that similarities and differences on channels with high weights are more meaningful to the overall concept similarity than those on channels with low weights.

$$sim(c_q, c_r) = \underbrace{\sum_{i=1}^{l}}_{\text{sum over channels}} \Bigg[\underbrace{(1 - H(\mathcal{N}(\mu_{qi}, \sigma_{qi}^2), \mathcal{N}(\mu_{ri}, \sigma_{ri}^2)))}_{\text{Hellinger similarity of distributions}} \underbrace{(1 - |\frac{\omega_{qi}}{\sum_{k=1}^{l} \omega_{qk}} - \frac{\omega_{ri}}{\sum_{k=1}^{l} \omega_{rk}}|)}_{\text{similarity of weights}} \underbrace{\frac{\frac{\omega_{qi}}{\sum_{k=1}^{l} \omega_{qk}} + \frac{\omega_{ri}}{\sum_{k=1}^{l} \omega_{rk}}}{2}}_{\text{average of weights}}\Bigg] \tag{2}$$

If the communicative interaction g was not successful, S will decrease the entrenchment score s_S by a fixed value. If a construction with the form U existed in the construction inventory of L, that is, $w = (U, c_L, s_L) \in I_L$, L will decrease its entrenchment score s_L with a fixed value and shift its concept representation c_L towards X_T given C. If no such construction existed, a new construction $w = (U, c, s)$ is added to I_L, with s being assigned a default initial value, and $c = ((\omega_1, \mu_1, \sigma_1), \ldots, (\omega_l, \mu_l, \sigma_l))$ being initialized with the exact values of the topic's feature vector X for $\mu_1 \ldots \mu_l$, with a default initial value for $\sigma_1 \ldots \sigma_l$, and with a default initial value for $\omega_1 \ldots \omega_l$.

The grounded concept learning methodology is directly applicable to any scenario where a population of agents needs to learn to communicate about entities that can be represented by means of continuously valued feature vectors. These vectors can originate from robotic simulations as referred to above or be derived from any tabular dataset that stores its entries in terms of continuously valued attributes. Botoko Ekila et al. (2024a) present three scenarios that show the broad applicability of the methodology on the one hand, and that illustrate the grounding of the agents' emergent linguistic knowledge in the communicative task and situated environment on the other. In the first scenario, referred to as CLEVR, the scenes in which the interactions take place are based on the images from the CLEVR dataset (Johnson et al. 2017). Each synthesized 3D image depicts three to ten geometrical objects that vary in shape, size, color, shininess, and 3D position. The conversion from images to scenes is performed using the visual preprocessing procedure described by Nevens, Van Eecke, and Beuls (2020), resulting, per scene, in a set of 20-dimensional feature vectors that each represent one object in the original image. The 20 dimensions of the vectors correspond to human-interpretable visual properties, such as the object's area, color channel values, position on the x and y axes, number of edges, and width-height ratio. A total of 1,000,000 training scenes and 100,000 test scenes were sampled from the original train and test splits, respectively. The scenes of the second scenario, referred to as WINE, were created based on the 4,898 physicochemical analyses of wine samples included in the Wine Quality dataset (Cortez et al. 2009). Each wine sample is described through an 11-dimensional feature vector in which each dimension corresponds to a particular physicochemical property of the sample, including for instance its acidity, amount of residual sugar, alcohol content, and amount of sulphates. Ninety percent of the wine samples were used to create 1,000,000 training scenes that each hold three to ten samples and the remaining 10% were used to create 100,000 test scenes in the same way. The scenes of the third scenario, referred to as CREDIT, were created based on the 284,807 financial transaction records included in the Credit Card Fraud Detection dataset (Dal Pozzolo et al. 2014). Each financial transaction is described along 28 dimensions that result from a principal component analysis. Again, 90% of the financial transaction records were used to create 1,000,000 training scenes that each hold three to ten records and the remaining 10% were used to create 100,000 test scenes. During the scene creation process, it was ensured that no scene holds the exact same entity more than once and that no entity that appears in a training scene is part of a test scene.

For all three scenarios, the parameters in the formal definition were set as follows. The population consists of 10 agents ($k = 10$) and each scene consists of a minimum of three and a maximum of 10 entities ($3 \leqslant n \leqslant 10$). Constructions initially receive an entrenchment score of 0.5. After a successful interaction, the scores of the used constructions are increased by 0.1 and the scores of their competitors are decreased by $0.02 * sim(c_q, c_r)$. After a failed interaction, the scores of the used constructions are decreased by 0.1. When it comes to the concept representations, initial channel weights (ω) are set to 0.5 and initial standard deviations (σ) to 0.01. Channel weights are rewarded or punished with a step of $+1$ and -5, respectively, on the sigmoid function $\frac{1}{1+e^{-1/2x}}$. Each experimental run consists of 1,000,000 communicative interactions, that is, one interaction for each scene in the training sets.

As the experimental runs unfold, the typical learning dynamics known from the language evolution literature (see, e.g., Blythe and Croft 2012) manifest themselves in all three scenarios. Figure 3 illustrates these dynamics for the case of the CLEVR experiment. The x axis represents the time dimension in terms of number of communicative interactions that have taken place. The solid line indicates, on the left y axis, the

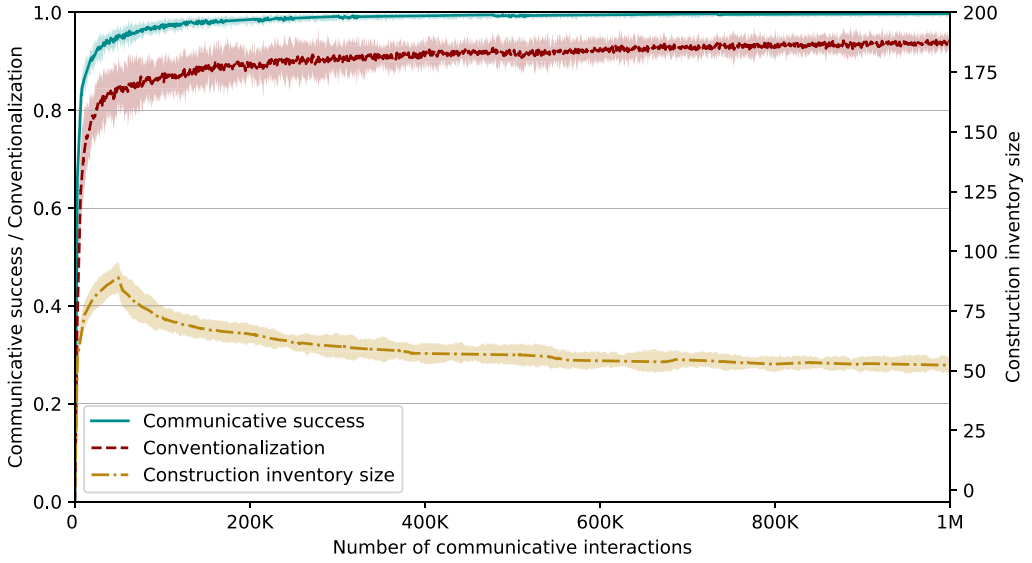

Figure 3
Evolutionary dynamics during the training phase of the CLEVR experiment, showing the degree of success on the communicative task, the level of conventionality of the emergent language, and the average number of constructions in active use as a function of the number of communicative interactions that have taken place. Mean and 2 standard deviations computed over 10 independent experimental runs with populations of 10 agents. This graph has been created based on experimental data from Botoko Ekila et al. (2024a).

