Loïs Dona*, Aslı Özyürek, Judith Holler and Limor Raviv

Investigating the role of pragmatic visual signals in language emergence using communication games and agent-based models

https://doi.org/10.1515/cog-2026-0009
Received January 16, 2026; accepted September 7, 2026; published online October 12, 2026

Abstract: Studying the emergence of language has often relied on communication games and agent-based models, with a strong focus on referential linguistic signals, often overlooking pragmatic signals that are frequently expressed visually, such as facial signals. These signals are a crucial component of natural interaction, supporting repair, coordination, and negotiation of meaning. We outline existing language emergence experiments, highlighting findings from communication games and research on pragmatic facial signals in interaction, demonstrating their relevance to the emergence of communication systems. Furthermore, computational models of language emergence, such as agent-based models, largely overlook that communication is layered across multiple communication channels. We argue that incorporating pragmatic visual signals into these approaches can help to provide a more nuanced understanding of the processes underlying language emergence in natural, multimodal interaction. Following these insights, we propose two complementary directions for future research. First, communication game paradigms can be adapted to systematically analyze visual pragmatic signals, linking them to existing commonly used measures such as convergence, communicative success, and adaptation of linguistic forms. Second, agent-based models can be extended to multichannel frameworks in which agents produce and interpret both referential and pragmatic signals, enabling investigation of how layered multimodal information shapes language emergence.

Keywords: multimodality; pragmatics; agent-based models; language evolution; facial signals; communication games

***Corresponding author: Loïs Dona**, Max Planck institute for Psycholinguistics, Nijmegen, The Netherlands, E-mail: lois.dona@mpi.nl. https://orcid.org/0000-0002-2867-2066
Aslı Özyürek, Judith Holler and Limor Raviv, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands; and Donders Institute for Brain, Cognition and Behavior, Radboud University, Nijmegen, The Netherlands. https://orcid.org/0000-0003-4914-2643 (A. Özyürek). https://orcid.org/0000-0003-0671-6651 (J. Holler). https://orcid.org/0000-0002-0716-3553 (L. Raviv)

Open Access. © 2026 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

1 Introduction

Studying the emergence of language is typically done using paradigms that simulate the process of interaction, learning, and diffusion of linguistic variants. Here we focus on two prevalent methods of simulating interaction: (1) laboratory communication games, where human participants create novel communication systems in controlled settings (Müller and Raviv 2025), and (2) agent-based computational models, which allow researchers to explore how linguistic conventions emerge and evolve among artificial agents (Ohmer & Cuskley 2025; Steels 2011). These paradigms can be used to test the role of different communicative pressures, as well as various cognitive and social factors that can impact the emergence of language. But to what extent do these experimental and computational paradigms actually incorporate relevant aspects of face-to-face communication?

In this review, we discuss research in these fields and highlight a significant blind spot: the limited consideration of pragmatic (also referred to as meta-communicative) visual signals in language emergence research. By these, we mean things like eyebrow raises, frowns, smiles, pragmatic hand-gestures, etc., which can support pragmatic inferences alongside referential signals. We here use the term "signal" to refer to bodily movements produced during communicative interaction, regardless of whether or not these movements are intentional. There is some evidence that people mark communicative intention when gesturing (Slonimska et al. 2026). Nevertheless, teasing apart intentional from unintentional signals, and deciding whether a signal carries communicative function remains a well known challenge in the field (Ben Mocha and Burkart 2021; Derchi et al. 2023; Ekman and Friesen 1969; Perusquía-Hernández et al. 2019). Previous experimental work has investigated this issue by manipulating addressee presence or visibility (e.g., Bavelas et al. 2008). However, such manipulations do not provide a definitive measure of communicative intent.[1] Moreover, the pragmatic function of a signal depends not only on the producer's intention but also on how it is interpreted by the receiver and the broader interactional context. Although some scholars use the term "cue" to refer specifically to unintended movements, we do not make this distinction.

Here, we integrate insights from work on visual pragmatic signals in natural conversation, focusing in particular on facial signals, which play a big role in signalling pragmatic information (Bavelas and Chovil 2018; Holler 2025), but are currently absent from classic communication games and agent-based models of language evolution. Our goal is to bridge the existing gap between these paradigms and the rich multimodal reality of human interaction, and to set the stage for new

1 Signals produced without a visible addressee may still be communicatively intended, and signals produced in the presence of an addressee may not necessarily have communicative intent.

lines of experimental and computational work that incorporate such pragmatic signals. Including pragmatic signals more seriously in language evolution research would not only make it more ecologically valid, but also advance our understanding of how these pragmatic signals (alongside referential signals) can support language emergence in the first place.

Human face-to-face interaction is inherently multimodal, integrating speech, gesture, facial signals, gaze, and many other bodily signals (Levinson and Holler 2014; Perniss 2018; Vigliocco et al. 2014). According to the dual track framework (Clark 1996), human interaction consists of referential signals in the primary track, which are signals that convey specific and denotable information about objects and events (e.g., words or signs like *table*, *happy*, *tomorrow*, *walking*), which have also been the focus of most language evolution research so far. Yet these referential signals are further layered with a second track, consisting of pragmatic signals that convey (un)certainty, understanding, trouble in understanding, or the intent to repair. Such "collateral" signals support making pragmatic inferences, and are often expressed not only through the auditory but also the visual modality using e.g., eyebrow raises, frowns, smiles, and manual gestures (Bavelas et al. 1992; Bavelas et al. 1995; Bavelas and Chovil 2018; Holler 2025; Kendon 2017; Micklos and Woensdregt 2023; Nölle et al. 2026; Swerts and Krahmer 2005). Importantly, these pragmatic signals are not inferior to the referential ones, seeing as they are central to managing the interaction, allowing interlocutors to make rapid inferences about what is already shared and what still needs clarification. In the context of language emergence, these signals are rather critical: when no shared language exists, interlocutors must rely heavily on such signals to negotiate meaning, identify misunderstandings and uncertainties, and adapt their communication strategies accordingly. By helping to establish and update common ground on the fly, these pragmatic signals may directly influence how novel communication systems (and referential signals) are established in the first place.

