



# Emergent language: a survey and taxonomy

Jannik Peters<sup>1</sup> · Constantin Waubert de Puiseau<sup>1</sup> · Hasan Tercan<sup>1</sup> ·  
Arya Gopikrishnan<sup>2</sup> · Gustavo Adolpho Lucas de Carvalho<sup>3</sup> ·  
Christian Bitter<sup>1</sup> · Tobias Meisen<sup>1</sup>

Accepted: 26 January 2025 / Published online: 7 March 2025  
© The Author(s) 2025

## Abstract

The field of emergent language represents a novel area of research within the domain of artificial intelligence, particularly within the context of multi-agent reinforcement learning. Although the concept of studying language emergence is not new, early approaches were primarily concerned with explaining human language formation, with little consideration given to its potential utility for artificial agents. In contrast, studies based on reinforcement learning aim to develop communicative capabilities in agents that are comparable to or even superior to human language. Thus, they extend beyond the learned statistical representations that are common in natural language processing research. This gives rise to a number of fundamental questions, from the prerequisites for language emergence to the criteria for measuring its success. This paper addresses these questions by providing a comprehensive review of relevant scientific publications on emergent language in artificial intelligence. Its objective is to serve as a reference for researchers interested in or proficient in the field. Consequently, the main contributions are the definition and overview of the prevailing terminology, the analysis of existing evaluation methods and metrics, and the description of the identified research gaps.

**Keywords** Emergent language · Emergent communication · Artificial intelligence · Reinforcement learning · Multi-agent

## 1 Introduction

Communication between individual entities is based on conventions and rules that emerge from the necessity or advantage of coordination. Accordingly, Lewis [1] formalized settings that facilitate the emergence of language as “coordination problems” [1] and introduced a simple signaling game. This game, in which a speaker describes an object and a listener confronted with multiple options has to identify the indicated one, extensively shaped the field of emergent language (EL) research in computer science. Early works examined narrowly defined questions regarding the characteristics of emergent communication (EC)

---

Arya Gopikrishnan and Gustavo Adolpho Lucas De Carvalho have work done during and after a DAAD RISE internship at Institute of Technologies and Management of Digital Transformation.

---

Extended author information available on the last page of the article

via hand-crafted simulations [2–12]. These approaches mostly utilized supervised learning methods and non-situated settings, limiting them in their ability to examine the origins and development of complex linguistic features [2]. However, EL research experienced an upsurge in the period between 2016 and 2018 [13–20] with a focus on MARL approaches [21–32] to enable the examination of more complex features.

One fundamental goal of EL research from the multi-agent reinforcement learning (MARL) perspective is to have agents autonomously develop a communication form that allows not only agent-to-agent but also agent-to-human communication in natural language (NL) style fashion [2, 16, 24, 29, 33, 34]. Therefore, reinforcement learning (RL) methods are attractive from two points of view. First, successful communication settings might lead to agents that are “more flexible and useful in everyday life” [35]. Furthermore, they may provide insights into the evolution of NL itself [36]. However, encouraging communication alone will not automatically produce a language with natural language characteristics [37]. Providing the right incentives for language development is therefore crucial.

EL is the methodological attempt to enable agents to not only statistically understand and use NL, like natural language processing (NLP) models that learn on text alone [38, 39], but rather to design, acquire, develop, and learn their own language [40, 41]. The autonomy and independent active experience of RL learning settings is a crucial difference to the data-driven approaches in the field of NLP [42–44] and its large language model (LLM). According to Browning and LeCun, “we should not confuse the shallow understanding LLM possess for the deep understanding humans acquire” [41] through their experiences in life. In EL settings, the agents experience the benefits of communication through goal-oriented tasks [45] just like it happens naturally [1] and therefore have the opportunity to develop a deeper understanding of the world [33, 46]. Hence, advances in EL research enable novel applications of multi-agent systems and a considerably advanced form of human-centric AI [35].

In the current state of EL research, numerous different methods and metrics are already established but they are complex to structure and important issues remain regarding the analysis and comparison of achieved results [29, 35, 47]. Therefore, we see a need for a taxonomy to prevent misunderstandings and incorrect use of established metrics. In this paper, we address these issues by providing a comprehensive overview of publications in EL research and by introducing a taxonomy for discrete EL that encompasses key concepts and terminologies of this field. Additionally, we present established and recent metrics for discrete EL categorized according to the taxonomy and discuss their utility. Our goal is to provide a clear and concise description that researchers can use as a shared resource for guidance. Finally, we create a summary of EL research that highlights its achievements and provides an outlook on future research directions. We base our work on a comprehensive and systematic literature search with reproducible search terms on well-known databases. We follow the PRISMA [48] specifications and show a corresponding flow diagram in Fig. 11 in Appendix B. The literature search and review process as well as its results are described in detail in Sect. 4. All identified work has been reviewed and categorized according to an extensive list of specific characteristics, e.g. regarding communication setting, game composition, environment configuration, language design, language metrics, and more.

Previous surveys of EL in computer science focused only on a subgroup of characteristics or specific parts of this research area. Some of these earlier surveys focus on specific learning settings [45, 49, 50], on methodological summaries and criticism [29, 40, 51–55], or provide a more general overview [24, 35, 36, 47, 56–58]. The most similar ones to our work are [35] and [58]. Lazaridou [35] gives an introduction and overview of the EL field

before 2021, however, it is mostly a summary of previous work and does not provide a taxonomy or review of existing metrics in the field as we do. Brandizzi [58] focuses on common characteristics in EC research and the development of emergent human–machine communication strategies. They discuss distinctions and connections of EC research to linguistics, cognitive science, computer science, and sociology, while we focus on emergent language and its analysis. We describe and discuss all relevant surveys in more detail in Sect. 3.

Based on this preliminary work, the current state of research on EL misses an overarching review and a comprehensive compilation and alignment of proposed quantification and comparability methods. Accordingly, the key contributions of the present survey are:

- A taxonomy of the EL field, in particular regarding the properties of discrete EL, see Sect. 5.
- A list of categorized quantification approaches and metrics in a consistent notation, see Sect. 6.
- A summary of open questions and an outlook on potential future work, see Sect. 7.

In addition, we introduce the fundamental concepts of NL and EC that underlie our survey in Sect. 2. As mentioned, we provide a detailed summary of related surveys in Sect. 3. Section 4 describes our study methodology, including the keywords and terms of our systematic literature search. Finally, Sect. 8 offers a concluding discussion and final remarks.

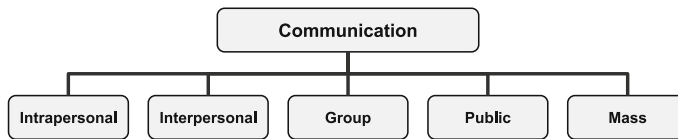
## 2 Background

To contextualize the presented taxonomy and analysis, this section summarizes the key concepts of communication and linguistics and provides an overview of EL research.

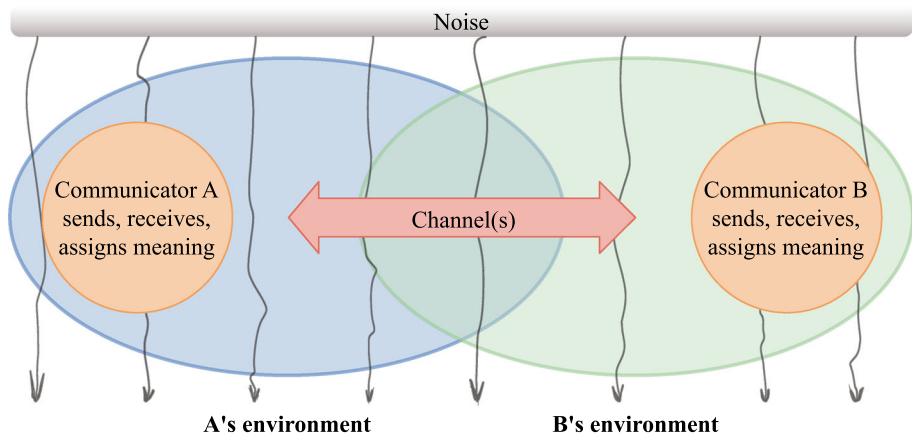
### 2.1 Communication

Communication at its very basis is the transfer or exchange of signals, which can be interpreted to form some information. These signals include both intended, such as deliberate utterances, and unintended, such as uncontrolled bodily reactions, and include both explicit and implicit parts [59]. According to Watzlawik’s “Interactional View” [60], “one cannot not communicate”. In this regard, communication is ubiquitous and necessary, occurring through various channels and modes [41, 61–64]. Depending on the specific channel and purpose, communication can be roughly divided into the five forms depicted in Fig. 1.

In the context of EL, two of these forms are actively studied, namely interpersonal communication and group communication. Interpersonal communication is communication between entities that mutually influence each other, and its general setting is depicted in Fig. 2. This form of communication is based on individual entities, each within its perceivable environment. Although these environments are agent-specific, they overlap and allow communication through a common channel. In addition, there may be noise in this process that affects the perception of the environment or the communication itself. Group communication, on the other hand, differs only in the number of entities involved and the communication goal. Usually, group communication is more formal and focuses on a common goal or group task while interpersonal communication has a social character and



**Fig. 1** The different forms of communication. They are divided by type of recipients and purpose. *Intrapersonal communication* encompasses self-centered communication like internal vocalization. The remaining forms of communication are directed externally and are utilized to transmit information to individuals, in the *interpersonal setting*, or groups of addressees. In *group communication* the participants usually have a common goal, whereas *public communication* focuses on the general transfer of information to a group of interested but not necessarily goal-aligned entities. Finally, *mass communication* is used to describe any form of communication that is directed towards a general audience and focuses availability, for example, through the use of various media, including the internet. Adapted from [65]



**Fig. 2** Interpersonal communication. Actors are communicator A and B, each depicted by orange circles. They are each situated in their individual environment, depicted by the blue and green ellipses. At the overlap point the red arrow indicates the available communication channel. The potential environmental noise, influencing the communication, is represented by grey arrows going through the entire image. Adapted from [76] (Color figure online)

might only relate to a goal or task of one of the participants. Accordingly, the group communication setting can be found in most population-based EL research. Intrapersonal communication (e.g., internal vocalization), public communication (e.g., lectures), and mass communication (e.g., blog entries) are not currently examined in the EL literature.

Communication has been studied in many different disciplines from many different perspectives, including animals [66–68], pre-linguistic infants [69, 70], and sign language [71]. However, in order to keep the present work concise, we will refer mainly to research in the fields of linguistics and computer science. Generally, communication can be seen as a utility to coordinate with others [72–75]. Conversely, the necessity for collaboration within a collective may be a fundamental precursor to the evolution and sustained functionality of explicit communication [4, 53]. This theory leads to an essential differentiation regarding context-dependent communication. Meaningful communication might emerge in a cooperative but not in a fully competitive or manipulative setting. However, a partially competitive setting might be vital for the emergence of resilient and comprehensive

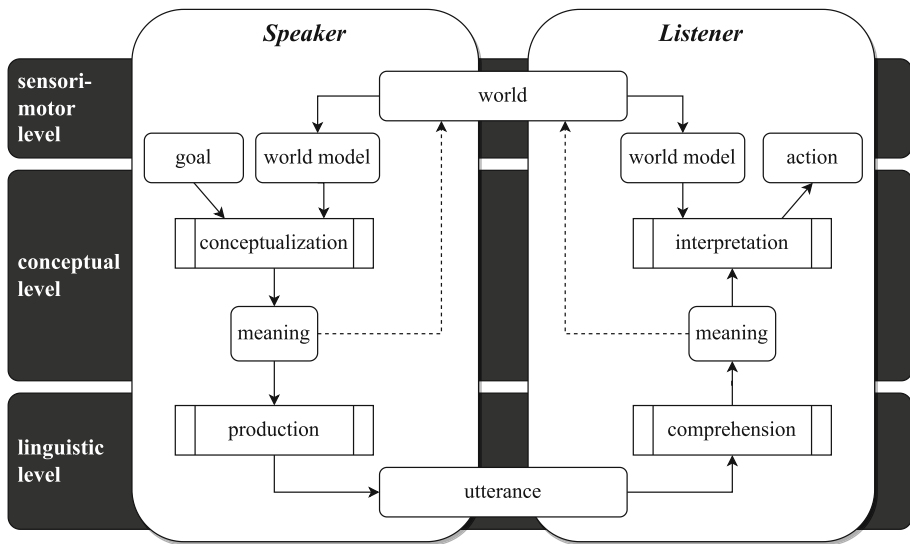
communication, e.g. to enable the detection and use of lies [34]. Accordingly, the level of cooperation is a defining element of the communication setting in EL research.