average degree of communicative success over the last 1,000 interactions. An interaction is counted as successful if the listener has indeed been able to identify the topic entity during the comprehension and interpretation step of the interaction script. Given that the construction inventories of the agents are empty at the beginning of the experiment, the degree of communicative success necessarily starts at 0. After 50,000 interactions, a degree of communicative success of about 90% is reached. This number then continues to increase, with the agents successfully communicating in 99.5% of the interactions after the 1,000,000 interactions of the experiment have taken place. The dashed line indicates, also on the left y axis, the average level of conventionality of the emergent language over the last 1,000 interactions. An interaction is counted as linguistically conventional if the listener would in principle have produced the same utterance if they would have been the speaker. Like the degree of communicative success, the level of conventionality will always start at 0. It increases at a somewhat slower pace than the degree of communicative success and reaches about 90% after 1,000,000 interactions. The dashdotted line shows, on the right y axis, the average number of distinct constructions used by the agents during the last 1,000 interactions in which they participated as the speaker. In the earlier stages of the experiment, a wide variety of concept representations and words emerges in the population, peaking at about an average of 90 constructions per agent after 50,000 interactions. As the emergent language becomes more conventional, the number of constructions in active use starts to decline, with an average of just over 50 constructions per agent after 1,000,000 interactions.

After running the experiments for 1,000,000 interactions, the emerged conventions are evaluated against the test portions of the datasets. The interaction script remains the

Table 1
Results of the CLEVR, WINE, and CREDIT experiments after evaluation on the test sets in terms of degree of communicative success, level of conventionality, and construction inventory size. The reported values represent the mean and 2 standard deviations computed over 10 independent experimental runs.

Scenario	Communicative success	Conventionality	Inventory size
CLEVR	99.65 ($\sim$ 0.13)	93.86 ($\sim$ 1.09)	46.72 ($\sim$ 2.45)
WINE	99.74 ($\sim$ 0.15)	88.67 ($\sim$ 1.92)	52.67 ($\sim$ 2.93)
CREDIT	99.67 ($\sim$ 0.13)	87.72 ($\sim$ 2.50)	51.43 ($\sim$ 2.49)

same, with the additional constraints that agents can no longer create new constructions to add to their inventories and that construction scores and concept representations are no longer updated. The test results are reported in Table 1 for all three scenarios. Degrees of communicative success of over 99.5% are achieved in each scenario, validating that the emergent languages indeed generalize to previously unseen entities and scenes. Levels of conventionality range from 87.72% for CREDIT, over 88.67% for WINE, to 93.86% for CLEVR. These numbers show that the emergent languages are not only effective at solving the communicative tasks, but that the agents in the population also manage to effectively align their linguistic systems. The average construction inventory size of the agents ranges from 51.43 in the case of CLEVR to 52.67 in the case of WINE. While further analysis would be required to provide a more theoretically substantiated interpretation of the resulting construction inventory sizes, the reported numbers at least corroborate the finding that the agents converge on a manageable set of widely applicable concepts.

The emergent conventions consist of holistic constructions that associate linguistic forms with grounded concept representations. Figure 4 shows, for each scenario, a construction from the inventory of one agent that has reached the maximum entrenchment score of 1.0 after 1,000,000 interactions. The DEMOXU-CXN shown in Figure 4a emerged during the CLEVR experiment. It associates the form *demoxu* to a concept representation with three relevant dimensions, that is, dimensions with a strictly positive weight value ω. The `area` dimension represents the number of pixels within an object's perimeter, normalized on a scale from 0 to 1. The `bb-area` dimension represents the normalized number of pixels within an object's rectangular bounding box. Finally, the `rel-area` dimension represents the ratio between the number of pixels within an object's perimeter and the total number of pixels in the image. Converting the normalized values back to raw pixel counts, the means and standard deviations on the three dimensions indicate that the concept representation of the DEMOXU-CXN prototypically represents entities with 1,344 pixels within their perimeter (σ = 76.8 pixels), 1,574 pixels within their rectangular bounding box (σ = 115 pixels), and which cover about 1% of the image. An analysis of the agent's communicative behavior reveals that they use the construction 73% of the time to refer to small objects with a spherical shape. These are indeed objects that cover a relatively small area of the image and fill about 80% of their bounding box. The ZAPOSE-CXN shown in Figure 4b emerged during the WINE experiment. It associates the form *zapose* to a concept representation with a single relevant dimension, namely, the normalized residual sugar content of a wine sample. The prototypical value on this dimension corresponds to 12.34 g/l (σ = 1.39 g/l), a residual sugar content typically associated with medium dry wines. The ZISENI-CXN shown in Figure 4c emerged

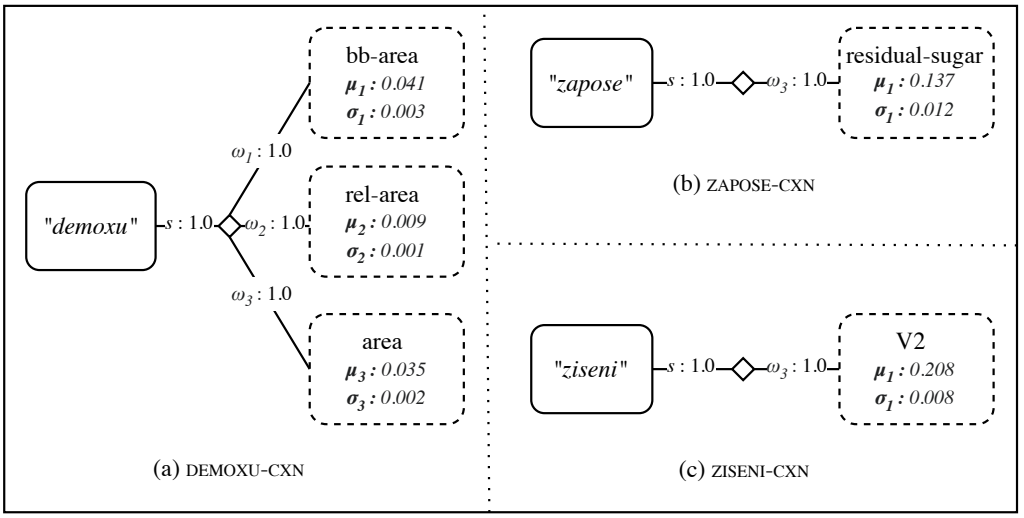

Figure 4
Examples of constructions that emerged in the CLEVR (a), WINE (b), and CREDIT (c) scenarios. The constructions associate atomic labels on their form side with situationally grounded and communicatively motivated concept representations on their meaning side. Figure adapted from Botoko Ekila et al. (2024b).

during the CREDIT experiment and associates the form *ziseni* to a low value range on the second PCA component of the transaction records.