However, despite the centrality of pragmatic signals to everyday communication, and despite the fact that coordination and pragmatic inference are crucial for getting novel communication systems off the ground, the role of pragmatic signals remains relatively underexplored in language evolution research. Not only are many experiments not even conducted face-to-face (e.g., Fay et al. 2018; Kirby et al. 2015), even those that are conducted face-to-face tend to treat pragmatic behavior as uncontrolled noise rather than a variable of interest, or choose not to analyze these signals at all. As a result, we currently know little about what such signals look like in the context of language emergence, and how they might accelerate or hinder the conventionalization of new referential signals through interaction. The same gap is present in agent-based models, which have been instrumental in uncovering how interaction, learning biases, and population structure shape emerging languages.

However, to date there are no models that simulate truly multimodal interaction, in which agents can choose between, combine, or layer referential and pragmatic signals across communication channels.

We start by reviewing existing communication games that investigate language emergence, focusing on multimodality and pragmatics. We then discuss the role of pragmatic visual signals, particularly facial signals, in natural conversation. By highlighting why these signals might be of interest for language emergence, we propose directions for future research on the role of pragmatic visual signals. We end with identifying gaps in current agent-based computational models, followed by proposed future directions for models in which agents communicate through multiple channels.

2 Communication games

2.1 Multimodality in communication games

Communication games can be used to simulate the emergence of communication systems in a laboratory setting. Generally, these paradigms involve dyads who take turns describing referents using artificial language. The participant who receives the signal has to guess what referent is being described upon seeing the signal. Through what medium signals are produced varies: several studies make use of silent gestures as the mode of communication between hearing participants (e.g., Motamedi et al. 2019), or let participants vocalize non-linguistically (e.g., Perlman et al. 2015). Other modes of communication have also been explored, such as typing (e.g., Raviv et al. 2019), drawing (e.g., Fay et al. 2013, 2018), or using a color spectrum (e.g., Roberts et al. 2024). While some of these modalities are less naturalistic than others, they all reveal how communication systems arise under and adapt to the constraints of a medium. For example, drawings tend to become less iconic through repeated interaction (Fay et al. 2013, 2018), whereas this is not necessarily the case for vocalizations (Perlman et al. 2015). Together, previous studies suggest different communicative affordances of each modality depending on properties of the referent and context, and the flexibility to recruit different communication channels may have played a key role in the emergence of language. This is also consistent with recent findings in animal communication of songbirds that adapt to noisy environments by using an increasingly visual communication system (Framond et al. 2025).

Some studies have also directly compared affordances of different modalities for constructing referential signals. For example, Fay et al. (2013) and Fay et al. (2014) conducted communication games in which pairs of participants were asked to repeatedly describe emotions, actions, or objects in one of three conditions: using

gesture only, vocalization only, or vocalizations and gestures. Findings indicated that communicative success and efficiency were higher in gestures compared to vocalization, but found no advantage for the combined condition over the gesture condition. Similarly, Zlatev et al. (2017) found that guessing accuracy was higher when participants saw video clips of pantomime without vocalization, albeit participants in the multimodal condition, with pantomime combined with vocalization, reported that vocalizations were helpful. A similar follow-up task, where the multimodal clips were shown with audio only or video only, showed that the visual component of the multimodal stimuli alone yielded similar accuracy to the clips with both audio and video. Additionally, emotional valence was reliably detected from the auditory component. In line with these findings, Macuch Silva et al. (2020) let participants communicate about abstract visual stimuli or auditory stimuli with either gesture, vocalization, or both. For both types of stimuli, multimodal signals were more efficient, and this advantage was most pronounced for the auditory stimuli. This was attributed to being able to combine mimicking sounds using vocalization and the expressive power of gesture. Even more interestingly, when participants were sitting face-to-face in the gestural and multimodal conditions, repair was much more frequent compared to the vocalization condition where participants were sitting back-to-back, possibly due to the access to visual facial signals that invite repair.

2.2 Pragmatics in communication games

While all of the studies mentioned in the previous section can inform us about how different modalities can support novel communication in distinct ways, they often do not examine interactional mechanisms such as turn-taking, repair, and alignment, through which conventions are negotiated in natural interaction. The importance of interaction in coordinating on a novel communication system is recognized (Fay et al. 2010; Healey et al. 2007; Macuch Silva and Roberts 2016; Micklos et al. 2018; Tamariz et al. 2018), but little work has actually focused on unpacking pragmatic signals that are involved beyond mappings between linguistic form and the referent. Below, we outline the work that has been done on pragmatics and language emergence.

Healey et al. (2007) investigated the role of feedback in the emergence of graphical communication systems in a task where participants made drawings to refer to music they heard. They found that when participants could modify each other's drawings, which is proposed to resemble conversational repair, signals became more abstract. Similarly, Fay et al. (2010) found that concurrent partner feedback led to simpler drawings over time. Macuch Silva and Roberts (2016) investigated the interplay between noise, repair, and language structure in a task where participants typed sequences of characters to describe referents. When

messages were disrupted by noise, this led to more redundancy in the signals over time. However, when participants had the opportunity to repair, there was less redundancy and signals became more systematic. Tamariz et al. (2018) tested the emergence of iconicity in written communication, and found that iconicity emerged only in a communicative context but not in reproduction alone, indicating that there is something about the interactive aspect that drives the emergence of iconicity. Collectively, these studies mostly demonstrate how the presence or absence of certain interactional processes affects the signal space of emerging communication systems. The present work, however, emphasizes that interaction is also fundamental to the coordination process through which communication systems are built, which has largely been overlooked.