NL is a tool that allows us to encode very complex information within a discrete and humanly manageable amount of utterances. A lot of artificial intelligence research aims to develop NL models, with applications ranging from translation to coherent full-text generation based on single-word input [44, 77–79]. However, current research is mostly based on LLM which “achieve a sophisticated level of inductive learning and inference” [80] but are also “far from human abilities in natural language inference, analogical reasoning, and interpretation” [80]. A lot of research from the EL community is based on the theory, that models which learn language statistically based on static datasets are limited in their communicative and cognitive abilities [18, 41, 72, 81, 82]. Recent publications have shown that LLM have “weak reasoning and decision-making abilities” [83], their “reasoning is fragile” [84], and that current LLM face “reliability issues” [85]. Correspondingly, the field of EL research in AI aims to enable agents to utilize intended communication in the same way humans use it to increase cooperation, performance, and generalization and, in the long run, enable direct meaningful communication between humans and artificial systems [18, 34]. In line with this, multiple explicit forms of EC in artificial intelligence research have been investigated as shown in Sect. 5.4.2. In contrast, work focusing on implicit communication, like the information content of spatial positioning of agents in a multi-agent setting [86], is not part of the present survey.

## 2.2 Natural language

NL is a prime example of a versatile and comprehensive form of communication designed to convey meaning [87]. The flexibility of NL allows humans to be exact but also deliberately ambiguous in their communication [88]. It is a vital feature that distinguishes us from other species and gives us a great advantage in terms of knowledge storage, sharing, and acquisition [88]. However, the origin and evolution of language is still a mystery [89, 90]. In the field of linguistics, many conflicting theories have been introduced so far [90–94], ranging from behavioral to biological explanations. Additionally, accompanying research in the field of computer science has a long history [31] with a comparable range of theories. Even though there is still a debate around this topic, it is commonly agreed upon that a very intricate evolutionary process was involved [88, 90]. This evolution most likely took place in two different areas simultaneously, biologically and linguistically. On the biological side, the human brain most likely developed specific areas and functionalities specifically for more complex language-based communication, that are studied in the scientific field of neurolinguistics [95]. On the linguistics side, this evolution can be seen in language development itself, which is a constantly ongoing process [88] that might be strongly connected to the development of cognitive skills [96] and the social environment [97]. Similarly, EL is concerned with the research of suitable model structures for the processing of language, while concurrently developing and evaluating language.

While the exact origin of language is highly debatable, the actual communication process via NL is generally easier to conceptualize. For example, it can be modeled by the semiotic cycle depicted in Fig. 3 [45, 98]. This depiction applies to multiple expressive channels, e.g. speech and writing. It assumes at least two involved parties, a speaker and a listener. The speaker produces an utterance based on the meaning to be conveyed. This meaning results from the combined conceptualization of the speaker’s goal and model of the world. On the other hand, the listener receives the utterance and comprehends it to derive a meaning,



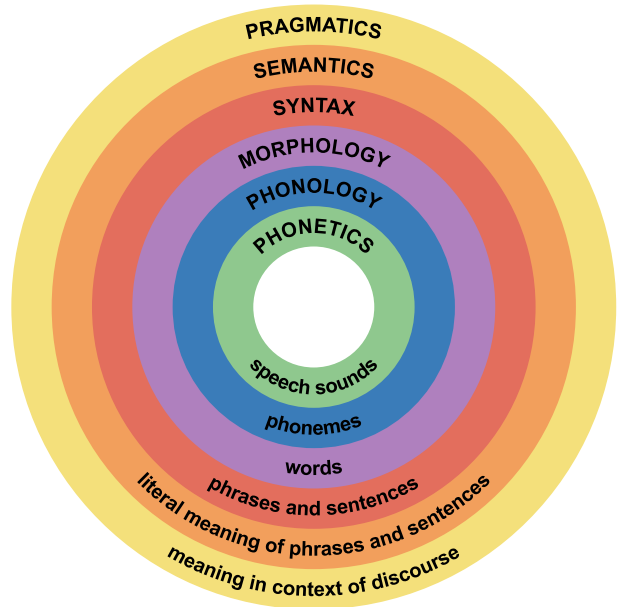
**Fig. 3** The semiotic cycle. This framework of a language-based exchange between two separate entities, called speaker and listener, categorizes the process into three levels. The *sensori-motor level* encompasses sensor and world-oriented components, the *conceptual level* includes internal and intangible parts like the individual world model and conceptualization capabilities, and the *linguistic level* consists of the production and comprehension of the linguistic exchange, which is the externalized connection between speaker and listener. Adapted from [45, 98]

which is not a direct copy of the initial one by the speaker but it still refers to the shared world. The interpretation of the meaning, which the listener's world model informs, leads to some action by the listener. At the center of this process are the shared world and the respective world models of speaker and listener that function as grounding for the information exchange via language. Further, both linguistic level components of production and comprehension allow the respective agent to participate in the language process.

The semiotic cycle puts the utterance as an externalized information carrier into focus. While the other components are internalized and thus difficult to define and measure, the utterance itself is external and available for analysis. Fundamentally, this specific utterance is based on the underlying communication process and specifically, the language used. Accordingly, most research papers investigate characteristics of the utilized language to analyze the communication possibilities and capabilities of users. To this end, linguistics subdivides the language structure into six major levels [99–101], as illustrated in Fig. 4. This structure was originally developed for spoken language, as indicated by the terms 'phonetics' and 'phonology' derived from the Greek word 'phon' meaning 'sound'. However, the levels are also applicable to written language in the context of EL. Therefore, the following description will address both spoken and written language within this framework.

The *phonetics* level includes the entire bandwidth of the chosen, often continuous, language channel. For example, it comprises the full range of possible speech sounds available to humans. Consequently, it is fundamental for the general transfer range and describes it without any limitation. At the *phonology* level are the atomic building blocks of the spoken or written language, defined as phonemes or graphemes. A phoneme or

**Fig. 4** Major levels of linguistic structure. This conceptual structure depicts the elements of a language as concentric rings that build upon each other. From the inner circle, *phonetics*, through *phonology*, *morphology*, *syntax*, and *semantics*, to *pragmatics*. Adapted from [99]



grapheme enables the creation of meaning as well as the necessary distinction at the lowest level of language. However, in a NL with an alphabetic writing system, phonemes and graphemes, which in this case correspond to letters, are often not a direct match and are only roughly related. Nevertheless, these individual units comprise the set of used elements from the continuous channel range for a specific language. These are used and combined at the *morphology* level to create and assign meaning by making words, in linguistics called lexemes. In this context, word-forming rules and underlying structures are of interest. Utilizing these meaningful building blocks, sentences can be realized at the *syntax* level. This level only concerns the structure of sentences and in particular, their assembly rules and the word categories that are used. The meaning of these sentences is relevant at the next level, *semantics*. At this level, the literal meaning of language constructions is of interest while the final level, *pragmatics*, focuses on how context contributes to the meaning. Accordingly, it analyzes how language is used in interactions and the relationship between the involved parties. Overall, the presented levels are not only important to describe language functionally and structurally but also to distinguish language characteristics and metrics. Thus, we use them to organize parts of the taxonomy in Sect. 5 and the metrics in Sect. 6.

### 2.3 Emergent language

EL refers to a form of communication that develops among artificial agents through interaction, without being explicitly pre-programmed. Thus, it is a bottom-up approach, arising from the agents' need to cooperate and solve tasks within a given environment [58]. This process involves the agents creating, adapting, and refining linguistic structures and meanings to enhance their ability to exchange information effectively and efficiently [49].



EL research aims to understand the principles and mechanisms underlying this spontaneous development of communication. It explores how linguistic elements such as syntax [102], semantics [103, 104], and pragmatics [29] can arise from the interaction of artificial agents and how these elements contribute to the agents' performance and cooperation.

A NL-like communication form would make artificial agents and computer systems, in general, more accessible, simpler to comprehend, and altogether more powerful [24, 34, 35]. EL research originally focused on the question of language origin [3]. Recently, this focus shifted to the more functional aspect of EL, focusing on how to enable agent systems to benefit from a mechanism that helped humanity thrive and how to achieve communication capabilities as close as possible to NL [35]. Today, EL within computer science is about self-learned [4], reusable [46], teachable [105, 106], interpretable [14], and powerful [18] communication protocols. In the long run, EL aims to enable machines to communicate with each other and with humans in a more seamless and extendable manner [35, 107].

Accordingly, various research questions and areas were derived. For example, recent papers have addressed issues around the nature of the setting, which can be semi-cooperative [34, 108], include adversaries [22, 32], have message-influencing noise [109], or incorporate social structures [110, 111]. Moreover, some are concerned with the challenge of grounding EL, e.g. using representation learning as basis [112], combining supervised learning and self-play [113], or utilizing EL agents as the basis for NL finetuning approaches [107]. Others tackle the direct emergence of language with NL characteristics, e.g. looking at internal and external pressures [114–118], evaluating factors to enforce semantic conveyance [53], looking at compositionality [119], generalization [25], or expressivity [120], or questioning the importance of characteristics like compositionality [121] and the connection between compositionality and generalization [122].

Based on these examples and the introduced goals and approaches, the difference in comparison to NLP research becomes apparent. Current approaches in NLP, namely LLM, learn language imitation via statistics, but they might not capture the functional aspects and the purpose of communication itself [18, 41]. In contrast, EL uses language not as the sole objective but as a means to achieve something with meaning [23]. Accordingly, agents have to learn their own EL to enable functionality beyond simple statistical reproduction. Specifically, agents should learn communication by necessity or benefits [114] and they need a setting that rewards or encourages communication, e.g., an at least partially cooperative setting [34].

While the EL concept sounds simple, it comes with many challenges. Encouraging communication alone can lead to simple gibberish that helps with task completion but does not represent the intended natural language characteristics [37]. Providing the right incentives for language development is therefore crucial. In addition, it is important to examine how agents use communication and the opportunity to send and receive information, raising the question of how to measure successful communication [29]. The measurability of language properties such as syntax, semantics, and pragmatics is also important for assessing the emergence of desirable language properties [122]. The following sections explore these challenges and related constructs and approaches in detail.

### 3 Related surveys

As briefly mentioned in Sect. 1, our literature review identified 19 publications that we classified as surveys. We adopted a broad definition of what constitutes a survey, categorizing any publication as a survey if it either explicitly described itself as such or provided a



particularly comprehensive and structured review of previous research. These publications conduct similar investigations on EL research but with different scopes. We focus on discrete language emergence, associated taxonomy, characteristics, metrics, and research gaps. In contrast, in our review of the existing survey work, three distinct interpretive directions emerge, which we categorize as summarized in Table 1: Surveys that focus on the learning *settings* [45, 49, 50], surveys that summarize and review utilized *methods* [29, 40, 51–55], and surveys that provide a *general* discussion or overview of the EL field [24, 35, 36, 47, 56–58]. The following section briefly summarizes these surveys within these categories.

### 3.1 Settings

Surveys within the *settings* category primarily focus on the design of language learning environments and the general structure of learning problems. For instance, van Eecke and Beuls [45] explored the language game paradigm, categorizing experiments and identifying properties critical for MARL research, such as symmetric agent roles and autonomous behavior. While their work offers a foundational perspective, our survey extends beyond the language game paradigm to analyze a broader range of approaches in greater depth (see Sect. 7). Similarly, Lipowska and Lipowski [49] emphasized sociocultural aspects, such as migration and teachability, within simple naming games. While these are part of our analysis, our review situates them within a unified framework, providing a more comprehensive perspective. Denamganai and Walker [50] introduced ReferentialGym as a tool for studying referential games and their associated metrics, like positive signaling and positive listening [29]. In contrast, our survey goes beyond referential games, offering a broader exploration of EL metrics and their applications across diverse frameworks.

### 3.2 Methods

The *methods* category encompasses surveys that primarily examine learning and evaluation methodologies in EL, each offering unique perspectives on key challenges. A recurring theme in this category is the need for more comprehensive evaluation metrics that capture the complexity of emergent communication [51–53]. For example, Korbak et al. [51] highlighted the limitations of existing compositionality metrics, introducing the *tree reconstruction error* to address semantic compositionality, a challenge we contextualize further in Section 5.4.5. In contrast, LaCroix [52] critiqued the overemphasis on compositionality, advocating for a shift towards reflexivity, though metrics for this remain unexplored. This gap underscores the fragmented nature of current evaluation practices.