In sum, the grounded concept learning experiment shows how a population of agents that take part in situated communicative interactions can self-organize a linguistic convention that can be used to refer to arbitrary entities in their environment. The convention consists of holistic constructions that associate atomic forms to concept representations. The meaningful, intentional, interactional, and communicative nature of the learning environment facilitated the establishment of a linguistic convention that is not only grounded in the (non-textual) environment of the agents, but which is also motivated and shaped by their communicative needs. While we have focused on the emergence of such a convention, agents that are added to the population at a later stage can acquire the established convention using the exact same adoption and alignment mechanisms while they take part in communicative interactions with more "mature" agents. As such, also existing natural languages can be learned in a constructivist manner through interactions with "tutor" agents that already master these languages, as will be demonstrated in the second experiment.

3.2.2 Experiment 2: Acquisition of Grammatical Structures. Through the second experiment, we aim to illustrate how linguistic structures of a higher morpho-syntactic and semantic complexity can be acquired in a communicative and situationally grounded learning environment, thereby transcending the level of atomic word forms and concept representations addressed in the first experiment. Concretely, we focus on a tutor-learner scenario in which the learner agent needs to acquire in a constructivist manner a linguistic system that is adequate to ask and answer English questions about the environment in which the agents are situated. The experiment, which has been extensively discussed

by Nevens et al. (2022) and Doumen, Beuls, and Van Eecke (2023), is set up as follows. The population consists of two agents, one being the tutor and the other being the learner. The communicative interactions between the two agents take place in randomly selected scenes from the CLEVR visual question answering dataset (Johnson et al. 2017). Each communicative interaction starts with the tutor asking an English question that is provided by the dataset for the selected scene. The task of the learner is to answer the question asked by the tutor. The learner starts without any linguistic knowledge apart from conceptual distinctions, such as different colors, materials, or sizes, which are assumed to have been acquired previously through a grounded concept learning experiment. The learner also possesses the ability to perform a number of primitive cognitive operations. These primitive operations can be thought of as the instruction set on which all complex reasoning processes need to build. In general, primitive operations can be cognitively inspired, rooted in theory of computation, or follow from a practical constraint such as a robot's API specification. In our experiment, we provide access to a number of basic operations, such as segmenting a scene, counting the number of elements in a set, computing unions and intersections of sets, querying attributes, and filtering sets according to prototypes.

Imagine that during their very first communicative interaction, the learner observes the question *How many blocks are there?*. The learner cannot understand the question and receives the answer to the question, say *4*, as feedback from the tutor. The learner then starts the process of intention reading. Based on the feedback provided by the tutor, the environment in which the utterance was observed, and the inventory of primitive cognitive operations that the agent can perform, the agent constructs a hypothesis about the intended meaning of the observed utterance. In order to do this, the agent searches for a network of primitive operations, referred to as a **procedural semantic network** (Woods 1968; Johnson-Laird 1977; Spranger et al. 2012; Verheyen et al. 2023), that, upon evaluation with respect to the scene, leads to the answer that was provided by the tutor. This network could, for example, consist of the operations `[segment image → filter cube → count]`. The mapping between the observed utterance *How-many-blocks-are-there?* and its hypothesized meaning `[segment image → filter cube → count]` is then stored by the learner agent as a holophrastic construction. Imagine that the learner later engages in a communicative interaction where they observe the utterance *How many spheres are there?*. Again, the learner cannot understand the question and needs to construct a hypothesis about the intended meaning of the utterance through the process of intention reading. Upon feedback by the tutor, in this case *3*, the learner constructs the hypothesis `[segment image → filter ball → count]`. Based on the previously acquired holophrastic construction that associates the form *How-many-blocks-are-there?* with the meaning `[segment image → filter cube → count]` and the current observation of the utterance *How-many-spheres-are-there?* and its hypothesized meaning `[segment image → filter ball → count]`, the learner can now use its pattern finding abilities to construct a more general item-based construction that pairs the form *How-many-Xs-are-there?* with the generalized meaning representation `[segment image → filter ?type → count]` in which *X* and `?type`, respectively, represent variable slots on the form and meaning sides of the construction. Importantly, the construction also captures that the referent of the filler of the *X*-slot on the form side will fill the `?type` slot on its meaning side. At the same time, the learner can acquire two holistic constructions that pair the forms *block* and *sphere* to their respective meanings, in this case the concepts of `cube` and `ball`. Along with the item-based construction and the two holistic constructions, the agent also learns that the BLOCK-CXN and the SPHERE-CXN provide suitable fillers for the X-slot in the HOW-MANY-XS-ARE-THERE-CXN. This information is added

to the agent's **categorial network**, which models emergent grammatical categories as links between constructional slots and their observed fillers, very much in the spirit of Radical Construction Grammar (Croft 2001). Pattern finding is implemented as an anti-unification-based generalization process (Plotkin 1970; Van Eecke 2018; Yernaux and Vanhoof 2019). A schematic representation of the intention reading and pattern finding processes involved in this example is shown in Figure 5. Note that the constructions that are learned constitute form-meaning mappings that can be used for both language comprehension, that is, mapping from utterances to their meaning representations, and language production, that is, mapping from meaning representations to utterances that express them.

As discussed in Section 1 in the context of human language acquisition, intention reading is an abductive reasoning process that introduces uncertainty into the overall learning process. In this experiment, meaning hypotheses are guaranteed to be consistent and plausible given the communicative interaction at hand, but might not generalize to other situations. This uncertainty directly percolates to the holophrastic, item-based, and holistic constructions that are learned. The evolutionary dynamics that