2.3 Pragmatic facial signals in communication

It becomes apparent that there is a gap: on the one hand, experiments about multimodality in language emergence often only focus on referential signals, while experiments about the role of pragmatic processes in language emergence often do not consider that communication is layered across several communication channels. In natural conversation, conversational turns usually also include pragmatic information that can be conveyed using multiple communication channels in addition to referential content (Bavelas et al. 1992; Bavelas et al. 1995; Bavelas and Chovil 2018; Holler 2025; Kendon 2017; Micklos and Woensdregt 2023; Nölle et al. 2026; Swerts and Krahmer 2005; Trujillo and Holler 2024). Rasenberg et al. (2022) zoom in on speech and gesture in other-initiated repairs in a dyadic referential communication task. They found more than half of the repair sequences contained both speech and gesture.[2] Apart from using speech and manual gesture for the purpose of repair, facial signals are a core device for conveying pragmatic information (Holler 2025). They receive far less attention because they have been traditionally researched in the context of emotion expression rather than communicative signalling (Crivelli and Fridlund 2018; Ekman 1993). Yet, facial signals are highly prevalent in coordinating interaction, such as prolonged blinking to signal understanding (Hömke et al. 2018) or frowning to indicate the need for clarification (Hömke et al. 2025). Importantly, there is evidence that pragmatic facial signals can drastically alter the interpretation of utterances in a non-additive and compositional manner (Trujillo and Holler 2024). For example, a spoken question accompanied by an eyebrow raise alone may be interpreted as an information request, but when the same eyebrow

2 Speech and gesture in this context were used for repair, but the signals themselves were referential in nature nonetheless.

raise is combined with a head tilt, the utterance may be interpreted as expressing skepticism. This highlights the importance of treating communication as a highly integrated multimodal system when investigating its emergence.

Specifically, Hömke et al. (2025) found that eyebrow furrows from an addressee signal misunderstanding: in a VR experiment with feedback given by a virtual human, speakers gave longer responses when avatars furrowed their eyebrows compared to when they nodded or did not respond. Even subtler movements, such as blink duration, can also be communicative: speakers gave shorter answers when their addressee produced longer blinks, perceived as signalling "I've received sufficient information for current purposes, please move on" (Hömke et al. 2018). Facial signals can also help anticipate conversational moves. In Emmendorfer and Holler (2025), participants judged the valence polarity of answers based on gaze direction and facial signals (e.g., brow raises, dimplers, and smiles for positive; frowns, nose wrinkles, and squints for negative), showing that such signals shape expectations about upcoming responses.

Beyond turn-by-turn interaction, facial signals can also convey epistemic stance. Swerts and Krahmer (2005) conducted a production followed by a perception study, where they found that low-certainty answers are often accompanied by raised eyebrows or "funny faces" in the production study. They used recordings from the production study in a perception study, where they found that participants can reliably distinguish high from low certainty from either visual, auditory, or both signals. Yet, different results were reported in a more recent perception study using a virtual human that answers questions (Nölle et al. 2026). There, widened eyes, closed eyes and raised brows were perceived as indicating higher certainty, while squints and frowns were perceived as indicating more doubt. The apparent difference in findings between what signals are produced to indicate uncertainty and what signals are perceived as uncertain could indicate a mismatch in communicative intention of the signal and how the signal is actually being perceived. This could be due to a number of factors, including the nature of questions and presented stimuli. Still, this potential mismatch is relevant when we consider these signals potentially playing a role in coordination between interlocutors.

2.4 Pragmatic facial signals in language emergence

The studies outlined above reveal that facial signals can offer rich information about understanding, level of certainty, and expected social valence of responses. In language emergence experiments, and communication games specifically, such signals

could support coordination: an expression of uncertainty on the receiver's face could prompt a producer to revise their utterance, while a confident expression could reinforce the creation of shared understanding. Still, the vast majority of existing work collapses these signals into a single channel or leaves them aside as uncontrolled noise, which is not a trivial decision, as visual pragmatic signals may contribute to the coordination process involved in forming a new communication system. Current experimental work mostly focuses on how different modalities can be leveraged to construct referential signals, but leave aside the potential of visual signals to convey pragmatic information.

Although scarce, a few studies have laid the groundwork to investigate visual pragmatic signals in the context of language emergence. Byun et al. (2018) studied emergent communication and repair in a more natural context, namely cross-signing, where two signers of different sign languages communicate. They annotated video recordings of free conversation between three pairs of signers of different sign languages at their first encounters. Among others, they annotated the data for repair sequences and try-markers: signals of uncertainty about whether the addressee will understand. Trouble in understanding is a continuous challenge in this context, and their analyses show that this is anticipated and often marked using eye contact and holding of the sign, potentially supporting communication by inviting repair sequences. They also found restricted repair initiations to be most prevalent in this context, probably due to the lack of shared code causing trouble in understanding individual signs. This is often marked by repetition of the sign and a puzzled facial expression. Reduced response times after repair initiations by the receiver suggest that these try-markers mobilize and facilitate other-initiated repairs. However, it has not been systematically investigated how these markers affect other behavioral outcomes during communication, such as communicative success. Additionally, Micklos (2015) includes an experiment where dyads communicate about nouns and verbs using silent gesture under two conditions: one where the director was allowed to do another attempt at gesturing the referent after communicative failure, and one where they were not. Still, both conditions allowed for interaction before the matcher made their guess, allowing for repair sequences. Careful investigation of the data revealed that repair sequences can promote the use of marking systems to distinguish nouns from verbs. They reported that trouble in understanding, which invites a repair sequence, was often signalled using furrowed eyebrows. Both these studies highlight the prevalence of visual pragmatic signals in repair sequences, especially facial expressions.