**Table 1** Previously published surveys on EL, organized according to their primary focus

Settings	van Eecke and Beuls [45], Lipowska and Lipowski [49], Denamganai and Walker [50]
Methods	Korbak et al. [51], LaCroix [52], Lemon [40], Lowe et al. [29], Mihai and Hare [53], Galke and Raviv [54], Vanneste et al. [55]
General	Hernandez-Leal et al. [47], Brandizzi and Iocchi [24], Moulin-Frier and Oudeyer [56], Galke et al. [36], Fernando et al. [57], Suglia et al. [123], Zhu et al. [124], Lazaridou and Baroni [35], Brandizzi [58]

Grounding and utility also feature prominently in this category. Lemon [40] emphasized the interplay of symbolic and conversational grounding, highlighting data-related challenges, while Lowe et al. [29] proposed pragmatic metrics such as positive signaling and positive listening, which inspired our taxonomy of language utility in Sect. 5.4.6. These works collectively underscore the necessity of balancing semantic depth with practical utility, a balance our survey seeks to achieve by integrating diverse perspectives into a unified framework.

Further, recent studies like those by Galke and Raviv [54] explored the role of linguistic pressures and biases in bridging the gap between EL and human NL, providing insights into the origins of NL phenomena in EL. Vanneste et al. [55] tackled discretization methods critical for EL learning, offering a comparative analysis that complements our work.

### 3.3 General

The *general* category includes surveys that provide broad overviews or address themes not confined to specific *settings* or *methods*. Key contributions in this category highlight the interdisciplinary perspectives, interaction paradigms, and structural dimensions of emergent communication research.

Hernandez-Leal et al. [47] offered a foundational overview of multi-agent deep reinforcement learning (MARL), including emergent behavior and communication. While their survey provides valuable historical context and practical challenges, our work builds upon this by focusing specifically on emergent language (EL) within MARL, analyzing it with finer granularity and from a metrics-driven perspective. Similarly, Brandizzi and Iocchi [24] emphasized the underrepresentation of human-in-the-loop paradigms, proposing novel interaction settings but lacking the systematic categorization and metrics-oriented discussion presented in our survey.

Moulin-Frier and Oudeyer [56], Fernando et al. [57], and Galke et al. [36] explored interdisciplinary connections and cognitive constraints in EL, with the latter focusing on the perceived gaps between EL and NL. While these works underscore key challenges, our survey contextualizes such limitations across a broader set of metrics and emergent properties, providing actionable insights for bridging these gaps.

Other surveys, such as those by Suglia et al. [123] and Zhu et al. [124], structured EL research into multimodal and dimensional frameworks, respectively. These works serve as useful complements to our survey, which introduces an extensive taxonomy (Sect. 5) that synthesizes and organizes findings from diverse sources. Similarly, Lazaridou and Baroni [35] and Brandizzi [58] provided comprehensive overviews of the field but lacked the detailed quantification and taxonomy of metrics that form the core of our work.

By synthesizing these contributions, our survey is distinguished by its focus on emergent language metrics and quantification, complemented by a systematic taxonomy to address fragmentation in the field. This integrated approach provides a structured roadmap for advancing EL research, with an emphasis on both practical measurability and interdisciplinary relevance.

## 4 Study methodology

The literature search that resulted in the body of work surveyed in this paper was conducted on the 17<sup>th</sup> of June 2024. The used libraries and databases are: [ScienceDirect](#), [IEEE Xplore](#), [ACM Digital Library](#), [WebOfScience](#), [arXiv](#), and [SemanticScholar](#). [SemanticScholar](#) is a

**Table 2** Literature databases and search queries used for the present survey and the number of results obtained for each

Source	Query	Results
ScienceDirect	“emergent language” Subject areas: Computer Science	5
IEEE Xplore	“emergent language”	9
ACM Digital Library	All: “emergent communication”	60
	All: “emergent language”	19
WebOfScience	TS=(“emerg* communication” or “emerg* language”) NOT TS=emergency NOT TS=5G NOT TS=wireless Refined by: WEB OF SCIENCE CATEGORIES: ( COMPUTER SCIENCE ARTIFICIAL INTELLIGENCE ) Timespan: All years	16
arXiv	all=“emergent communication”	207
	all=“emergent language”	66
	all=Emergent multi-agent communication	208
SemanticScholar	multi agent emergent language Filtered by topic Computer Science	23
References	–	23

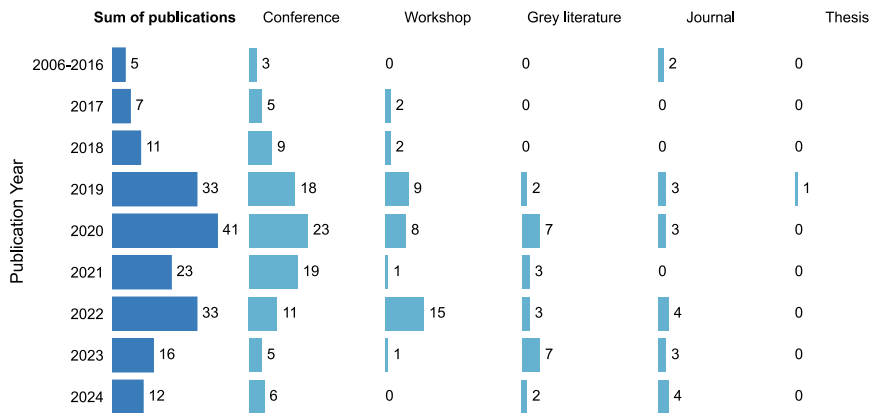
special case, due to the nature of the provided search machine that does not allow complex queries and filtering like the others. Consequently, we hand-picked suitable papers from the first 50 entries of the search result list. A PRISMA [48] flow diagram of the publication selection process is provided in Fig. 11 in Appendix B. Additionally, the individual queries and results of all services are summarized in Table 2. The queries delivered 613 hits in total which resulted in 516 unique papers. A first quick read of these papers led to 23 additional papers, referenced by some of the originally found work. Accordingly, the literature review started with a corpus consisting of 539 individual papers.

Of the 539 papers, 327 were sorted out due to the substantial divergence from the searched topic, often focusing on domains like 5G, networking, and radio. Of the remaining 212 papers, 181 directly address the field of interest, while 31 are only partially relevant. Papers were deemed partially relevant if they mentioned the surveyed topic but primarily focused on different areas such as datasets, language theory, simulation, or unrelated case studies. In conclusion, this survey mainly reviews 181 papers that directly discuss or contribute to the topic of EL in computer science.

Figure 5 presents the distribution of the 181 relevant publications over the years, categorized by publication type. The topic of EL has maintained a steady presence in conference publications, peaking in 2020. The subsequent decline in total publications may be attributed to the absence of recent topic-specific workshops. Additionally, the surge in interest in LLM technologies might have diverted attention from EL research. It is also worth noting that some recent studies may not have been openly published at the time of our literature search. We therefore expect the publication count to increase by 2024.

## 5 Taxonomy of emergent language

In the course of our comprehensive literature review, we identified recurrent instances of taxonomic inconsistencies due to missing standardization [125] and “ill-adapted metrics” [25]. Particular concern arises from the discrepancy between the concepts intended for measurement and their corresponding metrics, or the absence of such metrics [46, 51, 126–



**Fig. 5** Number of publications per year and by type. The number of publications per year is provided in the leftmost column, and the distribution of different publication types is shown in the remaining columns























128]. This section is dedicated to the formulation of a systematic taxonomy aimed at enhancing comparability and mitigating confusion within the field. This taxonomy forms the basis for the following sections and is designed to ensure consistent representation throughout the survey. It is created with the hope that it will serve as a cornerstone for future research, promoting the use of standardized terminology, particularly in the domain of language characteristics.




The taxonomy first describes the main factors influencing the EL, before categorizing the language characteristics. These influencing factors have a significant impact on the investigative possibilities of EL research and are therefore of particular importance when analyzing EL. Thus, the taxonomy introduces a classification system for the communication setting (Sect. 5.1) and communication games (Sect. 5.2) that agents encounter during language emergence. The communication setting encompasses factors such as the number of agents and the type of communication available to them. The communication game involves the environmental configuration and crucial factors influencing challenges and the complexity of multi-task learning. Furthermore, a short discussion on the concept of language priors is provided in Sect. 5.3, considering that the presence of a prior significantly influences the characteristics of the emerging language [14, 113]. We conclude this section with a comprehensive overview of the concepts and characteristics examined within EL research (Sect. 5.4). The taxonomy adheres to the six major linguistic structural levels introduced in Sect. 2.2 and illustrated in Fig. 4.

## 5.1 Communication setting

In the literature, several communication settings are represented. One distinguishing factor is the number of agents involved. We derived three classes—the single agent, dual agent, and population setting. While the single agent setting is rare, the other two are well represented in the examined literature, as shown in Table 3. A single agent is typically used to train human–machine interfaces [129] or fine-tune existing models [38]. In contrast, dual-agent settings are more common and often involve a pair of speaker-listener agents, with one agent designated as the speaker and the other as the listener exclusively [14]. The

Table 3 Classification of the communication settings in the literature reviewed

Agents	Cooperation	Symmetry	Recipients	Paper
		✗		[38, 103, 131]
		✓		[129, 132, 133]
		✗		[14–16, 25, 30, 31, 33, 34, 37, 46, 72, 102, 104, 106, 107, 109, 114, 115, 117, 119–122, 126, 134–199]
		✓		[19, 112, 113, 127, 130, 132, 133, 172, 178, 196, 200–210]
		✗		[34]
		✓		[208]
		✗		[34]
		✗		[25, 26, 30, 31, 106, 116–118, 164, 178, 179, 194, 211–216]
				[30, 31, 106, 117, 217, 218]
		✓		[21, 22, 105, 110–112, 130, 178, 201, 206, 219–229]
				[18, 20, 23, 27, 28, 112, 201, 206, 222, 226, 227, 229–239]
		✗		[32, 217, 240]
		✓		[17, 22, 108]

Agents:  Single,  Dual,  Population  
Cooperation:  Cooperative,  Semi-cooperative,  Competitive  
Symmetry: ✗ No, ✓ Yes  
Recipients:  Target,  Broadcast

population setting involves larger groups of agents in the language emergence process. This requires more computational resources but also enables more possibilities for regularization [21] and language evolution [117]. Accordingly, the population setting offers more opportunities to actively shape the process [21, 45, 130].

An additional factor that shapes the communication setting is the type of cooperation inherent in the setup. Determining the level of cooperation or competition feasible within the setting is a fundamental decision and closely related to the choice of the language game. We derived three options—the cooperative, semi-cooperative, and competitive type. In the literature reviewed, the majority of studies adopted a fully cooperative setting approach, where agents fully share their rewards and lack individual components. The emphasis on strongly cooperative settings is justified given that AI agents utilize a common language to coordinate and will not learn to communicate if they dominate without communication [34]. Only a few publications explore semi-cooperative settings that incorporate individual rewards alongside shared rewards, introducing the challenge of balancing tasks and rewards [29, 108]. A semi-cooperative setup can be compared to a simplified social scenario with overarching societal objectives, while also encompassing additional individual interests and goals. In contrast, investigations of fully competitive settings are rare, with only one work in which agents compete for rewards without a common goal [34]. This scarcity likely arises from the fact that such settings inherently favor deceptive language as the only advantageous strategy, making its emergence improbable without any cooperative element [34].

The third important factor in communications settings is symmetry. Agents should treat messages similarly to regular observations; otherwise, they risk devolving into mere directives [146]. Building on this premise, the symmetry is important for promoting robust language emergence, as opposed to languages that consist primarily of directives. An illustrative example of asymmetric settings is the commonly used, and aforementioned, speaker-listener paradigm [14, 51, 168]. Languages developed in such settings are severely limited compared to NL, lacking the capacity for diverse discourse or even basic information exchange beyond directives [146]. Contrary to promoting informed choices by the listener, the speaker-listener approach emphasizes obedience to commands. Conversely, a symmetric setting facilitates bi-directional communication, thereby allowing for more comprehensive language development [110, 202]. For instance, symmetry may result from agents being randomly assigned roles within the interaction [202]. Additionally, symmetry can emerge from tasks that are inherently balanced, such as negotiations between equal partners where both parties have equivalent roles and objectives [17].

At the population level, another important consideration is the choice of recipients, i.e., between targeted and broadcast communication. While broadcast communication facilitates broader information dissemination across the agent group, targeted communication promotes the development of social group dynamics and regularization [222, 226]. For example, targeted communication strategies can be learned through mechanisms such as attention [222], and agents can develop minimized communication strategies that optimize group performance [229].

Table 3 provides a summary of these settings and their variations. The setting categories presented and their implementation are not inherently tied to the language itself but are crucial in determining the likelihood of meaningful language emergence and in shaping the features and experimental possibilities. These initial choices dictate the options for the language development process, the opportunities for regularization [21, 130], and the requirements regarding computational resources.