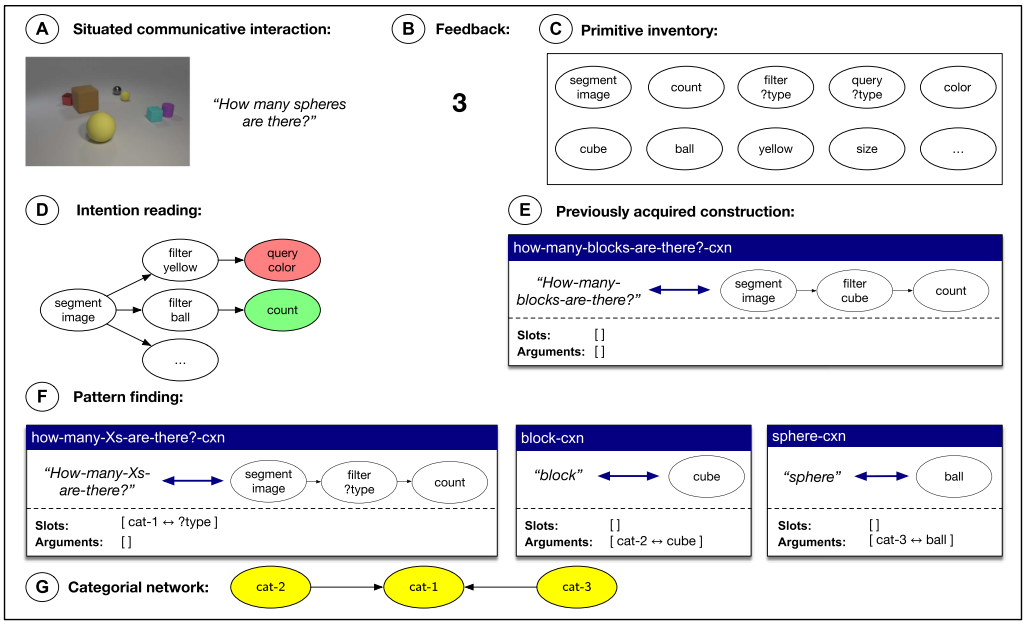

Figure 5
Schematic overview of the acquisition of an item-based construction and two holistic constructions through the processes of intention reading and pattern finding during a situated communicative interaction. An agent observes a question about its environment (A), cannot understand it, and receives the answer to the question as feedback (B). Based on its inventory of primitive operations (C), the agent uses its intention reading capabilities to make a hypothesis about the intended meaning of the observed utterance, which, upon evaluation with respect to the environment, is consistent with the feedback (D). Based on the observed utterance, the meaning hypothesis and a similar yet not identical construction that was previously acquired (E), the agent then applies its pattern finding capabilities to learn an item-based construction and two holistic constructions (F), as well as categorial relations between the constructional slots and their observed fillers (G).

overcome this uncertainty are modeled on the level of constructions through similar alignment dynamics as those discussed in Section 3.2.1 in the context of the grounded concept learning experiment. Concretely, all constructions carry an entrenchment score, which reflects the frequency of their past successes and failures in communication. During language comprehension and production, constructions with a higher score will be preferred over less entrenched constructions. At the end of a successful communicative interaction, the learner agent increases the score of the constructions that were used and decreases the score of their competitors (i.e., other constructions that would have led to successful communication). In the case of a failed interaction, the scores of the constructions that were used are decreased. As a consequence of these evolutionary dynamics, constructions that are applicable in a wider range of situations gradually gain the upper hand over constructions that either compete with them or hurt communication rather than support it. Not only does this provide a way to overcome suboptimal hypotheses resulting from the intention reading process, it also causes more general and abstract constructions to ultimately prevail over more specific ones.

The learning dynamics of the experiment are visualized in Figure 6, where the degree of communicative success and the construction inventory size are plotted as a function of the number of communicative interactions that have taken place. The degree of communicative success starts at 0 and reaches 1 after about 25,000 interactions. The construction inventory size exhibits the typical "overshoot pattern" that was also found in the grounded concept learning experiment. In the earlier stages of the experiment, many new constructions are learned. These constructions are often holophrastic constructions or item-based constructions with few variable slots. Moreover, many

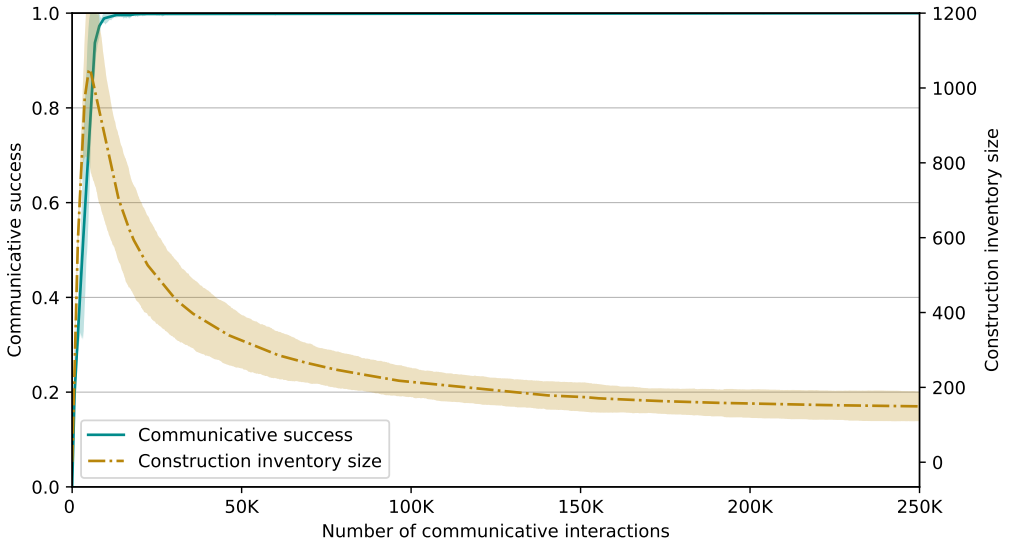

Figure 6
Learning dynamics of the experiment on the acquisition of grammatical structures through intention reading and pattern finding during situated communicative interactions. The degree of communicative success and the construction inventory size are plotted as a function of the number of communicative interactions that have taken place. This graph has been created based on experimental data from Nevens et al. (2022).

constructions result from pattern finding operations over suboptimal meaning hypotheses. As more and more interactions take place in a variety of environments, the evolutionary dynamics of strengthening and weakening constructions based on their successful or unsuccessful use in communication ensures that suboptimal constructions gradually disappear from the construction inventory of the learner agent.

In sum, the grammar acquisition experiment shows how linguistic structures that transcend the level of atomic word forms and concept representations can be acquired through the processes of intention reading and pattern finding during situated communicative interactions. The linguistic knowledge that is built up consists in an inventory of constructions at different levels of abstraction. These form-meaning mappings correspond to syntactico-semantic generalizations over the compositional and non-compositional aspects of language use observed in meaningful, intentional, and situated environments. The meaning side of the acquired constructions is thereby grounded in the physio-cognitive endowment of the agents, their shared environment, and their communicative tasks and intentions. The evolutionary dynamics of strengthening and weakening constructions based on their successful or unsuccessful use in communication provides a way to overcome the uncertainty involved in working out the underlying meanings and intentions of linguistic utterances.

The two experiments discussed in this section illustrate that the linguistic systems built up by artificial agents during situated communicative interactions are fundamentally different from language models learned through a text prediction task. For one thing, the agents' linguistic knowledge is directly grounded in their physio-cognitive endowment and environment. The constructions they acquire are motivated by their function in solving communicative tasks rather than by the observed collocational behavior of linguistic forms. For another, the processes of language comprehension and production no longer rely on (sub)word prediction, but on finding combinations of constructions that optimally map from linguistic forms to their meaning representations and vice versa. Due to their grounded nature, these meaning representations can be conceptualized and interpreted with respect to the communicative tasks and environment of the agents. As a consequence, the limitations of LLMs that result from their generative nature do not apply as such to the linguistic systems that are acquired in this setting. The agents do not hallucinate in the sense that LLMs do. They might well fail to achieve communicative success, but the utterances they produce are always motivated by more factors than the collocational behavior of the linguistic forms that they contain. The situated and interactional nature of the learning environment ensures that the constructions learned by the agents are communicatively motivated, which will ultimately be required to draw human-like logical and pragmatic inferences (see Section 2). The richness of the learning environment also lifts the constraint that all aspects of meaning, reasoning, and world knowledge need to be learned through the intermediary of their effect on the distribution of (sub)words, bearing the promise to avoid the need for exposure to hundreds of billions of tokens. Finally, the linguistic systems that are acquired will still mirror the biases present in the learning environment, but these will likely be easier to mitigate than those resulting from uncuratable amounts of textual data. By no means is this discussion intended to dismiss the impressive results obtained through the LLM paradigm, nor would we claim that achieving similar results through agent-based simulations is within reach in the near future. We do argue, however, that more faithfully modeling the situated, communicative, and interactional environments in which human languages are acquired provides a promising path to overcome the limitations of systems that essentially rely on the distributional hypothesis, as powerful as it might be.