We propose to build upon these ideas, and investigate how pragmatic signals, which are often signalled visually, are linked to the coordination involved in the emergence of communication systems. The next section outlines ongoing work and provides further promising directions for future research.

2.5 Proposed future directions: analyzing visual signals in communication games

To investigate the role of visual pragmatic signals in language emergence, existing communication games can be taken as a starting point. Traditional analyses of compositionality, iconicity, convergence, communicative success etc. can be supplemented with the use of methods from multimodal communication research to analyze visual pragmatic signals, such as manual annotations or kinematic analysis of the video recordings. This would allow us to not only describe what these signals look like, but also to systematically investigate how they relate to the aforementioned measures in the context of language emergence.

Traditionally, much attention is given to analysis of the referential language that participants create, in terms of structure, iconicity, compositionality etc. Our proposal is to look beyond the referential content, but to the process of interaction as a whole, revealing how other aspects of face-to-face interaction, such as pragmatic facial signals, can contribute to the coordination process that is central to language emergence. To implement this approach, researchers can video-record future experiments to allow the annotation of bodily signals. Recordings can be manually annotated for facial signals using carefully developed annotation schemes, for example according to the Facial Action Coding System (FACS) (Ekman and Friesen 1978) or custom annotation schemes depending on the research aim. In addition, advances in kinematic technologies and algorithms allow for automated detection and annotation of movement patterns of interest to some extent (Akamine et al. 2026; Baltrušaitis et al. 2016; Esselink et al. 2023; Ghaleb et al. 2024; Pouw et al. 2026). Analysis of visual signals in video data can be time-consuming, especially if they require manual annotation of gestures or facial signals. At the same time, such analyses are already the standard in many studies on multimodal communication (e.g., Hömke et al. 2025; Nota et al. 2021; Rasenberg et al. 2022), which often have similar sample sizes as communication games in the field of language evolution. In the field of multimodal communication, annotations are often done manually, and in recent years also increasingly through automated approaches. Indeed, rapid advances in automated methods (including the detection and classification of visual signals, motion tracking, and the extraction of kinematic measures) have substantially reduced the time and effort required for video analysis. As these techniques continue to improve, we believe that integrating kinematic analyses into communication games is becoming an increasingly feasible approach. While this does require additional computational and analytical resources and knowledge, we believe that the potential gains justify this investment.

Linking this rich data containing both information on referential and pragmatic visual signals to behavioral outcomes would make it possible to ask a wide range of questions, for example:

- How does the presence of certain pragmatic facial signals affect convergence and communicative success between partners?
- Do certain facial or other bodily signals affect the extent to which a linguistic form is adapted?
- Which facial signals are used for what purpose, and how does that differ from natural conversation using established language?

Such analyses could reveal more about the mechanisms behind the negotiation of a shared communication system, not only by creating referential signals, but also through pragmatic signalling and the management of interaction.

Only very recently, ongoing research has started to systematically investigate the use of pragmatic facial signals in a communication game (Dona et al. 2026). It involves a communication game paradigm where participants take turns typing labels to refer to referents and guessing which referent belongs to a presented label, investigating the effects of facial visibility on a range of behavioral measures and analyzing the use of pragmatic facial signals. Snapshots of this experimental work are included below, showing that such signals are indeed being produced in these set-ups (see Figure 1).

The proposed approach raises a broader methodological standpoint about the current state of communication game paradigms. Many intentionally strip away multimodal resources to control experimental variables, but this also strips away mechanisms that are central to real-world language emergence. Incorporating and



Figure 1: Examples of facial signals produced during a communication game.

analysing these signals does not mean completely abandoning experimental control. Instead, it would move the study of language emergence closer to the conditions in which human communication systems naturally emerge, and would open the door to testing new hypotheses about the role of visual pragmatic signals in building novel communication systems.

Such experiments can provide valuable insights into how communication systems emerge under different conditions. However, they also have clear limitations: they typically involve small, short-lived communities, and participants who already use a language and thus possess both implicit and explicit linguistic knowledge. These constraints make it difficult to observe the true origins of language-like systems from scratch. This is where the added value of computational modeling becomes clear. In the following section, we propose how agent-based models could be leveraged to investigate the role of pragmatic signals from a multimodal perspective.

3 Computational simulations of language emergence

3.1 Agent-based models and swarm robotics

Agent-based models allow researchers to simulate populations of artificial agents who start with no pre-existing linguistic competence, and observe how shared communication systems can gradually emerge through repeated interactions over time (Kirby 2002; Steels 1995). These interactions are often structured as so-called language games, in which a sender agent produces a signal for a receiver agent, who in turn interprets the signal based on their own repertoire of signals (see Ohmer and Cuskley (2025) for a more detailed outline). Agent-based models offer a controlled and scalable way to investigate the fundamental mechanisms that might underlie language emergence, something that is difficult to achieve with human participants. In these models, researchers explicitly and carefully define the cognitive architectures of agents. We argue that this explicit control over cognitive architectures is not necessarily a weakness, but rather a strength when used to test competing hypotheses about mechanisms underlying emergent behavior. By implementing alternative cognitive architectures and comparing their behavioral outcomes to one another, as well as to empirical data from humans, these models can be used e.g., to explore competing hypotheses about the cognitive mechanisms underlying communication. In addition, the need for precise formalizations of cognitive architecctures also makes underlying assumptions explicit.