**Table 4** Overview of the distribution of game types in the reviewed literature

Type	Paper
Referential	[14–16, 25, 26, 31–34, 37, 38, 46, 49–51, 53, 57, 58, 72, 102, 104, 106, 107, 110, 111, 113, 114, 116, 117, 120, 126, 130, 137–141, 144, 145, 147–152, 154–158, 160–162, 165, 167, 168, 170, 171, 173–175, 177, 178, 180–182, 184, 186–191, 193, 195, 199, 206, 211, 212, 214, 215, 220, 221, 224, 225, 230]
Reconstruction	[37, 49, 50, 57, 58, 109, 113, 118, 119, 121, 122, 134, 135, 137, 142, 143, 153, 155, 157, 162–164, 166, 179, 183, 197, 200, 203, 210, 218, 219, 228]
Question-answer	[19, 21, 46, 58, 105, 108, 127, 169, 194, 202]
Grid world	[18, 20, 22, 27, 28, 30, 33, 55, 58, 103, 112, 115, 131–133, 136, 137, 146, 148, 154, 159, 161, 176, 190, 192, 198, 205, 209, 213, 222, 226, 227, 229–234, 237, 239–244]
Continuous World	[58, 172, 207, 223, 227, 229, 234, 236, 238, 239]
Other	[17, 20, 23, 28–30, 55, 140, 158, 196, 204, 208, 217, 234, 235, 238, 242]

## 5.2 Language games

Distinct communication settings are implemented through different communication games. In this section, we provide an overview of the games used in EL literature. Specifically, we focus on a subset of these games known as language games, that emphasize explicit communication via a predefined language channel. The literature identifies several categories of language games, such as referential games, reconstruction games, question-answer games, grid-world games, among others. Our review indicates that these categories represent the most commonly used game types. To give a comprehensive view, Table 4 lists the publications that focus on these game types. In the following, we offer a concise overview of each category to provide a clearer understanding of their characteristics.

*Referential game* Generally, a referential game, also called signaling game, consists of two agents, a sender and a receiver [14]. The objective of this game is for the receiver to correctly identify a particular sample from a set, which may include distractors, solely based on the message received from the sender. This set can consist of images [14, 16, 149], object feature vectors [140], texts [140], or even graphs [186]. To accomplish this selection task, the sender must first encode a message that contains information about the correct sample. In game design, a fundamental decision arises regarding whether the sender should only view the correct sample or also some distractors that may differ from those presented to the receiver [14]. Another design decision concerns the receiver's side, specifically the number of distractors and whether to provide the original sample shown to the sender or only a similar one for selection [160]. However, only the encoded message is transmitted to the receiver, who then selects an item from their given collection.

*Reconstruction game* The reconstruction game is similar to the referential game, but with a key difference: the receiver does not have a collection to choose from. Instead, the receiver must construct a sample based on the message from the sender, aiming to replicate the original sample shown to the sender as closely as possible [118, 122]. Consequently, this game setup resembles an autoencoder-based approach, with a latent space tailored to mimic or facilitate language [162]. Therefore, the key distinction between reconstruction and



referential games, often used interchangeably in early literature, lies in the collection's presence (referential) or absence (reconstruction) for the receiver to select from [155].

*Question-answer game* The question-answer game is a variant of the referential game, but without strict adherence to previously established rules. It operates as a multi-round referential game, allowing for iterative and bilateral communication [202]. Unlike referential and reconstruction games, the question-answer game explicitly incorporates provisions for multiple rounds with follow-up or clarifying queries from the receiver [19, 21]. Question-answer games have introduced intriguing inquiries and avenues for exploring the symmetry of EL, although they are not as widely adopted [19].

*Grid world game* Grid world games use a simplified 2D environment to model various scenarios like warehouse path planning [22], movement of objects [205], traffic junctions [226, 231], or mazes [20, 115]. They offer design flexibility, allowing agents to be part of the environment or act as external supervisors. Design choices also include environment complexity and the extent of agents' observations. Although common in the literature surveyed, implementations of grid world games vary widely in their design choices and are thus a very heterogeneous group.

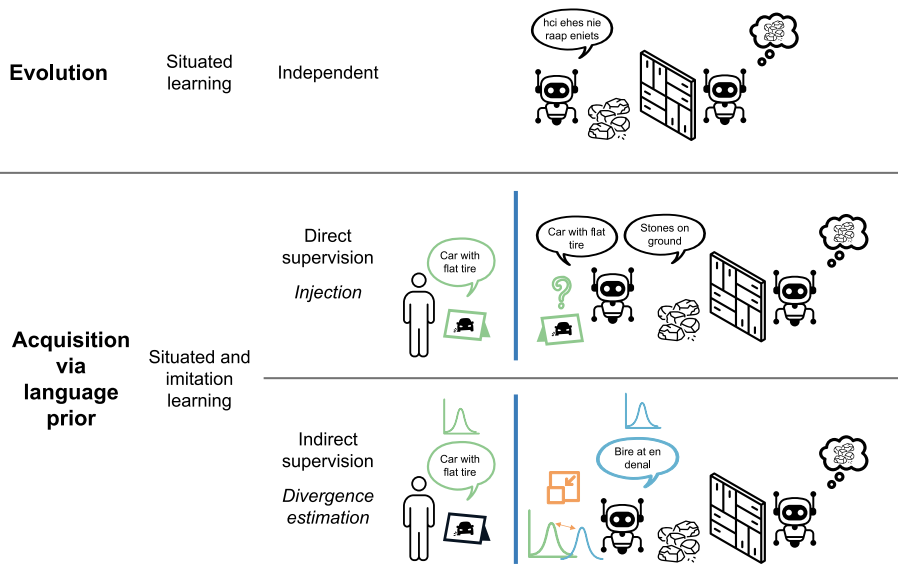
*Continuous world game* Continuous environments add complexity to the learning process [229, 236]. In EL approaches, the learning landscape involves multi-task settings where one task is tackled directly within the environment while another involves language formation. Playing continuous world games, whether in two or three dimensions, presents challenges and adds a greater sense of realism and intricacy. These environments have the potential to make it more feasible to deploy EL agents in real-world scenarios compared to discrete environments [207].

*Other* The literature on EL also covers various other game types besides those mentioned earlier, such as matrix communication games [29, 30], social deduction games [23, 235], or lever games [20, 234]. These game types contribute to the creation of new language emergence settings, often designed to target specific aspects or characteristics of language development. They are valuable tools to explore and understand the complexities of EL in different contexts.

In summary, although many language games have been developed, comparing different games can be complex and understanding the nuances of each game can prove challenging. A promising direction would be for the research community to collectively agree on a standardized subset of these games as benchmarks. By focusing on a representative set of games from different categories, researchers could systematically explore different settings, ensuring that new approaches are rigorously tested and their results are directly comparable across studies. This would accelerate the maturation of the field of EL research, foster collaboration, and enable the community to better identify and address key challenges.

### 5.3 Language prior

EL research occasionally utilizes a concept known as a *language prior* to incorporate structures from human NLS into the emerging language. A language prior is used to impose specific linguistic structures on the emerging language, making it easier to align with human NL and improve interpretability and performance. This prior can be implemented through supervised learning [14, 19, 113], also known as injection, or through divergence estimation [15]. An overview of prior usage in the literature surveyed is given in Table 11 in Appendix A.



**Fig. 6** Language evolution and the different language prior options for acquisition approaches. The *evolution* of language and the guided *acquisition* are contrary approaches. Language evolution is based on a situated learning environment and the *independent* design of an appropriate language scheme. Acquisition, on the other hand, is based on the existence of a prior and the combination of situated and imitation learning. The prior, generally a NL, is either introduced into the learning process via injection or divergence estimation. *Injection* is a direct supervision approach that uses examples of prior usage to inform and train the language learner. *Divergence estimation* is an indirect supervision method which utilizes a distribution representation of the prior and learner language. The goal is to limit the divergence of these distributions

Given this context, research on EL can be divided into two main areas. The first area focuses on independent situated learning and does not use priors, so that communication and language emerge spontaneously [18]. The second area explores imitation learning-based approaches, which aim to replicate NL behavior in artificial agents using priors [103]. However, it is important to note that these approaches differ from LLM because language acquisition in EL is generally task-oriented. In academic literature, the independent situated learning environment is often referred to as the evolution-based approach, while the imitation learning-related approach is commonly known as the acquisition-based approach. The term *evolution* implies starting from scratch, while *acquisition* involves learning an existing language [168]. The terminology and different approaches are depicted in Fig. 6.

In addition, the concepts of community and generational learning are closely related [21, 116, 130]. In these methods, language emerges through iterative learning across and within agent sub-groups called communities. Generational learning additionally involves older generations of agents training younger ones using previously developed communication as a foundation [105, 117]. Language transfer across groups or generations can be interpreted as an iterative prior. However, this method remains a fully evolutionary approach in the absence of a deliberately designed prior.

**Table 5** Overview of the use of channel types in the reviewed literature

Type	Paper
Discrete	[14–19, 21, 23, 25, 26, 29–33, 37, 38, 46, 49, 51, 53, 55, 72, 102, 103, 105–122, 126, 127, 129–132, 134, 135, 137–139, 141–152, 154–168, 170–173, 175–178, 180–188, 190–195, 197, 199, 200, 202–206, 209, 210, 212–215, 217, 218, 220, 224, 225, 229–232, 235, 240, 241, 243]
Continuous	[20, 22, 28, 57, 104, 136, 169, 174, 179, 207, 211, 216, 219, 222, 223, 226–228, 236–239]
Both	[27, 34, 35, 58, 140, 153, 189, 196, 198, 221, 233, 234]

## 5.4 Language characteristics

As discussed in Sect. 2.2, language is a complex, multifaceted system [88, 90]. Therefore, it is essential to establish a comprehensive taxonomy of its properties to provide a unified framework for EL research. This taxonomy will not only facilitate the unambiguous categorization of metrics used in EL studies (cf. Section 6) but will also enhance the comparability and comprehensibility of approaches and results within the field. As shown previously in Fig. 4, NL can be divided hierarchically into distinct characteristics [99–101]. The following sections provide a categorization of the reviewed publications along these characteristics, occasionally breaking them down into smaller sub-characteristics if relevant.

### 5.4.1 Phonetics

The phonetics of a language inherently represents its medium, delineating the constraints of the specific communication channel [3, 101]. These media or channels can be either discrete or continuous; for example, an audio channel is continuous, while a symbolic channel is typically discrete. Regardless of the type, they lay the foundation for the nature of communication. However, for EL research the discrete case is of particular importance, as it closely mirrors NL as we understand it [35]. Although humans use a continuous phonetic medium for communication, some degree of discretization is essential to establish a common ground for efficient communication [100].

**Table 6** Overview of the use of vocabulary types in the reviewed literature

Vocabulary	Paper
Binary	[14, 16, 23, 27, 30, 31, 49, 55, 58, 109, 110, 112, 119, 130, 136–140, 144, 146, 151, 155, 163, 188, 190, 199, 203, 206, 209, 210, 215, 218, 229–235, 243]
Token	[15, 17–20, 22, 23, 25, 28, 29, 32–34, 37, 46, 49, 51, 53, 58, 72, 102, 105–108, 111, 113–118, 120–122, 126, 127, 129, 131, 132, 134, 135, 140–143, 145, 147–150, 153–158, 160–162, 164–168, 170–173, 175–187, 189, 191–193, 195–198, 202, 204, 205, 207, 211–214, 216, 217, 219–227, 233, 234, 236–241]
NL	[15, 19, 21, 26, 38, 58, 103, 152, 159, 194]
Sound	[58, 153, 200]
Picture	[57, 58, 104, 169, 174]

Table 5 provides an overview of the reviewed papers, categorized according to the continuous or discrete approach. Notably, some papers explore both approaches, providing valuable insights for researchers interested in the basic aspects of phonetics research in EL.

### 5.4.2 Phonology

Phonology encompasses the actively used vocabulary and determines the part of the medium that is utilized for communication. We identified five different types of vocabulary actively researched, however, some of them are rare to find in the literature. Table 6 summarizes the results of our survey regarding vocabulary types in EL research. One commonly used phonological type in EL is a binary encoding, while an even more prominent type is a token-based vocabulary. However, these two phonological classes are not always distinct, as a token-based vocabulary often builds upon a binary encoded representation [112].

The other three types, which are distinct from the two most prominent, are rarely mentioned in the literature reviewed. One of these types involves using NL vocabulary, such as all the words from an English dictionary. While this approach enforces the NL resemblance of the EL, it also drastically limits the emergence and associated benefits [19]. Essentially, this phonological preset strips the agents of the possibility to shape phonology and morphology. The other two vocabulary types being referred to are sound and graphics. The former enables agents to produce and process sound [200], while the latter focuses on enabling agents to draw and analyze graphical representations [104, 174]. Both mediums present challenges in ensuring discretization, which may be the reason why they are not as extensively researched in EL.