4. Conclusion

The primary argument of this article has been that the way in which humans acquire their native languages is fundamentally different from the way in which LLMs learn their linguistic capabilities. We have argued that these differences have consequential repercussions on the linguistic knowledge that is built up, as well as on how this knowledge can be used to drive the processes of language comprehension and production.

We have highlighted the fact that humans learn language by actively taking part in meaningful and intentional communicative interactions that are situated in their everyday environment. During these communicative interactions, they make hypotheses about the meanings and communicative intentions that underlie the utterances they perceive, relying on clues provided by the communicative environment in which the interactions take place. Constructions of different degrees of abstraction are learned as syntactico-semantic generalizations over combinations of perceived utterances, their reconstructed meanings and intentions, and previously acquired constructions. As such, humans rely on their situation-based intention reading and syntactico-semantic pattern finding capabilities to bootstrap a productive linguistic system that is tied to their own physical and cognitive endowment, grounded in their environment, shaped by their past experiences, and motivated by their communicative needs.

By contrast, LLMs acquire their linguistic capabilities by learning to predict (sub)words based on the textual environment in which they appear. They meticulously capture the (sub)words' collocational behavior in a fine-grained manner and on an inconceivable scale. Stretching the distributional hypothesis to previously unimaginable dimensions, they are able to generate texts that do not only exhibit human-like lexical and morpho-syntactic correctness, fluency, and eloquence, but that also exhibit a remarkable discourse coherence and semantic adequacy. The models even seem to be capable of solving natural language understanding tasks that require substantial non-linguistic reasoning abilities.

However, the fact that LLMs are learned in the absence of meaningful and intentional situated communicative interactions lies at the root of a number of inherent limitations. A first limitation concerns their so-called hallucinations. Due to their grounding in text generation, by which they use collocation-based patterns and structures observed in input texts to generate output texts that exhibit similar patterns and structures, they seamlessly mix fact and fiction in the output they produce. A second limitation concerns their struggle to perform human-like logic and pragmatic reasoning. For one thing, this struggle is ascribable to a lack of situational grounding during the training process, by which the distribution of objects and events in the world about which they need to reason differs from the distribution of (sub)words in the texts they are trained on. For another, it can be ascribed to the absence of communicative intent during the training process. While LLMs have been trained to capture fine-grained collocational patterns and structures in enormous corpora of language use, they have never had access to clues about why particular utterances were produced in the first place. A third limitation concerns their data-hungriness, which results from the fact that all aspects of meaning, reasoning, and world knowledge that they are expected to capture need to be learned indirectly through the intermediary of their effect on the distribution of (sub)words in corpora. Finally, LLMs are sensitive to biases, as they mirror the collocational patterns and structures present in uncuratable amounts of human-written texts.

Our comparison between the ways in which humans and LLMs learn their linguistic capabilities along with our analysis of the limitations inherent to LLMs have led us to conclude that the key towards more human-like language learning in machines lies

in more faithfully modeling the meaningful, intentional, situated, communicative, and interactional aspects of human linguistic communication. In order to investigate how this could be operationalized in practice, we have first discussed approaches that extend the LLM paradigm beyond a purely text-based prediction task. The integration of input from other modalities than text provides a way to incorporate richer representations of the situational embedding of observed utterances, whereas the alignment of language models to text-external criteria confers the possibility of integrating task-oriented, communicative, and interactional objectives into the training process of LLMs. While these extensions have already proven to be valuable additions to the LLM paradigm, the extent to which situated communicative interactions can be approximated by non-gameable reward functions still remains an open question.

We have then moved on to discuss a line of work that leaves the text-based prediction paradigm altogether and takes situated communicative interactions as the starting point for modeling language acquisition in machines. We have discussed two experiments that model how artificial agents can acquire a language through the processes of intention reading and pattern finding during situated communicative interactions. The first experiment focused on grounded concept learning and shows how a population of autonomous agents can self-organize a linguistic convention that can be used to refer to arbitrary entities in their environment. The meaningful, intentional, interactional, and communicative nature of the learning environment facilitates the emergence and evolution of an environmentally grounded and communicatively motivated convention. The second experiment focused on the acquisition of linguistic structures that transcend the level of atomic forms and concept representations. It implements a tutor-learner scenario in which a learner agent acquires linguistic knowledge that is adequate to ask and answer English questions about their environment. This linguistic knowledge consists of constructions at different levels of abstraction, which can combine to map between English utterances and situationally grounded executable meaning representations. These meaning representations are grounded in the physio-cognitive endowment of the agents, their shared environment, and their communicative tasks and intentions.

The experiments have shown that the linguistic systems that are built up by the agents through situated communicative interactions are of a fundamentally different nature than the linguistic knowledge that is captured by LLMs. The agents' acquired constructions are directly motivated by their communicative function, rather than by the collocational behavior of linguistic forms. The constructions can thereby be used to comprehend and produce linguistic expressions without relying on textual prediction, and the meanings of linguistic expressions can be conceptualized and interpreted with respect to the communicative tasks and environment of the agents. They thereby steer clear of the hallucination effects that characterize text prediction-based models. As a result of the rich nature of the learning environment, constructions can be acquired in a more data-efficient manner and they are able to capture the communicative function of language, which will ultimately be required to draw human-like logical and pragmatic inferences.

The experiments that we have discussed in the final part of this article definitely do not offer an application-level alternative to today's LLMs, nor were they ever designed with that goal in mind. We hope, however, to have convinced the reader that a better integration of the situated, communicative, and interactional aspects of human linguistic communication constitutes the key to overcoming the limitations of current LLMs, and that more faithfully modeling the situated communicative interactions through which humans acquire their native languages provides a promising path towards more human-like language processing in machines.

Acknowledgments

We would like to thank Jérôme Botoko Ekila and Jens Nevens for their assistance in creating the graphs shown in Figures 3 and 6. We are grateful to Marie-Catherine de Marneffe and Remi van Trijp for their role in the discussions that led up to the writing of this article, and to Lara Verheyen, Remi van Trijp, and three anonymous reviewers for their insightful feedback on earlier versions of the manuscript. The research reported on in this article was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 951846, the Flemish Government under the Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen programme, the AI Flagship project ARIAC by DigitalWallonia4.ai, and the F.R.S.-FNRS-FWO WEAVE project HERMES I under grant numbers T002724F (F.R.S.-FNRS) and G0AGU24N (FWO).