Agent-based modeling is a method that is widely used in the field of language evolution (Ohmer & Cuskley 2025). Early research inspired a wide range of subsequent agent-based models exploring aspects such as compositionality (e.g., Kirby et al. 2015), iconicity (e.g., Mudd et al. 2022), and the role of population structure (e.g., Josserand et al. 2024). Some have focused on aspects that are relevant to specific modalities, such as combinatoriality in phonological systems (de Boer and Zuidema 2010) and the population structure of emerging sign languages (Kirby and Tamariz 2021; Richie et al. 2014).

Taking agent-based models of sign language emergence as an example, Mudd et al. (2022) started from the observation that tight-knit communities often display more lexical variation in signs than more dispersed communities. They developed a model to test the idea that community structure, particularly the amount of shared cultural context and population size, influences lexical variability in sign languages. Their results showed that shared cultural background can sustain variation, as culturally dependent iconicity enables meaning inference without strict conventionalization. Similarly, Richie et al. (2014) compared homesign systems with broader community sign languages, finding that richer social networks in emerging sign languages promote conventionalization more than the sparser networks of homesign systems.

In a similar spirit, Josserand et al. (2024) investigated variation and population structure, focusing on the influence of social network structure by manipulating network parameters across a large number of simulations. They show that variation is more likely to emerge in populations made up of distinct communities, demonstrating how the ease with which information spreads shapes the emergence of linguistic conventions. This factor is related to, but distinct from, population size and density, which have traditionally been considered to be the main drivers of linguistic variation.

All these models represent language as a simple direct mapping between signals and meanings represented as one-dimensional vectors, focusing only on the referential aspect of communication. However, both spoken and signed communication is inherently multimodal, and utterances are constructed using multiple pieces of information that are distributed across different communication channels: individuals produce and process gestures or signs, facial expressions, and other visual signals to convey referential or pragmatic information. For example, the manual modality allows for multiple articulators to represent different aspects of meaning simultaneously, shaping the way meaning is conveyed (Slonimska et al. 2022). This adds complexity to how conventions and variation might emerge. What is needed to address questions related to these affordances is computational models with agents that can communicate through not one, but multiple channels that represent both referential and pragmatic aspects of the communicative signal.

A complementary line of research using swarm robotics extends these kinds of simulations into the embodied domain (Cambier et al. 2020; Cambier and Miletitch

2025), something that serves a lot of potential when we consider multimodality and the use of different communication channels. In robotic systems, researchers determine the available sensors, the processing mechanisms applied to sensory information, and the mapping from perception to behaviour, although the sensory input itself generally emerges from interactions with the environment. Swarm robotics experiments let large numbers of embodied robots tackle a task together, and can be used to study language emergence by studying emergent communication protocols between the robots in different contexts, such as path finding or foraging (Cambier et al. 2023; Cambier et al. 2024; Miletitch et al. 2022). Robots that are used in these simulations can have different sensors, picking up on e.g., auditory, visual, or proximity information. The possibility of manipulating the sensors and actuators through which robots interact holds considerable promise for future research on multimodality and the layering of different types of signals. This enables researchers to investigate how different physical communication channels interact, complement one another, or compete depending on the task and context, providing a richer, embodied perspective on how multimodal communication may emerge.

Taken together, agent-based models have yielded insights into how linguistic structure can emerge from the interplay of learning, communication, and social organization. These models have been extended to embodied, task-based contexts using swarm robotics paradigms. Yet, despite the inherently multimodal nature of real-world communication, most models treat signals as occurring in a single undifferentiated channel, representing referential content only. Addressing this gap by incorporating multiple channels that carry pragmatic information into artificial agents would open new possibilities for explaining their role in coordination, adaptation, and innovation of communication systems across diverse settings. Existing models, such as work by Kirby et al. (2015) or Steels (2016), provide a useful starting point. While originally applied to abstract, unimodal signals, these frameworks could be extended by integrating additional channels that represent pragmatic information coming from speech, gesture, and facial expressions, which can be incorporated into the learning and inference algorithms. Such an approach would allow agents to draw on multiple channels when negotiating meaning, mirroring the way real interlocutors coordinate and adapt in everyday interaction. In the next section, we propose promising avenues for this research.

3.2 Proposed future directions: multi-channel agent-based models

The second proposal concerns agent-based modeling frameworks. The work that has been undertaken so far, whether it is using Bayesian learners (e.g., Kirby et al. 2015),

neural network agents (e.g., Chaabouni et al. 2021), rule-based agents (e.g., Mudd et al. 2022; Steels 2016), or embodied agents (Miletitch et al. 2022), comes with a significant disadvantage. Namely, these models do not incorporate multimodality as it occurs in natural communication: agents communicate using a single-channel signal that is usually referential (Grifoni et al. 2021). To model communication more naturally and test hypotheses about pragmatic signals, we propose extending existing modelling frameworks to allow agents to produce and interpret combinations of signals from multiple channels. An example starting point could be the Bayesian framework of Kirby et al. (2015). This framework could be extended so that each agent can interpret, produce, and learn from multiple channels instead of one (e.g., a referential channel and a pragmatic channel), as illustrated in Figure 2.