### 5.4.3 Morphology

Morphology governs the rules for constructing words and sentences, meaning the overall ability to combine individual elements, also called tokens, into words and to combine those words into sentences [101]. This is particularly relevant in the field of EL due to the prominent division of existing work based on morphological setup and options. The most significant differentiation is between the use of a fixed or flexible message length. Table 7 demonstrates that much of the existing work employs fixed message lengths, despite this setup not being comparable to NL [15]. For instance, NL users, such as humans, have the ability to adjust the length of their message to fit their intention, which may vary depending

**Table 7** Overview of the characteristics of message length in the literature reviewed

Length	Paper
Fixed	[14, 16, 17, 20, 22, 23, 25, 27–32, 34, 55, 102, 104–106, 108–113, 115, 117–122, 127, 130, 134–137, 139, 143–146, 150, 151, 154–159, 164–167, 169, 171, 172, 174, 176, 179, 180, 182, 184–190, 192, 194–196, 198–200, 202–205, 207, 209, 210, 212–216, 218–223, 225–227, 230–241, 243]
Variable	[15, 18, 19, 21, 26, 33, 37, 38, 46, 51, 53, 72, 103, 107, 114, 116, 126, 129, 131, 132, 138, 141, 142, 147–149, 152, 153, 160–163, 168, 170, 173, 175, 177, 178, 181, 183, 191, 193, 197, 211, 217, 229]
Both	[140, 206]

on the audience, medium, or communicative goal. When communicating with colleagues, they may use shorter sentences to be efficient, while more detailed explanations may be used when conversing with friends.

Accordingly, this characteristic can be measured using metrics that assess word formation and vocabulary. Based on the metrics found in the literature, distinct features of language morphology can be quantified. Specifically, this refers to the compression of language and the presence of redundancy or ambiguity.

*Compression* Compression [105], also known as combinatoriality [224], refers to the ability of a communication system to combine a small number of basic elements to create a vast range of words that can carry meaning. This feature of discrete communication is crucial in producing comprehensive and flexible communication with limited resources, and is an essential characteristic of NL. We assume that using compressed language is generally favorable for language learners as it reduces the burden of learning [105].

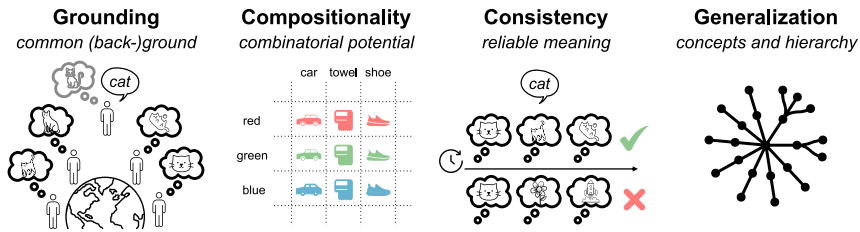
*Redundancy or ambiguity* In NL, words and phrases can have redundant or ambiguous meanings. Redundancy occurs when multiple words convey the same meaning, while ambiguity arises from a limited vocabulary [14, 114]. The addition of this characteristic in the morphology subsection rather than the semantics subsection may be controversial. We argue that any metric measuring redundancy or ambiguity provides more useful information about the morphology, encompassing the form and size of the vocabulary, than it does about the semantic range and capabilities of the language. However, to quantify redundancy or ambiguity, we must establish semantic meaning first.

#### 5.4.4 Syntax

The syntax of a language establishes the grammatical rules that govern sentence formation. Consequently, syntax plays a central role in establishing a functional correspondence between emerged language and NL [245]. This specific characteristic of a discrete language is underrepresented in current EL literature. However, we found two examples in the body of literature discussing syntax in EL. Ueda et al. [102] introduced a method to examine the syntactic structure of an EL using categorial grammar induction (CGI), which is based on the induction of categorial grammars from sentence-meaning pairs. This method is straightforward in simple referential games. Additionally, van der Wal et al. [193] introduced unsupervised grammar induction (UGI) techniques for syntax analysis in EL research. We discuss the methods they use to measure and analyze syntax in an EL briefly in Sect. 6.3.

#### 5.4.5 Semantics

Semantics is concerned with the literal meaning of language constructs and is a dominant topic in current EL research, as shown in Table 12 in Appendix A. EL studies often focus on establishing useful and meaningful communication between agents, making semantics a central feature [35]. It serves as a crucial tool for distinguishing actual information exchange from mere noise utterances [168, 202]. Given the complexity of capturing the meaning of literal language in a single metric, several features have been introduced to measure the semantics of EL. In particular, these features include grounding, compositionality, consistency, and generalization, as shown in Fig. 7. Table 8 provides an overview of the literature addressing the individual semantic features in EL.



**Fig. 7** Semantics features of language addressed in EL research: *grounding*, *compositionality*, *consistency*, and *generalization*

**Grounding** A language is considered grounded when it is deeply intertwined with the environment, for example, when it is tightly bound to environmental concepts and objects [149, 184, 246]. Grounding is essential for the interoperability of individuals and is particularly important in NL communication, where meaningful interaction requires shared understanding [33, 36, 112]. While in theory, an EL can establish a unique form of grounding using self-emerged concepts distinct from those in NL, deriving a useful metric for such a scenario proves challenging. This difficulty arises from the need to compare ELs to existing and comprehensible grounding principles typically found in NLs [50, 112, 127].

**Compositionality** When a language exhibits compositionality, its components can be rearranged or replaced by conceptually equivalent words without changing the overall meaning [121, 122]. Compositionality facilitates the construction of higher-level concepts, using conceptual foundations to enable efficient language expression [46, 51, 135]. For example, NLs partition concepts such as objects and their attributes to allow compositional constructions [18, 155]. As a result, we can describe variations of a single object using different words from the same semantic concept, such as ‘blue towel and ‘red towel for the object towel and the semantic concept of color. Similarly, we can attribute specific properties to different objects using the same phrase, as in ‘green towel and ‘green car. Ultimately, compositionality is beneficial for the learning process [36, 122] and promotes efficient and rich language use, even in systems with limited memory capacity [18, 108, 119].

**Consistency** Merely having grounded words in a language does not necessarily guarantee its semantic quality. In addition, consistency is essential for a language to convey meaningful and practical information effectively [36, 247]. If the words within a language lack consistency in their literal meanings, they will not facilitate effective communication. Therefore, even if a language is semantically grounded and compositional, its utility is compromised if the words exhibit inconsistent literal meanings [127]. While words can change their general meaning to fit the context, their literal meaning should remain consistent to keep their usefulness [33].

**Generalization** Generalization serves as a cornerstone of NL, allowing humans to communicate about topics ranging from simple to complex, broad to specific, and known to unknown, all with a relatively limited vocabulary [25, 135]. A language that excels at generalization enables its users to navigate different levels of complexity, facilitating hierarchical descriptions of concepts and relationships [37]. Consequently, generalization and compositionality are closely related, as they both contribute to the flexibility and expressiveness of language [36, 117, 168]. This ability to generalize not only enriches communication but also underscores the adaptability and robustness of human language.

**Table 8** Semantic features discussed in the reviewed literature

Feature	Paper
Grounding	[14–19, 21, 26, 46, 58, 104, 106, 107, 114, 127, 144, 146, 148, 151–153, 165, 169, 171, 174, 175, 185, 192, 198, 207, 216, 231, 234]
Compositionality	[15, 18, 21, 25, 35, 37, 46, 50, 51, 58, 72, 102, 105–107, 114, 116–119, 121, 122, 127, 134, 135, 138, 141, 147, 149, 153–157, 159, 160, 164, 166–168, 177, 179, 180, 182–184, 186, 187, 191, 195, 197, 198, 213, 224]
Consistency	[18, 21, 28, 29, 32, 33, 37, 51, 58, 72, 104, 108, 110, 114, 116, 118, 134, 136, 150, 151, 159, 160, 164, 167, 171, 173, 176, 177, 185, 192, 195, 202, 225, 230, 231, 240]
Generalization	[14, 25, 32, 37, 51, 58, 72, 103, 104, 107, 114, 117, 118, 120, 122, 127, 135, 139, 147, 149, 153–155, 157, 164, 167, 168, 176, 177, 182, 183, 185, 186, 189, 194, 195, 197, 211, 223, 228]

### 5.4.6 Pragmatics

The final dimension of EL research is pragmatics. This field of study examines how language is employed in context, particularly in interactions, and how it conveys information [101, 160, 248]. By evaluating the pragmatics of the linguistic structure, we can ascertain whether EL is itself useful and utilized effectively. While this assessment may be feasible based on rewards in a standard RL setting, integrating communication into such environments increases the complexity. This is because most setups do not separate the agent's environment interaction from its communication capabilities, thereby expanding the network's capacity, and making it difficult to attribute an increase in reward directly to EL [35].

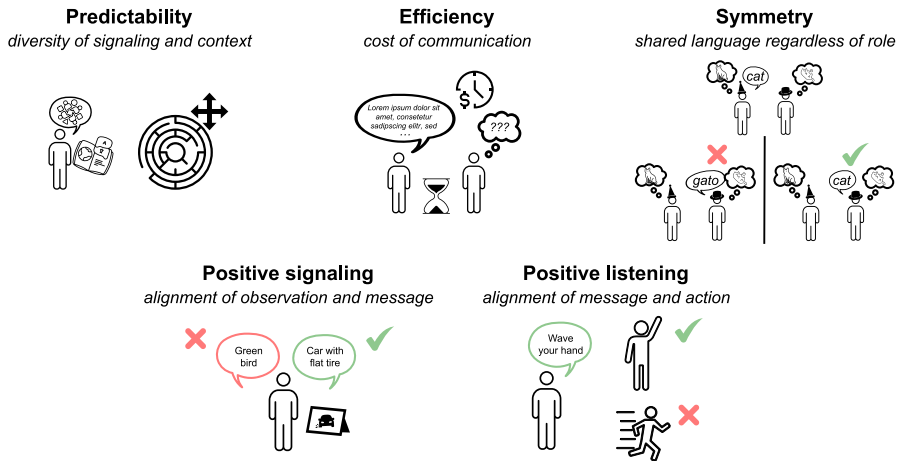
As outlined in Table 9 and depicted in Fig. 8, five distinct features have been identified: predictability, efficiency, positive signaling, positive listening, and symmetry. These features are essential for assessing the constructive impact and utilization of EL. Understanding how agents employ language is crucial in evaluating its effectiveness and overall benefit.

**Predictability** Predictability is concerned with the assessment of the complexity of the context, including the action space within the environment. When actions exhibit less diversity, it becomes more feasible to coordinate without communication [110]. For instance, in a simple grid-based environment where agents have only two possible actions—moving left or right—agents can often achieve their objectives without the need for communication. In such a scenario, the limited action space reduces the necessity for EL, as agents can predict each other's movements based on past behavior or simple rules. However, in a more complex environment where agents have multiple actions, such as navigating a maze with numerous paths and obstacles, the need for effective communication

**Table 9** Pragmatic features discussed in the reviewed literature

Feature	Paper
Predictability	[110]
Efficiency	[115]
Positive signaling	[17, 29, 34, 35, 58, 108–110, 124, 146, 181, 186, 187, 200, 202, 207, 230, 234, 240]
Positive listening	[29, 34, 35, 58, 108–110, 112, 124, 146, 176, 186, 187, 200, 202, 207, 230, 234, 240]
Symmetry	[105, 110]





**Fig. 8** Pragmatics features of language in EL research. These illustrations are intended to promote an intuitive understanding of the categories *predictability*, *efficiency*, *positive signaling*, *positive listening*, and *symmetry*

increases. Here, EL can significantly enhance coordination by allowing agents to share information about their positions, plans, or discoveries, thus improving their overall performance in navigating the maze. Therefore, it is essential to compare the diversity of signaling and context attributes to evaluate the potential benefit of EL.

**Efficiency** Efficiency is a critical aspect considered whenever communication entails a cost. This is particularly true in the context of modeling the emergence of NL and the broader objective of employing EL for human-computer interaction(HCI). In EL settings, the achievement of concise communication is contingent upon the presence of an opportunity cost [115]. Without such a cost, there is no incentive to communicate concisely, making EL ineffective as an intermediary for HCI. When communication is accompanied by a cost the necessity for efficiency in communication becomes paramount. In such scenarios, the objective is to minimize the cost while maximizing the effectiveness of communication within a given task.

**Positive signaling** The concept of positive signaling is concerned with the degree of alignment between the observations, knowledge, and experience of the message producer and their communication output [29]. The objective is to guarantee the transmission of useful information, or at the very least, information that the speaker can discern through knowledge or observation [181]. This feature assumes that all communication should be relevant to something observable, known or tangible to the speaker. Thus, it stipulates that the situational information content of the produced signal is crucial for a language that can be used in a contextually meaningful way.