References

Apidianaki, Marianna. 2022. From word types to tokens and back: A survey of approaches to word meaning representation and interpretation. *Computational Linguistics*, 49(2):465–523.

Aslin, Richard N. 2017. Statistical learning: A powerful mechanism that operates by mere exposure. *Wiley Interdisciplinary Reviews: Cognitive Science*, 8(1–2):e1373. https://doi.org/10.1002/wcs.1373, PubMed: 27906526

Austin, John L. 1962. *How to Do Things with Words*. Oxford University Press, London.

Baevski, Alexei Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312.

Bai, Yuntao, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pages 238–247. https://doi.org/10.3115/v1/P14-1023

Batali, John. 1998. Computational simulations of the emergence of grammar. In Chris Knight, James R. Hurford, Michael Studdert-Kennedy, editors, *Approaches to the Evolution of Language: Social and Cognitive Bases*. Cambridge University Press, Cambridge, UK, pages 405–426.

Behrens, Heike. 2021. Constructivist approaches to first language acquisition. *Journal of Child Language*, 48(5):959–983. https://doi.org/10.1017/S0305000921000556, PubMed: 34382923

Bender, Emily M. and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

BigScience Workshop. 2022. BLOOM: A 176B-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Blythe, Richard A. and William Croft. 2012. S-curves and the mechanisms of propagation in language change. *Language*, 88:269–304. https://doi.org/10.1353/lan.2012.0027

Boden, Margaret A. 1978. Artificial intelligence and Piagetian theory. *Synthese*, 38(3):389–414. https://doi.org/10.1007/BF00486637

Botoko Ekila, Jérôme, Jens Nevens, Lara Verheyen, Katrien Beuls, and Paul Van Eecke. 2024a. Decentralised emergence of robust and adaptive linguistic conventions in populations of autonomous agents grounded in continuous worlds. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multi-Agent Systems*, pages 2168–2170.

Botoko Ekila, Jérôme, Jens Nevens, Lara Verheyen, Katrien Beuls, and Paul Van Eecke. 2024b. Decentralised emergence of robust and adaptive linguistic conventions in populations of autonomous agents grounded in continuous worlds. *arXiv preprint arXiv:2401.08461*.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information*

Processing Systems 33 (NeurIPS 2020), pages 1877–1901.
Bruner, Jerome. 1983. *Learning to use language*. Oxford University Press, Oxford, United Kingdom.
Bybee, Joan. 2010. *Language, Usage and Cognition*. Cambridge University Press, Cambridge, United Kingdom. https://doi.org/10.1017/CBO9780511750526
Chang, Tyler A. and Benjamin K. Bergen. 2024. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1):293–350. https://doi.org/10.1162/coli_a_00492
Choudhury, Sagnik Ray, Anna Rogers, and Isabelle Augenstein. 2022. Machine reading, fast and slow: When do models "understand" language? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 78–93.
Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.
Clark, Herbert H. 1996. *Using Language*. Cambridge University Press, Cambridge, United Kingdom. https://doi.org/10.1017/CBO9780511620539
Coelho Mollo, Dimitri and Raphaël Millière. 2023. The vector grounding problem. *arXiv preprint arXiv:2304.01481v1*.
Cortez, Paulo, Antonio Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. 2009. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553. https://doi.org/10.1016/j.dss.2009.05.016
Croft, William. 1991. *Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information*. University of Chicago Press, Chicago, IL, USA.
Croft, William. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press, Oxford, United Kingdom. https://doi.org/10.1093/acprof:oso/9780198299554.001.0001
Dal Pozzolo, Andrea, Olivier Caelen, Yann-Aël Le Borgne, Serge Waterschoot, and Gianluca Bontempi. 2014. Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, 41(10):4915–4928. https://doi.org/10.1016/j.eswa.2014.02.026
Darwin, Charles R. 1871. *The Descent of Man, and Selection in Relation to Sex*, 1st edition, volume 1. John Murray, London, United Kingdom. https://doi.org/10.1037/12293-000
Dasgupta, Ishita, Andrew K Lampinen, Stephanie C. Y. Chan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051v1*.
Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171-4186.
Diessel, Holger. 2017. Usage-based linguistics. In Mark Aronoff, editor, *Oxford Research Encyclopedia of Linguistics*. Oxford University Press, Oxford, United Kingdom. https://doi.org/10.1093/acrefore/9780199384655.013.363
Doumen, Jonas, Katrien Beuls, and Paul Van Eecke. 2023. Modelling language acquisition through syntactico-semantic pattern finding. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1317–1327. https://doi.org/10.18653/v1/2023.findings-eacl.99
Doumen, Jonas, Veronica J. Schmalz, Katrien Beuls, and Paul Van Eecke. 2025. The computational learning of construction grammars: State of the art and prospective roadmap. *Constructions and Frames*, 17. https://arxiv.org/pdf/2407.07606
Driess, Danny, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. PaLM-E: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, pages 8469–8488.
Fan, Yue, Jing Gu, Kaizhi Zheng, and Xin Wang. 2023. R2H: Building multimodal navigation helpers that respond to help requests. In *Proceedings of the 2023*

Conference on Empirical Methods in Natural Language Processing, pages 14803–14819. https://doi.org/10.18653/v1/2023.emnlp-main.915

Firth, John R. 1957. A synopsis of linguistic theory, 1930–1955. In *Studies in Linguistic Analysis*. Basil Blackwell, Oxford, pages 1–31.

Frank, Michael C. 2023. Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, 27(11):990–992. https://doi.org/10.1016/j.tics.2023.08.007, PubMed: 37659919

Gabrieli, Gianmarco, Michal Muszynski, Edouard Thomas, David Labbe, and Patrick W. Ruch. 2022. Accelerated estimation of coffee sensory profiles using an AI-assisted electronic tongue. *Innovative Food Science & Emerging Technologies*, 82:103205. https://doi.org/10.1016/j.ifset.2022.103205

Givón, Talmy. 1995. *Functionalism and Grammar*. John Benjamins, Amsterdam, Netherlands. https://doi.org/10.1075/z.74

Goldberg, Adele E. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press, Oxford, United Kingdom.

Goodhart, Charles. 1975. Problems of monetary management: The U.K. experience. In *Papers in Monetary Economics*. Reserve Bank of Australia, Sydney, pages 1–20.

Grice, Paul. 1967. Logic and conversation. In Paul Grice, editor, *Studies in the Way of Words*. Harvard University Press, Cambridge, MA, USA, pages 41–58.

Harris, Zellig S. 1954. Distributional structure. *Word*, 10(2–3):146–162. https://doi.org/10.1080/00437956.1954.11659520

Hill, Felix, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695. https://doi.org/10.1162/COLI_a_00237

Hockett, Charles F. and Charles D. Hockett. 1960. The origin of speech. *Scientific American*, 203(3):88–97. https://doi.org/10.1038/scientificamerican0960-88

Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pages 30016–30030.