For example, in each interaction, an agent might have two channels: one for representing the referential signal, and another to represent a pragmatic signal. The latter could be conceptualized as a facial signal of their certainty level that is based on their internal probabilistic state. The receiver would interpret both channels jointly, updating its internal model of the language and its model of the sender's

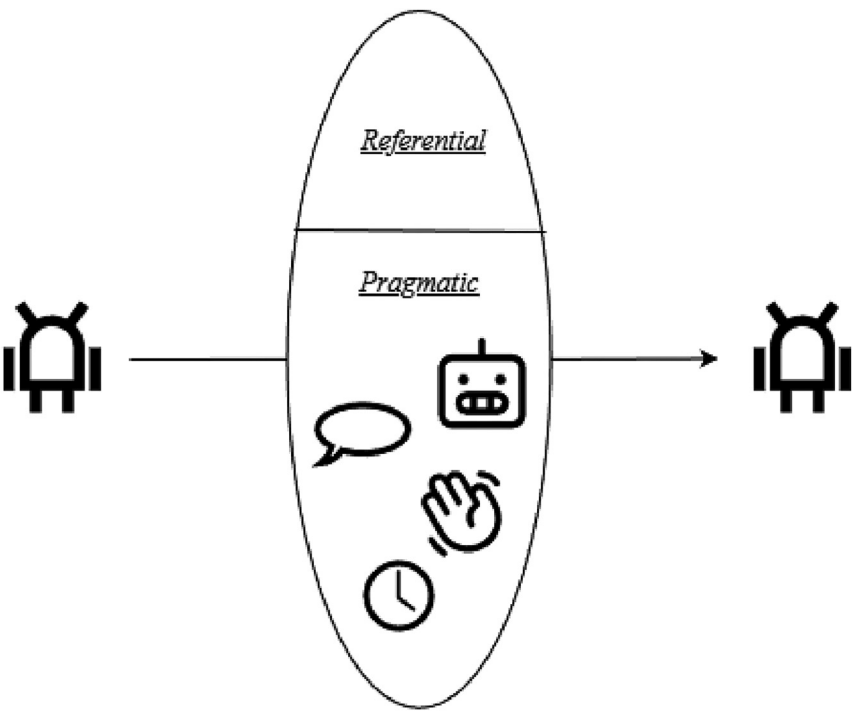

Figure 2: The proposed multi-channel agent-based modeling approach.

knowledge. Taking these more abstract models as a starting point, additional questions about the role of embodiment could be explored by extending these to the study of swarm robotics.

Such models open up possibilities for answering questions such as:

- How does access to extra pragmatic information on top of referential content affect convergence and communicative success under different theoretical frameworks of how these channels of information are integrated?
- How does noise in one channel (e.g., having to rely on the visual modality in a foggy environment) affect the use of other channels?
- How does access to pragmatic information affect structural characteristics of the emerging language?

There are, of course, challenges. Modeling interaction involving multiple communication channels that represent referential as well as pragmatic information requires defining how these channel may be independently or jointly interpreted, produced and how they influence each other, which requires making theoretical commitments. However, making these assumptions explicit also provides a means of formalizing and testing hypotheses about how different communicative channels interact.

By integrating multiple communication channels into agent-based models, we can simulate and experimentally manipulate the complex interplay between different modalities and between referential and pragmatic signals. This approach enables us to test different hypotheses about the role of multimodality and pragmatic communication in language evolution, as well as generate novel predictions to guide future laboratory research.

4 Conclusions

In this work, we highlighted a significant gap in current experimental and computational research on language emergence: the limited attention paid to pragmatic signals that are often expressed visually, especially facial signals, in both experimental paradigms and agent-based models. While existing studies have explored how novel communication systems arise through gestures, vocalization, and interaction, they predominantly focus on referential content and overlook the essential role of pragmatic signals in coordination and managing the interaction. We argue that integrating insights from natural face-to-face conversation, in which pragmatic signals such as facial signals can support coordination processes, can deepen our understanding of the dynamics underlying language emergence. To this end, we proposed new directions for research: multimodal analyses of communication

games that capture both linguistic and interactional processes, and multi-channel agent-based modeling approaches that simulate how agents use and learn from multiple, layered communication channels. Together, these offer a framework for advancing the study of how communication systems evolve in natural, interactive, and multimodal settings.

Acknowledgements: We would like to thank Marieke Woensdregt for contributing to conceptual discussions and Roman Miletitch for insightful discussions on the application of swarm robotics to language evolution. We also thank the reviewers for their constructive and insightful comments.
Data availability: Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

References

Akamine, S., M. Dingemanse & A. Özyürek. 2026. Validating dynamic time warping as a measure of gesture form similarity. *Behavior Research Methods* 58(4). 110.

Baltrušaitis, T., P. Robinson & L.-P. Morency. 2016. OpenFace: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)* 1–10. https://doi.org/10.1109/WACV.2016.7477553.

Bavelas, J. & N. Chovil. 2018. Some pragmatic functions of conversational facial gestures. *Gesture* 17(1). 98–127.

Bavelas, J., N. Chovil, L. Coates & L. Roe. 1995. Gestures specialized for dialogue. *Personality and Social Psychology Bulletin* 21(4). 394–405.

Bavelas, J., N. Chovil, D. A. Lawrie & A. Wade. 1992. Interactive gestures. *Discourse Processes* 15(4). 469–489.

Bavelas, J., J. Gerwing, C. Sutton & D. Prevost. 2008. Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language* 58(2). 495–520.

Ben Mocha, Y. & J. M. Burkart. 2021. Intentional communication: Solving methodological issues to assigning first-order intentional signalling. *Biological Reviews* 96(3). 903–921.

Byun, K.-S., C. de Vos, A. Bradford, U. Zeshan & S. C. Levinson. 2018. First encounters: Repair sequences in cross-signing. *Topics in Cognitive Science* 10(2). 314–334.