**Positive listening** Positive listening, in contrast to positive signaling, focuses on the role of the message receiver, to evaluate the active processing of incoming information [29]. From a pragmatic point of view, it makes sense to process incoming messages in a meaningful way. However, the definition of meaningful processing is broad. Positive listening as defined in this taxonomy, contrary to earlier work [29, 202], does not necessarily require a connection between the incoming message and the subsequent action; it is much more about active processing, which may or may not lead to inclusion in the choice of

action. Thus, active engagement followed by rejection or disregard is also considered positive listening in the context of this taxonomy.

**Symmetry** Symmetry in EL is defined as the consistency in language usage among participating agents [105, 110]. This concept applies to MARL settings where agents can assume multiple roles, such as message producer and message receiver. Symmetry plays a crucial role in achieving convergence on a shared and aligned EL. For instance, if an agent employs language differently depending on whether it is sending or receiving messages so that words have varying meanings based on the assigned role the EL setting is considered asymmetric. In such instances, rather than learning a collectively and contextually grounded language, agents develop individual protocols specific to their respective roles [110]. This would suggest that there is no common language, but rather separate codes that can only be applied to specific combinations and conditions. For this reason, this pragmatic feature is particularly relevant, since the aim of EL is a common language.

## 5.5 Summary of the taxonomy

Our proposed taxonomy systematically categorizes the key features of EL systems, including communication settings, language games, language priors, and language characteristics. The latter is particularly detailed, with sub-characteristics and their features aligned with the major levels of linguistic structure, as previously illustrated in Fig. 4. This comprehensive taxonomy enables a standardized comparison of approaches in the EL literature, highlighting the opportunities and properties associated with individual options and topics in EL research. Specifically, by applying this taxonomy, especially in terms of language characteristics, we can uncover the capabilities and potentials of various EL approaches. This facilitates a more detailed, comparable, and insightful analysis of EL.

## 6 Metrics

This section provides a comprehensive categorization and review of existing metrics used in EL research. The section is organized along the same categorization used in Sect. 5.4. Note that the categories of phonetics and phonology are excluded from this discussion, as these aspects are predetermined settings in the current EL literature and thus not yet targeted by metrics.

We begin by introducing the notational system used for all metrics to ensure consistency and facilitate ease of use. We then describe the metrics within each category, detailing the individual metric and adapting it to our notation. For each metric, we provide references to both original sources and additional literature, if available, to enable further exploration beyond the scope of this work. Figure 9 provides a visual summary of the existing metrics and their correspondence to the language characteristics. An extended version including all references for the individual metrics is provided in Fig. 12 in Appendix B

### 6.1 Notation

Given the complexity and variability within the EL field, it is crucial to establish a unified and coherent notation system. In this section we present a standardized mathematical notation designed to be consistent across the various aspects of EL research, thereby facilitating clearer communication and comparison of results within the community. This approach aligns with our broader goal of advancing the field through a common taxonomy

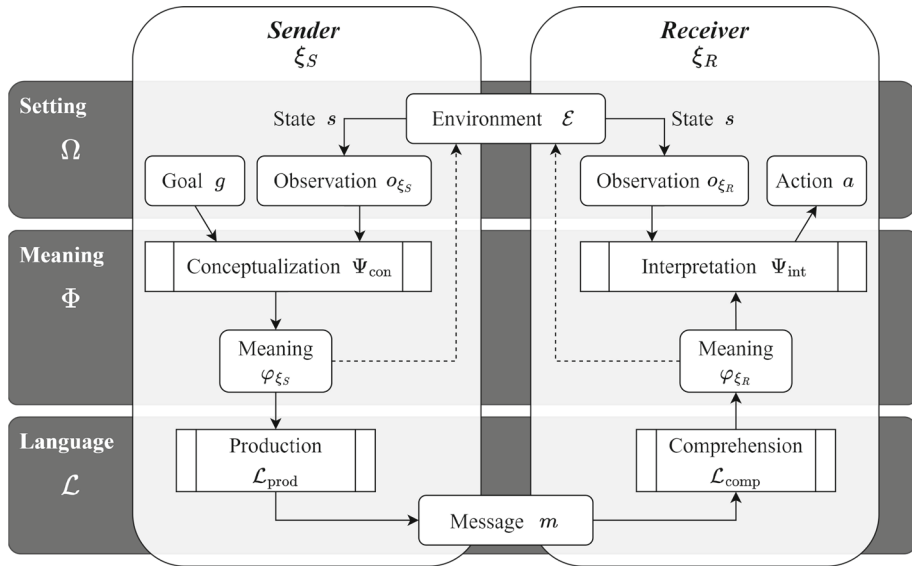


**Fig. 9** Graph presenting a visual representation of the metrics identified in the surveyed literature, sorted by language characteristics. ●●●●: Language characteristics (inner nodes) ●●●●●: Individual metrics (outermost nodes) (Color figure online)

that supports the development of measurable and interpretable ELs. Throughout this section we focus on finite and discrete languages, although some of the definitions and metrics discussed here are also applicable to continuous languages. These languages offer a more straightforward mapping to NLS, making them particularly relevant to the study of EL systems.

### 6.1.1 Definition

In alignment with the semiotic cycle introduced in Sect. 2.2, our notation is organized into three interconnected spaces: setting, meaning, and language. The setting space encompasses the typical elements of RL, providing the foundational environment in which agents operate. The meaning space incorporates a representation learning endeavor, whereby sensory input is integrated with decision-relevant information to generate a coherent internal representation. Finally, the language space encompasses both the production and comprehension of discrete messages, encapsulating the communication process. These components, illustrated in Fig. 10, will be introduced and explored in detail in the following paragraphs.



**Fig. 10** The proposed notation is integrated within the semiotic cycle framework (cf. Figure 3). The language  $\mathcal{L}$ , along with its associated mapping functions  $\mathcal{L}_{prod}$  (production) and  $\mathcal{L}_{comp}$  (comprehension), forms the core of EL research. These linguistic components, however, are underpinned by the representations learned in the meaning space  $\Phi$ , which play a crucial role in guiding and shaping effective communication. All elements in the sender or receiver domain are entity-specific, but we only index duplicate symbols to keep it concise. Adapted from [45, 98]

**Setting** The overall setting, consisting of the environment, actions, goals, and other typical RL elements, is denoted by  $\Omega$ . Let  $\xi$  denote the set of all entities in the system, with an individual entity represented as  $\xi_i \in \xi$ . Each entity can assume specific roles, such as the sender ( $S$ ) or receiver ( $R$ ) in a communication scenario. An entity can assume several roles over the course of the entire communication scenario. However, for an individual message exchange, an entity assumes one specific role. We represent the role of an individual entity  $i$  by  $\xi_{i,j} \in \xi_i$ , where  $j$  specifies the role (e.g.,  $j = S$  or  $j = R$ ).

Entities interact with their environment  $\mathcal{E}$  through actions, denoted as  $a$ , which belong to the set of possible actions  $A$ , such that  $a \in A$ . The action taken by a specific entity  $\xi_i$  is represented as  $a_{\xi_i}$ . The state of the environment at any given time is denoted by  $s$ , which is an element of the state space  $S$ , so that  $s \in S$ . As the system progresses over time, denoted by discrete points in time  $[0, \dots, t]$ , the sequence of states and actions forms a trajectory  $\tau$ , generally expressed as  $\tau = \{s^0, a^0, \dots, s^t, a^t\}$ . It is important to note that the entities described here do not necessarily correspond to autonomous agents in the traditional sense; they could also represent ground truth models, human participants, or abstract constructs that lack the direct interaction capabilities typically associated with agents. Despite this distinction, for the sake of clarity and consistency, we will refer to these entities as agents in the following sections.

Given the importance of partial observability in EL research [18, 50, 172, 228], it is essential to consider that agents only have access to their own observations, denoted as  $o_{\xi}$ , which are derived from the underlying state  $s$ . An individual observation  $o_{\xi}$  is an element of the collection of observations of an agent  $O_{\xi}$ , which is a subset of the observation space  $O$ ,

so that  $o_\xi \in O_\xi \subseteq O$ . In our framework, an observation  $o_\xi$  effectively replaces the ‘world model’ component from the traditional semiotic cycle, highlighting the localized and subjective nature of an agent’s perception in partially observable environments.

Referential games (cf. Table 4) are frequently employed in EL literature. They often operate on individual, static samples that are drawn from a corresponding dataset or distribution. In doing so, they differ from traditional RL setups that emphasize sequential decision-making and environmental interactions over time. In such cases, rather than speaking of a state  $s$  or an observation  $o$ , we use the term sample  $k$ , which is an element of the collection of all samples  $K$ , so that  $k \in K$ . The specific nature of a sample depends on the environment; for example, in an image-based sender-receiver game, the sample would be an image. Each sample is represented by its feature vector  $f$ , which belongs to the feature space  $F$ , so that  $f \in F$ . The feature vector corresponding to a specific sample  $k$  is denoted by  $f_k$ .

In EL settings, the communicative goal  $g$  of an agent may differ from the (reinforcement) learning task goal. In addition, depending on the game, the sender and receiver may have distinct goals. These are important factors to consider when evaluating the communicative behavior.

**Meaning** In our notation, the meaning space, denoted by  $\Phi$ , serves as the critical intermediary between the setting space and the language space. The meaning space represents the semantic connections derived from the provided information. Each element within this space, represented by a specific meaning vector  $\varphi \in \Phi$ , captures the essence of concepts or objects as understood by the agent. These meaning vectors are critical to the processes of language comprehension and production, as well as to the processes of conceptualization and interpretation, that allow an agent to effectively use inputs and generate outputs in the setting space (cf. Figure 10).

The representation mappings  $\Psi$  within the meaning space are agent-specific and referred to as  $\Psi_{\text{con}}$  and  $\Psi_{\text{int}}$ , given in Eq. 1. These mappings enable the transition between an arbitrary space  $\chi$ , such as sensory inputs or raw data, and the meaning space, where the data acquires semantic meaning.  $\Psi_{\text{con}}$  refers to the conceptualization process that transforms raw, uninterpreted data into meaningful representations within  $\Phi$ . Conversely,  $\Psi_{\text{int}}$  denotes the interpretation process that translates these meaning vectors back into the arbitrary space that can represent any external or internal stimuli. These mappings are critical to the agent’s ability to both understand its environment and communicate effectively within it through language that is both grounded in and reflective of the underlying reality with which the agents interact.

$$\Psi = \begin{cases} \Psi_{\text{con}} : \chi \rightarrow \Phi \\ \Psi_{\text{int}} : \Phi \rightarrow \chi \end{cases} \quad (1)$$

**Language** In our proposed framework, a message  $m$  belongs to the message space  $M$ , such that  $m \in M$ . Each message encapsulates semantic and pragmatic content, serving as a vehicle for meaningful communication between agents. A message is composed of individual words  $w$ , which are elements of a finite collection  $W$ , commonly referred to as vocabulary, lexicon, or dictionary. In this context, each word is considered a semantic unit that carries (intrinsic) meaning. At the lowest level, a word is composed of characters or symbols  $v \in \Upsilon$ . These atomic characters, while essential for constructing words, do not independently carry semantic meaning. Instead, they function as elements of a finite set  $\Upsilon$  from which any number of meaningful words can be composed.

Building on the formalization from [119], we describe the message space  $M_\xi$  of an agent  $\xi$ , which represents the agent's language capabilities from a compositional standpoint. The message space  $M_\xi \subseteq M$  is composed of a set of messages or strings  $m_\xi$ , each constructed from words within  $W_\xi$ , as shown in Eq. 2. Further, each  $w_\xi \in m_\xi$  is composed of a set of characters  $v_\xi \in Y_\xi \subseteq Y$  utilized by the agent, given by Eq. 3.

$$\begin{aligned} m_\xi &\subseteq M_\xi \\ &= \{w_\xi \mid w_\xi \in W_\xi \subseteq W \wedge |w_\xi| \geq 0\} \end{aligned} \quad (2)$$

$$\begin{aligned} w_\xi &\subseteq W_\xi \\ &= \{v_\xi \mid v_\xi \in Y_\xi \subseteq Y \wedge |v_\xi| \geq 0\} \end{aligned} \quad (3)$$

A language  $\mathcal{L}$  encompasses a set of mapping functions that facilitate the transformation between the message space  $M$  and other arbitrary spaces  $\chi$ . These mappings are agent-specific and enable both the production of messages, denoted as  $\mathcal{L}_{\text{prod}}$ , and the comprehension of messages, denoted as  $\mathcal{L}_{\text{comp}}$ . This framework aligns with the linguistic level description of the semiotic cycle presented in Fig. 3. Within this context, we formally define a language  $\mathcal{L}$  in Eq. 4.

$$\mathcal{L} = \begin{cases} \mathcal{L}_{\text{prod}} : \chi \rightarrow M \\ \mathcal{L}_{\text{comp}} : M \rightarrow \chi \end{cases} \quad (4)$$

These emerging mapping functions are not necessarily injective, meaning that distinct inputs from the space  $\chi$  could potentially be mapped to an identical message within  $M$  [117, 120]. Conversely, distinct messages within  $M$  could also be mapped to the same value in  $\chi$ . While this non-injectivity adds a layer of complexity to the expressiveness of the language, it also introduces a degree of flexibility that can be advantageous in certain communication scenarios. For example, it allows for synonymy (where different messages convey the same meaning), which can provide redundancy and flexibility in communication, and homonymy (where the same message may have multiple interpretations depending on context), which can facilitate more nuanced and context-dependent communication. These natural phenomena, though challenging, are well-documented in NLPs and are of particular interest in the design and evaluation of artificial communication systems [101, 249]. However, managing these complexities effectively is crucial, as unchecked non-injectivity could lead to ambiguities that complicate communication rather than simplifying it.