Hong, Xudong, Margarita Ryzhova, Daniel Adrian Biondi, and Vera Demberg. 2023. Do large language models and humans have similar behaviors in causal inference with script knowledge? *arXiv preprint arXiv*.07311.

Hu, Jennifer, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213. https://doi.org/10.18653/v1/2023.acl-long.230

Hupkes, Dieuwke. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926. https://doi.org/10.1613/jair.1.11196

Jiang, Nanjiang, and Marie-Catherine de Marneffe. 2021. He thinks he knows better than the doctors: BERT for event factuality fails on pragmatics. *Transactions of the Association for Computational Linguistics*, 9:1081–1097. https://doi.org/10.1162/tacl_a_00414

Johnson, Justin, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2901–2910. https://doi.org/10.1109/CVPR.2017.215

Johnson-Laird, Philip N. 1977. Procedural semantics. *Cognition*, 5(3):189–214. https://doi.org/10.1016/0010-0277(77)90001-4

Joos, Martin. 1950. Description of language design. *The Journal of the Acoustical Society of America*, 22(6):701–707. https://doi.org/10.1121/1.1906674

Jurafsky, Daniel and James H. Martin. 2024. Vector semantics and embeddings. In *Speech and Language Processing: An Introduction to Natural Language Processing,*

Computational Linguistics, and Speech Recognition with Language Models. 3rd ed.
Kravtchenko, Ekaterina and Vera Demberg. 2022. Informationally redundant utterances elicit pragmatic inferences. *Cognition*, 225:105159. https://doi.org/10.1016/j.cognition.2022.105159, PubMed: 35580451
Kuhl, Patricia K., Feng-Ming Tsao, and Huei-Mei Liu. 2003. Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences*, 100(15):9096–9101. https://doi.org/10.1073/pnas.1532872100, PubMed: 12861072
Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar: Theoretical Prerequisites*, volume 1. Stanford University Press, Stanford CA, USA.
Lauriola, Ivano, Alberto Lavelli, and Fabio Aiolli. 2022. An introduction to deep learning in natural language processing: Models, techniques, and tools. *Neurocomputing*, 470(C):443–456. https://doi.org/10.1016/j.neucom.2021.05.103
Lazaridou, Angeliki, and Marco Baroni. 2020. Emergent multi-agent communication in the deep learning era. *arXiv preprint arXiv:2006.02419*.
Li, KunChang, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. VideoChat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.
Lieven, Elena. 2014. First language learning from a usage-based approach. In Thomas Herbst, Hans-Jörg Schmid, and Susen Faulhaber, editors, *Constructions Collocations Patterns*. De Gruyter Mouton, Berlin, pages 9–32. https://doi.org/10.1515/9783110356854.9
Liu, Ruibo, Ge Zhang, Xinyu Feng, and Soroush Vosoughi. 2022. Aligning generative language models with human values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 241–252. https://doi.org/10.18653/v1/2022.findings-naacl.18
Löhr, Guido. 2022. What are abstract concepts? On lexical ambiguity and concreteness ratings. *Review of Philosophy and Psychology*, 13(3):549–566. https://doi.org/10.1007/s13164-021-00542-9
Lyu, Chenyang, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-LLM: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*.
MacWhinney, Brian. 2014. Item-based patterns in early syntactic development. In Thomas Herbst, Hans-Jörg Schmid, and Susen Faulhaber editors, *Constructions Collocations Patterns*. De Gruyter Mouton, Berlin, pages 33–69. https://doi.org/10.1515/9783110356854.33
Mandelkern, Matthew and Tal Linzen. 2024. Do language models refer? *Computational Linguistics*, 50(3):1191–1200. https://doi.org/10.1162/coli_a_00522
Manheim, David and Scott Garrabrant. 2018. Categorizing variants of Goodhart's Law. *arXiv preprint arXiv:1803.04585*.
Maynard Smith, John and Eörs Szathmáry. 1999. *The Origins of Life: From the Birth of Life to the Origin of Language*. Oxford University Press, Oxford, United Kingdom. https://doi.org/10.1093/oso/9780198504931.001.0001
McCoy, R. Thomas. Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L. Griffiths. 2023. Embers of autoregression: Understanding large language models through the problem they are trained to solve. *arXiv preprint arXiv:2309.13638*.
McKenna, Nick, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2758–2774. https://doi.org/10.18653/v1/2023.findings-emnlp.182
Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations (ICLR 2013), Workshop Proceedings*.
Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pages 1–9.
Mitchell, Melanie and David C. Krakauer. 2023. The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120. https://doi.org/10.1073/pnas.2215907120, PubMed: 36943882

Mitchell, Melanie, Alessandro B. Palmarini, and Arsenii Kirillovich Moskvichev. 2024. Comparing humans, GPT-4, and GPT-4v on abstraction and reasoning tasks. In *AAAI 2024 Workshop on "Are Large Language Models Simply Causal Parrots?"*, 9 pages. https://llmcp.cause-lab.net/pdf/LLMCP_4.pdf.

Navigli, Roberto, Simone Conia, and Bjorn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *Journal of Data and Information Quality*, 15(2):1–21. https://doi.org/10.1145/3597307

Nelson, Katherine. 1998. *Language in Cognitive Development: The Emergence of the Mediated Mind*. Cambridge University Press, Cambridge, United Kingdom.

Nevens, Jens, Jonas Doumen, Paul Van Eecke, and Katrien Beuls. 2022. Language acquisition through intention reading and pattern finding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 15–25.

Nevens, Jens, Paul Van Eecke, and Katrien Beuls. 2020. From continuous observations to symbolic concepts: A discrimination-based strategy for grounded concept learning. *Frontiers in Robotics and AI*, 7(84):1–20. https://doi.org/10.3389/frobt.2020.00084, PubMed: 33501251

Ngo, Richard, Lawrence Chan, and Sören Mindermann. 2024. The alignment problem from a deep learning perspective: A position paper. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*.

Nissim, Malvina, Rik van Noord, and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497. https://doi.org/10.1162/coli_a_00379

Oliphant, Michael. 1999. The learning barrier: Moving from innate to learned systems of communication. *Adaptive Behavior*, 7(3–4):371–383. https://doi.org/10.1177/105971239900700309

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Park, Kyungseo, Hyunwoo Yuk, M. Yang, Junhwi Cho, Hyosang Lee, and Jung Kim. 2022. A biomimetic elastomeric robot skin using electrical impedance and acoustic tomography for tactile sensing. *Science Robotics*, 7(67):eabm7187. https://doi.org/10.1126/scirobotics.abm7187, PubMed: 35675452

Pavlick, Ellie. 2023. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A*, 381(2251):20220041. https://doi.org/10.1098/rsta.2022.0041, PubMed: 37271171

Piaget, Jean. 1923. *Le langage et la pensée chez l'enfant*. Delachaux & Niestlé, Neuchâtel/Paris, Switzerland/France.

Piantadosi, Steven T. and Felix Hill. 2022. Meaning without reference in large language models. *arXiv preprint arXiv:2208.02957v2*.