Cambier, N., A. E. Eiben & E. Ferrante. 2023. Emergent naming system in an unstructured environment: A shortest-path discovery case study. In *ALIFE 2023: Ghost in the Machine: Proceedings of the 2023 Artificial Life Conference* 47. https://doi.org/10.1162/isal_a_00644.

Cambier, N. & R. Miletitch. 2025. Task-driven language evolution with swarm robotics. In L. Raviv & C. Boeckx (eds.), *The Oxford handbook of approaches to language evolution*, 371–390. Oxford: Oxford University Press.

Cambier, N., R. Miletitch, A. Benítez-Burraco & L. Raviv. 2024. Pros and cons of prosociality across ecological niches: A swarm robotics approach. *Research Square Preprint*. https://doi.org/10.21203/rs.3.rs-3778159/v1.

Cambier, N., R. Miletitch, V. Frémont, M. Dorigo, E. Ferrante & V. Trianni. 2020. Language evolution in swarm robotics: A perspective. *Frontiers in Robotics and AI* 7. https://doi.org/10.3389/frobt.2020.00012.

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

Clark, H. H. 1996. *Using language*. Cambridge University Press.

Crivelli, C. & A. J. Fridlund. 2018. Facial displays are tools for social influence. *Trends in Cognitive Sciences* 22(5). 388–399.

de Boer, B. & W. Zuidema. 2010. Multi-agent simulations of the evolution of combinatorial phonology. *Adaptive Behavior* 18(2). 141–154.

Derchi, C. C., E. Mikulan, A. Mazza, S. Casarotto, A. Comanducci, M. Fecchio, J. Navarro, G. Devalle, M. Massimini & C. Sinigaglia. 2023. Distinguishing intentional from nonintentional actions through EEG and kinematic markers. *Scientific Reports* 13(1). 8496.

Dona, L., A. Özyürek, J. Holler & L. Raviv. 2026. Facial visibility and the emergence of communication systems: Insights from behavioral measures and facial signals. In *The Evolution of Language: Proceedings of the 16th International Conference (EVOLANG XVI)* 95–98.

Ekman, P. 1993. Facial expression and emotion. *American Psychologist* 48(4). 384–392.

Ekman, P. & W. V. Friesen. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica* 1(1). 49–98.

Ekman, P. & W. V. Friesen. 1978. Facial action coding system. https://doi.org/10.1037/t27734-000.

Emmendorfer, A. K. & J. Holler. 2025. Facial signals shape predictions about the nature of upcoming conversational responses. *Scientific Reports* 15(1). 1381.

Esselink, L., M. Oomen & F. Roelofsen. 2023. Exploring new methods for measuring, analyzing, and visualizing facial expressions. *FEAST. Formal and Experimental Advances in Sign Language Theory* 5. 35–48.

Fay, N., M. Arbib & S. Garrod. 2013. How to bootstrap a human communication system. *Cognitive Science* 37(7). 1356–1367.

Fay, N., S. Garrod, L. Roberts & N. Swoboda. 2010. The interactive evolution of human communication systems. *Cognitive Science* 34(3). 351–386.

Fay, N., C. J. Lister, T. M. Ellison & S. Goldin-Meadow. 2014. Creating a communication system from scratch: Gesture beats vocalization hands Down. *Frontiers in Psychology* 5. https://doi.org/10.3389/fpsyg.2014.00354.

Fay, N., B. Walker, N. Swoboda & S. Garrod. 2018. How to create shared symbols. *Cognitive Science* 42(S1). 241–269.

Framond, L. de, S. P. Sharp, K. Duclos, T. Beleyur & H. Brumm. 2025. Stream noise induces song plasticity and a shift to visual signals in a riverine songbird. *Current Biology* 35(18). 4570–4576.

Ghaleb, E., I. Burenko, M. Rasenberg, W. Pouw, P. Uhrig, J. Holler, I. Toni, A. Özyürek & R. Fernández. 2024. Co-Speech gesture detection through multi-phase sequence labeling. In *2024 IEEE/CVF winter conference on applications of computer vision (WACV)*, 3995–4003.

Grifoni, P., A. D'Ulizia & F. Ferri. 2021. When language evolution meets multimodality: Current status and challenges toward multimodal computational models. *IEEE Access* 9. 35196–35206.

Healey, P. G. T., N. Swoboda, I. Umata & J. King. 2007. Graphical language games: Interactional constraints on representational form. *Cognitive Science* 31(2). 285–309.

Holler, J. 2025. Facial clues to conversational intentions. *Trends in Cognitive Sciences* 29(8). 750–762.

Hömke, P., J. Holler & S. C. Levinson. 2018. Eye blinks are perceived as communicative signals in human face-to-face interaction. *PLoS One* 13(12). e0208030.

Hömke, P., S. C. Levinson, A. K. Emmendorfer & J. Holler. 2025. Eyebrow movements as signals of communicative problems in human face-to-face interaction. *Royal Society Open Science* 12(3). 241632.

Josserand, M., M. Allassonnière-Tang, F. Pellegrino, D. Dediu & B. de Boer. 2024. How network structure shapes languages: Disentangling the factors driving variation in communicative agents. *Cognitive Science* 48(4). e13439.

Kendon, A. 2017. Pragmatic functions of gestures: Some observations on the history of their study and their nature. *Gesture* 16(2). 157–175.

Kirby, S. 2002. Natural language from artificial life. *Artificial Life* 8(2). 185–215.

Kirby, S. & M. Tamariz. 2021. Cumulative cultural evolution, population structure and the origin of combinatoriality in human language. *Philosophical Transactions of the Royal Society B: Biological Sciences* 377(1843). 20200319.