### 6.1.2 Important notes

The notation presented here is designed to be comprehensible and thorough; however, it may not be directly applicable in all cases to existing works, as these employ different wordings. A lot of existing work uses the term 'word', which in our notation describes element carrying semantic meaning, and 'symbols', which in our notation serve as fundamental building blocks without inherent semantic meaning, interchangeably [19, 35, 134, 150, 186]. Furthermore, a considerable proportion of existing literature utilizes a multitude of different definitions for concepts such as 'meaning space' [128, 138, 181], 'ground-truth oracle' [134, 148], and other pivotal elements. In our endeavor to establish a unified framework, we have occasionally adopted terminology that differs from that used by the original authors. While this may initially lead to some

confusion, we intend to mitigate this by providing transparent and detailed descriptions. Our objective is a consistent application of these concepts across the field of EL research, thereby promoting coherence between different studies. The following sections attempt to align existing research and metrics with the proposed framework. While this alignment has required some linguistic adjustments to existing terminology and procedures, it is important to note that no substantive changes have been made to the underlying methodologies.

## 6.2 Morphology

Morphological metrics aim to evaluate the structure and formation of words within a language, as well as the richness and diversity of its vocabulary. The identified metrics focus on aspects such as language compression, redundancy, and ambiguity. The morphology of a language significantly influences the complexity of language based tasks [250]. Therefore, the evaluation of morphological features is a crucial component for understanding and evaluating the effectiveness of ELs.

### 6.2.1 Compression

The concept of compression within a language refers to its ability to efficiently combine and reuse a limited set of characters to generate a large collection of words or meanings [105, 224]. Several metrics can be used to quantify compression in ELs. A straightforward approach for these metrics is to use statistical measures, as shown in the following paragraphs. These metrics provide insight into the efficiency of the language, indicating how well it minimizes redundancy while maximizing expressiveness. Efficient compression is a key indicator of a communication system, especially in scenarios where resources (such as memory or bandwidth) are constrained.

#### *Distinct appearances*

The metric of distinct appearances ( $DA$ ) was proposed by Loreto et al. [224]. It is formalized in Eq. 5 and designed to quantify the capacity of a communication system to name a diverse set of objects or categories using its available symbols [224]. Specifically, this metric evaluates how frequently characters  $v \in Y$  are reused across different words or names  $w$  within the lexicon  $W$ . By examining the set  $W_v$ , which includes all words containing a given character  $v$ , we can assess the system's flexibility in recombining basic units to generate a broad spectrum of expressions.

A high  $DA$  value, approaching 1, indicates that the characters are highly versatile and reused extensively across different words, thereby reflecting a flexible communication system. Conversely, a low  $DA$  value suggests limited reuse of characters, which may imply constraints in the system's expressiveness or a less efficient use of its symbolic resources. This metric provides insights into how efficiently a system can balance the trade-off between a compact character set and the richness of its vocabulary.

$$DA = \frac{\sum_{v \in Y} (|W_v| - 1)}{(|W| - 1)|Y|} \quad \text{with} \quad W_v = \{w \mid v \in w \wedge w \in W\} \quad (5)$$

#### *Average message length*

Another way to assess the degree of compression achieved by agents in their communication is to analyze the average message length [72, 114, 142, 163, 169, 193]. This metric, which appears for the first time in Choi et al. [72], captures the typical length of generated messages and provides insight into the efficiency of the EL in terms of information density



[114]. By tracking the average number of words in the messages, we can quantify how effectively the agents compress their language. This metric is computed at the word level, meaning each word within a message is counted. The average message length  $\overline{|m|}$  for a set of messages  $M$  is calculated as follows:

$$\overline{|m|} = \frac{1}{|M|} \sum_{m \in M} |m| \quad \text{with} \quad |m| = \sum_{w \in m} 1 \quad (6)$$

### Active words

The active words metric, introduced by Lazaridou et al. [14], complements the average message length by quantifying the diversity of word usage within the vocabulary [114]. Specifically, this metric measures the variety and utilization of distinct words in a communication system. A high number of active words indicates a diverse vocabulary, reflecting a more complex or redundant EL. Conversely, a lower number suggests that the communication system relies on a limited set of words, which may indicate a more efficient and compressed language with less synonyms [18]. This metric is widely used in the literature [14, 18, 32, 114, 116, 119, 137, 140, 142, 144, 151, 187]. Mathematically, the active word value  $AW$  for an agent  $\xi_i$  can be defined as the size of the collection of words actively used by the agent  $W_{\xi_i}$ , as given in Eq. 7. In multi-agent setups, this metric can be averaged across all agents to provide a collective measure of vocabulary diversity within the joint system.

$$AW(\xi_i) = |W_{\xi_i}| \quad \text{with} \quad W_{\xi_i} \subset W \quad (7)$$

## 6.2.2 Redundancy or ambiguity

Redundancy in language occurs when multiple words are associated with the same meaning, providing alternative expressions for the same concept. Conversely, ambiguity occurs when a single word is associated with multiple meanings, creating the potential for different interpretations depending on the context. Both redundancy and ambiguity are characteristic features of NLS, reflecting the complexity and flexibility inherent in human communication [114, 132].

### Perplexity

Perplexity, introduced by Havrylov and Titov [15], measures how often a word was used in a message to describe the same object [72, 114]. “A lower perplexity shows that the same [words] are consistently used to describe the same objects.” [114]. Mathematically,  $P(w|\varphi)$  represents the probability or score of a word for a specific concept or meaning, e.g., derived from an affine transformation of the sender’s hidden state [114] or from a ground truth label [72]. Thus, perplexity, given in Eq. 8, quantifies the predictability of word usage, with lower values reflecting a less redundant communication system. It is usually calculated based on a sampled set of meanings  $\Phi_{\text{test}}$  for which the word probability can be generated.

$$Ppl = \exp \left( - \sum_{w \in W} [P(w|\varphi) \cdot \log(P(w|\varphi))] \right) \quad \forall \varphi \in \Phi_{\text{test}} \subseteq \Phi \quad (8)$$

### Singular value decomposition

Another approach to quantitatively assess the redundancy of the vocabulary used in a communication system is outlined by Lazaridou et al. [14]. This method involves constructing a matrix where the rows correspond to distinct meanings, the columns represent individual words, and the matrix entries indicate the frequency with which each word is

used for a given meaning. The rows are thus constructed based on a predefined ground truth classification. By applying Singular Value Decomposition (SVD) to this matrix, we can examine the dimensionality of the underlying communication strategy. If the communication system relies on a limited set of highly synonymous words, we would expect the SVD to reveal a low-dimensional structure. Conversely, a higher-dimensional decomposition would indicate a more diverse use of vocabulary, reflecting a potentially less synonymous and more redundant language.

### **Message distinctness**

Message distinctness evaluates the linguistic representation of distinct features and thus aims to quantify ambiguity [72, 114, 168, 225]. The metric, first suggested in Lazaridou et al. [168] and Choi et al. [72], quantifies the diversity of messages generated by the agent by assessing how well it differentiates between various inputs. Specifically, message distinctness  $MD$  is calculated as the ratio of the number of unique messages generated within a batch (cf. 9) to the batch size (cf. 10). A higher message distinctness indicates less ambiguity of the language.

$$M_{\text{unique}} = \{m_i \mid m_i \in M_{\text{test}} \wedge m_i \neq m_j \forall m_j \in M_{\text{test}}, i \neq j\} \quad (9)$$

$$MD = \frac{|M_{\text{unique}}|}{|M_{\text{test}}|} \quad (10)$$

## **6.3 Syntax**

Despite the significance of structural properties in ELs, particularly regarding their syntax and its relation to semantics, research in this area remains limited [245]. Recurrent syntactical patterns are central to the robustness and versatility of NLs [50]. Exploring these properties within the context of EL could provide valuable insights into their development and alignment with NL.

### **6.3.1 Syntax tree**

Van der Wal et al. [193] introduced unsupervised grammar induction (UGI) techniques for syntax analysis in EL research, describing a two-stage approach to deriving grammar and syntax. The first phase involves the induction of unlabeled constituent tree structures, explained below, and the labeling of these structures. The second phase extracts a probabilistic context-free grammar (PCFG) from the labeled data. Two methods were compared for constituency structure induction: the Common Cover Link (CCL), a pre-neural statistical parser that makes assumptions about NL such as the Zipfian distribution, and the Deep Inside-Outside Recursive Auto-encoder (DIORA), a neural parser. For the labeling process, Van der Wal et al. [193] used Bayesian Model Merging (BMM), to consolidate probabilistic models to label the induced syntax trees.

In syntax trees, the structure of the language is represented in a hierarchical manner, where nodes represent grammatical constructs (such as sentences, phrases, and words) and edges represent the rules or relationships that connect these constructs. Analysis of these trees helps to understand how well grammar induction methods match the true syntactic nature of ELs. There are several metrics associated with syntax trees that are used to measure the complexity of the grammar [193]. First, tree depth measures the maximum

distance from the root of the tree to its deepest leaf. Tree depth reflects the hierarchical complexity of the grammar. Shallow trees indicate a simpler grammar, while deeper trees suggest a more complex syntactic structure. Second, the number of unique preterminal groups is a metric that counts the different sets of preterminals (intermediate symbols) that appear to the right of production rules in a grammar. A larger number of unique preterminal groups indicates a richer and more diverse syntactic organization, suggesting that the grammar can generate a greater variety of structures.

### 6.3.2 Categorical grammar induction

Ueda et al. [102] proposed a novel approach for analyzing the syntactic structure of ELs using Categorical Grammar Induction (CGI). This technique focuses on deriving categorical grammars from message-meaning pairs, making it particularly well-suited for simple referential or signaling games.

In this method, derivation trees are constructed using lexical entries and application rules, mapping messages to atomic syntactical representations. Given that multiple derivations might exist for a single message, “the most likely derivation [is selected] using a log-linear model” [102]. CGI is particularly valuable for assessing the syntactic structure of an EL using the generated trees.

## 6.4 Semantics

Capturing the semantic properties of ELs is inherently complex, making it difficult to encapsulate nuances in a single metric. To address this, several key features have been introduced, including grounding, compositionality, consistency, and generalization. These are important because agents can develop representations that are well aligned with task performance but fail to capture the underlying conceptual properties [16]. Thus, an EL might enable successful task completion without truly encoding semantic meaning. Therefore, evaluating these semantic features is essential to evaluate the value and validity of the EL.

### 6.4.1 Grounding

Grounding is essential for the development of meaning and for systematic generalization to novel combinations of concepts [123]. It forms the basis of human-agent communication [33], and without proper grounding, meaningful communication cannot be effectively learned [112]. However, in general dialog settings, grounding does not emerge naturally without specific regularization techniques [127]. The grounding problem, which concerns how words acquire semantic meaning, is central to this challenge [184]. Thus, grounding metrics are vital as they largely define the usability of a language. However, a significant limitation of these metrics is their reliance on some form of oracle or a NL-grounded precursor [15, 26, 134, 207].