Plotkin, Gordon D. 1970. A note on inductive generalization. *Machine Intelligence*, 5(1):153–163.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763.

Rehder, Bob, Missy E. Schreiner, Michael B. W. Wolfe, Darrell Laham, Thomas K. Landauer, and Walter Kintsch. 1998. Using latent semantic analysis to assess knowledge: Some technical considerations. *Discourse Processes*, 25(2–3):337–354. https://doi.org/10.1080/01638539809545031

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866. https://doi.org/10.1162/tacl_a_00349

Ruis, Laura Eline, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by LLMs. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, pages 20827–20905.

Saffran, Jenny R., Richard N. Aslin, and Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science*,

274(5294):1926–1928. https://doi.org/10.1126/science.274.5294.1926, PubMed: 8943209

Schleicher, August. 1869. *Darwinism Tested by the Science of Language. English Translation of Schleicher 1863, translated by Alex V. W. Bikkers*. John Camden Hotten, London, United Kingdom. https://doi.org/10.5962/bhl.title.49464

Schütze, Hinrich. 1992. Dimensions of meaning. In *Proceedings of the 1992 ACM/IEEE Conference on Supercomputing*, pages 787–796. https://doi.org/10.1109/SUPERC.1992.236684

Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Shanahan, Murray. 2024. Talking about large language models. *Communication of the ACM*, 67(2):68–79. https://doi.org/10.1145/3624724

Shichman, Mollie, Claire Bonial, Austin Blodgett, Taylor Hudson, Francis Ferraro, and Rachel Rudinger. 2023. Use defines possibilities: Reasoning about object function to interpret and execute robot instructions. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 284–292.

Shiffrin, Richard and Melanie Mitchell. 2023. Probing the psychology of AI models. *Proceedings of the National Academy of Sciences*, 120(10):e2300963120. https://doi.org/10.1073/pnas.2300963120, PubMed: 36857344

Skalse, Joar, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pages 9460–9471.

Sperber, Dan and Deirdre Wilson. 1986. *Relevance: Communication and cognition*. Harvard University Press, Cambridge, MA, USA.

Spranger, Michael, Simon Pauw, Martin Loetzsch, and Luc Steels. 2012. Open-ended procedural semantics. In Luc Steels and Manfred Hild, editors, *Language Grounding in Robots*. Springer, New York, NY, USA, 153–172. https://doi.org/10.1007/978-1-4614-3064-3_8

Steels, Luc. 1995. A self-organizing spatial vocabulary. *Artificial Life*, 2(3):319–332. https://doi.org/10.1162/artl.1995.2.3.319, PubMed: 8925502

Steels, Luc. 2011. Modeling the cultural evolution of language. *Physics of Life Reviews*, 8(4):339–356. https://doi.org/10.1016/j.plrev.2011.10.014, PubMed: 22071322

Sullivan, Jessica, Michelle Mei, Andrew Perfors, Erica Wojcik, and Michael C. Frank. 2021. SAYCam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open Mind*, 5:20–29. https://doi.org/10.1162/opmi_a_00039, PubMed: 34485795

Sutskever, Ilya, Oriol Vinyals, and Quoc Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, pages 3104–3112.

Tan, Xiaoyu, Shaojie Shi, Xihe Qiu, Chao Qu, Zhenting Qi, Yinghui Xu, and Yuan Qi. 2023. Self-criticism: Aligning large language models with their understanding of helpfulness, honesty, and harmlessness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 650–662. https://doi.org/10.18653/v1/2023.emnlp-industry.62

Tomasello, Michael. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, Harvard, MA, USA.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Trott, Sean, Cameron Jones, Tyler Chang, James Michaelov, and Benjamin Bergen. 2023. Do large language models know what humans know? *Cognitive Science*, 47:e13309. https://doi.org/10.1111/cogs.13309, PubMed: 37401923

Turney, Peter D. and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188. https://doi.org/10.1613/jair.2934

Vallor, Shannon. 2024. *The AI Mirror: How to Reclaim our Humanity in an Age of Machine Thinking*. Oxford University Press, Oxford, United Kingdom. https://doi.org/10.1093/oso/9780197759066.001.0001

Van Eecke, Paul. 2018. *Generalisation and Specialisation Operators for Computational Construction Grammar and Their Application in Evolutionary Linguistics Research*. Ph.D.

thesis, Vrije Universiteit Brussel, Brussels: VUB Press.

Van Eecke, Paul, Lara Verheyen, Tom Willaert, and Katrien Beuls. 2023. The Candide model: How narratives emerge where observations meet beliefs. In *Proceedings of the 5th Workshop on Narrative Understanding (WNU)*, pages 48–57. https://doi.org/10.18653/v1/2023.wnu-1.7

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 6000–6010.

Verheyen, Lara, Jérôme Botoko Ekila, Jens Nevens, Paul Van Eecke, and Katrien Beuls. 2023. Neuro-symbolic procedural semantics for reasoning-intensive visual dialogue tasks. In *Proceedings of the 26th European Conference on Artificial Intelligence (ECAI 2023)*, pages 2419–2426. https://doi.org/10.3233/FAIA230544

Vong, Wai Keen, Wentao Wang, A. Emin Orhan, and Brenden M. Lake. 2024. Grounded language acquisition through the eyes and ears of a single child. *Science*, 383(6682):504–511. https://doi.org/10.1126/science.adi1374, PubMed: 38300999

Vulić, Ivan, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. 2020. Multi-SimLex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897. https://doi.org/10.1162/coli_a_00391

Vulić, Ivan, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240. https://doi.org/10.18653/v1/2020.emnlp-main.586

Webb, Taylor, Keith J. Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541. https://doi.org/10.1038/s41562-023-01659-w, PubMed: 37524930

Weidinger, Laura, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Weissweiler, Leonie, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. The better your syntax, the better your semantics? Probing pretrained language models for the English comparative correlative. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10882. https://doi.org/10.18653/v1/2022.emnlp-main.746

Welford, Barry Payne. 1962. Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3):419–420. https://doi.org/10.1080/00401706.1962.10490022

West, Peter, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D. Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, Benjamin Newman, Pang Wei Koh, Allyson Ettinger, and Yejin Choi. 2024. The generative AI paradox: "What it can create, it may not understand." In *The Twelfth International Conference on Learning Representations*.

Woods, William A. 1968. Procedural semantics for a question-answering machine. In *Proceedings of the December 9–11, 1968, Fall Joint Computer Conference, Part I*, pages 457–471, New York, NY, USA. https://doi.org/10.1145/1476589.1476653

Yamazaki, Takato, Tomoya Mizumoto, Katsumasa Yoshikawa, Masaya Ohagi, Toshiki Kawamoto, and Toshinori Sato. 2023. An open-domain avatar chatbot by exploiting a large language model. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 428–432. https://doi.org/10.18653/v1/2023.sigdial-1.40

Yernaux, Gonzague and Wim Vanhoof. 2019. Anti-unification in constraint logic programming. *Theory and Practice of Logic Programming*, 19(5–6):773–789. https://doi.org/10.1017/S1471068419000188

Ziegler, Daniel M., Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.