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

Levinson, S. C. & J. Holler. 2014. The origin of human multi-modal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences* 369(1651). 20130302.

Macuch Silva, V., J. Holler, A. Özyürek & S. G. Roberts. 2020. Multimodality and the origin of a novel communication system in face-to-face interaction. *Royal Society Open Science* 7(1). 182056.

Macuch Silva, V. & S. G. Roberts. 2016. Language adapts to signal disruption in interaction. In *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG XI)*.

Micklos, A. 2015. *Emerging communication systems: Interaction for language evolution and transmission*. Los Angeles: University of California.

Micklos, A., V. Macuch Silva & N. Fay. 2018. The prevalence of repair in studies of language evolution. In *The Evolution of Language: Proceedings of the 12th International Conference (EVOLANG XII)* 316–318.

Micklos, A. & M. Woensdregt. 2023. Cognitive and interactive mechanisms for mutual understanding in conversation. In *Oxford Research Encyclopedia of Communication*. Oxford: Oxford University Press.

Miletitch, R., A. Reina, M. Dorigo & V. Trianni. 2022. Emergent naming conventions in a foraging robot swarm. *Swarm Intelligence* 16(3). 211–232.

Motamedi, Y., M. Schouwstra, K. Smith, J. Culbertson & S. Kirby. 2019. Evolving artificial sign languages in the lab: From improvised gesture to systematic sign. *Cognition* 192. 103964.

Mudd, K., C. de Vos & B. de Boer. 2022. Shared context facilitates lexical variation in sign language emergence. *Languages* 7(1). 31.

Müller, T. F. & L. Raviv. 2025. Communication experiments: Social interaction in the formation of novel communication systems. In L. Raviv & C. Boeckx (eds.), *The Oxford handbook of approaches to language evolution*, 41–62. Oxford: Oxford University Press.

Nölle, J., Y. Wu, P. Arias Sarah, O. G. B. Garrod, P. Schyns & R. Jack. 2026. Dynamic facial movements shape how confident speakers appear. *PsyArXiv Preprint*. https://doi.org/10.31234/osf.io/8hqy9_v1.

Nota, N., J. P. Trujillo & J. Holler. 2021. Facial signals and social actions in multimodal face-to-face interaction. *Brain Sciences* 11(8). 8.

Ohmer, X. & C. Cuskley. 2025. Communication games: Modelling language evolution through dyadic agent interactions. In L. Raviv & C. Boeckx (eds.), *The Oxford handbook of approaches to language evolution*, 301–318. Oxford: Oxford University Press.

Perlman, M., R. Dale & G. Lupyan. 2015. Iconicity can ground the creation of vocal symbols. *Royal Society Open Science* 2(8). 150152.

Perniss, P. 2018. Why we should study multimodal language. *Frontiers in Psychology* 9. https://doi.org/10.3389/fpsyg.2018.01109.

Perusquía-Hernández, M., S. Ayabe-Kanamura & K. Suzuki. 2019. Human perception and biosignal-based identification of posed and spontaneous smiles. *PLoS One* 14(12). e0226328.

Pouw, W., S. A. Shaikh, J. Trujillo, B. Yung, A. Rueda-Toicen, G. de Melo & B. Owoyele. 2026. EnvisionHGdetector: A computational framework for Co-Speech gesture detection, kinematic analysis, and interactive visualization. *PsyArXiv Preprint*. https://doi.org/10.31234/osf.io/psg5f_v2.

Rasenberg, M., W. Pouw, A. Özyürek & M. Dingemanse. 2022. The multimodal nature of communicative efficiency in social interaction. *Scientific Reports* 12(1). 1.

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

Richie, R., C. Yang & M. Coppola. 2014. Modeling the emergence of lexicons in homesign systems. *Topics in Cognitive Science* 6(1). 183–195.

Roberts, G., J. Kuang & R. Clark. 2024. Communicative factors in the emergence of phonological dispersion. In *Proceedings of the Annual Meeting of the Cognitive Science Society* 46.

Slonimska, A., A. Özyürek & E. Campisi. 2026. Adults mark the communicative relevance of their gestures more for children than for other adults. *Discourse Processes* 63(3). 185–209.

Slonimska, A., A. Özyürek & O. Capirci. 2022. Simultaneity as an emergent property of efficient communication in language: A comparison of silent gesture and sign language. *Cognitive Science* 46(5). https://doi.org/10.1111/cogs.13133.

Steels, L. 1995. A self-organizing spatial vocabulary. *Artificial Life* 2(3). 319–332.

Steels, L. 2011. Modeling the cultural evolution of language. *Physics of Life Reviews* 8(4). 339–356.

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

Swerts, M. & E. Krahmer. 2005. Audiovisual prosody and feeling of knowing. *Journal of Memory and Language* 53(1). 81–94.

Tamariz, M., S. G. Roberts, J. I. Martínez & J. Santiago. 2018. The interactive origin of iconicity. *Cognitive Science* 42(1). 334–349.

Trujillo, J. P. & J. Holler. 2024. Conversational facial signals combine into compositional meanings that change the interpretation of speaker intentions. *Scientific Reports* 14(1). 2286.

Vigliocco, G., P. Perniss & D. Vinson. 2014. Language as a multimodal phenomenon: Implications for language learning, processing and evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences* 369(1651). 20130292.

Zlatev, J., S. Wacewicz, P. Zywiczynski & J. Van De Weijer. 2017. Multimodal-first or pantomime-first? Communicating events through pantomime with and without vocalization. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems* 18(3). 465–488.