#### *Divergence*

Havrylov and Titov [15] proposed a weak form of grounding. Weak grounding means that the same word can correspond to completely different concepts in the induced EL and NL. They used the Kullback–Leibler divergence  $D_{KL}$  (cf. Equation 11) of an EL and a NL distribution to ensure that the statistical properties of EL messages resemble those of NL. They introduced this approach as an indirect supervision measure during training but it can also serve as a metric for evaluating the alignment between EL and NL. For a given sample

$k$  and the message  $m_{\xi_s}$  produced by the sender, the grounding divergence  $G_{Div}$  calculation is shown in Eq. 12. Since the true NL distribution  $P_{NL}(m_{\xi_s})$  is inaccessible, a language model is trained to approximate this distribution. The KL divergence yields a value in the range  $[0, \infty)$ , with lower values indicating a closer resemblance between the generated messages and NL.

$$D_{KL}(P||Q) = \sum_x P(x) \log \left( \frac{P(x)}{Q(x)} \right) \quad (11)$$

$$G_{Div} = D_{KL}(P(m_{\xi_s} | k) || P_{NL}(m_{\xi_s})) \quad (12)$$

### Purity

Purity, proposed by Lazaridou et al. [14], is a metric used to assess the alignment between predefined semantic categories and those observed in an EL. It measures the effectiveness of a communication system in consistently mapping signals or words to specific concepts [198]. Thus, purity quantifies the extent to which the clustering of words reflects meaningful and coherent categories, as determined by ground-truth labels. To assess purity, we first form clusters by grouping samples based on the most frequently activated words to describe them. The quality of these clusters is then evaluated using the purity metric, which calculates the proportion of labels in each cluster that match the majority category of that cluster. A higher purity score indicates that the sender is producing words that are semantically aligned with predefined categories, as opposed to arbitrary or agnostic symbol usage, as demonstrated in [14]. However, this metric requires the existence of predefined ground-truth labels, limiting its applicability in scenarios where such labels are unavailable or ambiguous.

Formally, given a set of clusters  $\{C_k\}$  where each cluster of samples  $C_k$  has a corresponding majority ground-truth label  $c_k$ , the purity of a cluster  $C_k$  is defined as:

$$\text{purity}(C_k) = \frac{|\{w_c \mid w_c \in C_k \wedge w_c = c_k\}|}{|\{w \mid w \in C_k\}|} \quad (13)$$

Here,  $\{w \mid w \in C_k\}$  is the collection of all words used to describe the samples in the cluster and  $\{w_c \mid w_c \in C_k \wedge w_c = c_k\}$  is the collection of words within the cluster that fit the majority label of that cluster. The purity metric ranges from 0 to 1, where a value of 1 indicates perfect alignment with the ground-truth categories.

### Representational similarity analysis

Representational Similarity Analysis (RSA) emerged in the field of neuroscience and was proposed by Kriegeskorte et al. [251]. It has since been adapted for the evaluation of the similarity of neural representations across different modalities, including computational models and brain activity patterns. This technique has been effectively applied in EL research [16, 114, 216], where the focus shifts from analyzing neural activity to exploring the structural relationships between different embedding spaces. For example, RSA has been employed to compare the similarity of embedding space structures between input, sender, and receiver in a referential game [16, 114]. By calculating pairwise cosine similarities within these spaces and then computing the Spearman correlation between the resulting similarity vectors, we can calculate an RSA score that measures the global agreement between these spaces, independent of their dimensionality. The agreement of an agent's embedding space with the input embedding space as such provides an intuitive measure of the grounding of the EL.

This approach offers the advantage of being applicable to heterogeneous agents and arbitrary input spaces. In our framework, this corresponds to any ground truth structured embedding  $e(o_\xi)$  of an agent's observation  $o_\xi$  and its internal meaning representation  $\varphi_\xi$ . Nevertheless, a significant limitation is the necessity for an embedding, which provides a structured description of the observation oriented towards a ground truth, for example, based on a NL model. Furthermore, RSA is not directly applicable to the language itself, particularly for discrete languages. Instead, it operates at the level of earlier meaning representations. Despite this, RSA provides valuable insights into whether the EL can be grounded by evaluating the grounding of the meaning space.

The methodology of [16] utilizes a collection  $K$  of samples, comprising  $k$  observations, images, or feature vectors, to compute representational similarities between input and meaning space. First, we generate input or ground truth embeddings  $e_{GT} = e(o_\xi)$  using an appropriate model and generate the corresponding internal representations  $\varphi_\xi$  from the appropriate architecture part of agent  $\xi$ . Next, we compute pairwise similarities within each embedding space, denoted as  $S_e$  for the ground truth embeddings and  $S_\varphi$  for the agent representations, typically using cosine similarity  $S_{\cos}$  as defined in Eq. 14. This yields a similarity vector of size  $N \cdot (N - 1)$  for each embedding space. The vectors are converted into rank vectors  $R(S_e)$  and  $R(S_\varphi)$ . Finally, we calculate the Spearman rank correlation  $\rho$  [252] between the ranked similarity vectors, using the covariance  $cov$  and standard deviation  $\sigma$ , to assess the alignment between the input and agent representation spaces (cf. Equation 15). The correlation coefficient  $\rho$  takes on values between  $-1$  and  $1$ . A high absolute value of this coefficient indicates a strong alignment between the two variables.

$$S_e = S_{\cos}(e_i, \varphi_i) \quad \forall i, j \in k, i \neq j \quad \text{with} \quad S_{\cos}(a, b) = \frac{a \cdot b}{||a|| \cdot ||b||} \quad (14)$$

$$\rho = \frac{cov(R(S_e), R(S_\varphi))}{\sigma_{R(S_e)} \sigma_{R(S_\varphi)}} \quad (15)$$

## 6.4.2 Compositionality

In EL research, achieving compositionality often requires deliberate guidance, as it does not naturally arise without specific interventions [119]. For instance, training models on diverse tasks and varying environmental configurations can facilitate the development of compositional structures. This occurs as atomic concepts, learned in simpler contexts, are recombined in more complex scenarios [18]. When a language is truly compositional, its components can be systematically rearranged or substituted with conceptually equivalent components without altering the overall meaning [121, 122].

The formalization of compositionality can be framed using the comprehension  $\mathcal{L}_{\text{comp}}$  or production  $\mathcal{L}_{\text{prod}}$  function that map expressions from a language  $\mathcal{L}$  to a space of meanings  $\Phi$  or vice versa [138]. For example, the function  $\mathcal{L}_{\text{comp}} : \mathcal{L} \rightarrow \Phi$  reflects “all the things that the language can denote” [138]. A language is compositional if these functions act as a homomorphism, e.g., there exist binary operators  $\circ$  on  $\mathcal{L}_{\text{comp}}$  and  $\times$  on  $\Phi$  such that for any expression composed of two constituents  $m_1$  and  $m_2$  in  $\mathcal{L}$ , the following condition holds:

$$\mathcal{L}_{\text{comp}}(m_1 \circ m_2) = \mathcal{L}_{\text{comp}}(m_1) \times \mathcal{L}_{\text{comp}}(m_2) \quad (16)$$

### Topographic similarity

Topographic similarity (topsim), originally proposed by Brighton and Kirby [253] and first applied to EL by Lazaridou et al. [168], is a metric designed to quantify the structural alignment between the internal representations of meanings and the corresponding generated messages in a communication system. Unlike RSA (cf. Section 6.4.1), which compares the meaning space against a ground truth, topsim focuses on the internal alignment within an agent's meaning and message spaces. "The intuition behind this measure is that semantically similar objects should have similar messages" [168]. It has become a widely used metric in the study of EL, as depicted in Fig. 12 in Appendix B).

To compute topsim, we start by sampling  $k$  meaning representations denoted by  $\varphi$ , typically embedded feature vectors, from the meaning space  $\Phi$ . Let  $\phi = \{\varphi_1, \dots, \varphi_k\}$  denote the collection of these samples, with  $\varphi \in \Phi$ . Using the sender's policy  $\pi_{\xi_s}^M$ , we generate corresponding messages  $m_i = \pi_{\xi_s}^M(\varphi_i)$  for each sample  $\varphi_i \in \phi$ . We then compute distances within the meaning and language spaces using suitable distance functions for language  $\Delta_{\mathcal{L}}$  and meaning  $\Delta_{\Phi}$  space. The choice of distance function  $\Delta$  depends on the nature of the spaces involved. For discrete communication, typical choices include Hamming [254] or Levenshtein [255] distance, whereas for continuous spaces, cosine or Euclidean distance are often used [51]. Finally, we compute the Spearman rank correlation  $\rho$  [252] using the ranked distances to get the topsim value of the language:

$$\rho = \frac{\text{cov}(R(\Delta_{\mathcal{L}}(m_i)), R(\Delta_{\Phi}(\varphi_i)))}{\sigma_{R(\Delta_{\mathcal{L}}(m_i))} \sigma_{R(\Delta_{\Phi}(\varphi_i))}} \quad \forall \varphi_i \in \phi \quad (17)$$

### Positional disentanglement

Positional Disentanglement (posdis) was introduced by Chaabouni et al. [122] as a metric to evaluate the extent to which words in specific positions within a message uniquely correspond to particular attributes of the input. This metric operates on an order-dependent strategy, which is normalized by the message length and calculated as the ratio of mutual information to entropy. The underlying assumption is that the language leverages positional information to disambiguate words, such that "each position of the message should only be informative about a single attribute" [122]. Thus, "posdis assumes a message whose length equals the number of attributes in the input object, and where each message token, in a specific position, represents a single attribute" [180]. This order-dependence is a characteristic feature of NL structures and is essential for the emergence of sophisticated syntactic patterns [122].

The metric begins by identifying each word  $w_p$  at position  $p$  in a message  $m$ , where  $f$  represents the feature vector of the ground truth. The mutual information  $I(w_p, f_i)$  between  $w_p$  and a specific feature  $f_i$  is calculated to determine how informative the position  $p$  is about the attribute  $f_i$  (cf. Equation 18). The two most informative features  $f_i^1$  and  $f_i^2$  are then identified based on the mutual information value (cf. Equation 19). To quantify positional disentanglement, the mutual information difference between the two most informative features is normalized by the entropy  $H(w_p)$  of the word at position  $p$ , as defined in Eq. 20 and Eq. 21. Finally, the overall posdis value for a language is calculated by averaging the posdis scores across all positions in the messages within the dataset. For messages of varying lengths, the posdis score is normalized by the average message length  $\overline{|m|}$ , as given

in Eq. 22.

$$I(w_p, f_i) = \sum_{w_p \in m} \sum_{f_i \in f} P(w_p, f_i) \log \left( \frac{P(w_p, f_i)}{P(w_p)P(f_i)} \right) \quad (18)$$

$$f_i^1 = \operatorname{argmax}_{f_i \in f} I(w_p, f_i) \quad \text{and} \quad f_i^2 = \operatorname{argmax}_{f_i \in f \wedge f_i \neq f_i^1} I(w_p, f_i) \quad (19)$$

$$H(w_p) = - \sum_{w_p \in m} P(w_p) \log(P(w_p)) \quad (20)$$

$$\text{posdis}_p = \frac{I(w_p, f_i^1) - I(w_p, f_i^2)}{H(w_p)} \quad (21)$$

$$\text{posdis} = \frac{1}{|m|} \sum_p \text{posdis}_p \quad \text{with} \quad |\overline{m}| = \frac{1}{|M|} \sum_{m \in M} |m| \quad (22)$$

### Bag of symbols disentanglement

Bag of Symbols Disentanglement (bosdis) is a metric introduced by Chaabouni et al. [122] to assess the degree to which words in a language unambiguously correspond to different input elements, regardless of their position within a message. While positional disentanglement (posdis) relies on the assumption that positional information is crucial for disambiguating words (cf. Section 6.4.2), bosdis relaxes this assumption and captures the intuition behind a permutation-invariant language. In such a language, the order of words is irrelevant, and only the frequency of words carries meaning [122]. The metric normalizes the mutual information between symbols and input features by the entropy summed over the entire vocabulary. This approach maintains the requirement that each symbol uniquely refers to a distinct meaning, but shifts the focus to symbol counts as the primary informative element.

$$I(w, f_i) = \sum_{w \in m} \sum_{f_i \in f} P(w, f_i) \log \left( \frac{P(w, f_i)}{P(w)P(f_i)} \right) \quad (23)$$

$$f_i^1 = \operatorname{argmax}_{f_i \in f} I(w, f_i) \quad \text{and} \quad f_i^2 = \operatorname{argmax}_{f_i \in f \wedge f_i \neq f_i^1} I(w, f_i) \quad (24)$$

$$H(w) = - \sum_{w \in m} P(w) \log(P(w)) \quad (25)$$

$$\text{bosdis}_w = \frac{I(w, f_i^1) - I(w, f_i^2)}{H(w)} \quad (26)$$



$$bosdis = \frac{1}{|W|} \sum_{w \in W} bosdis_w \quad (27)$$

### Tree reconstruct error

Tree Reconstruct Error (TRE) assumes prior knowledge of the compositional structure within the input data, enabling the construction of tree-structured derivations [134]. As defined by Andreas [134], a language is considered compositional if it functions as a homomorphism from inputs to their representations. The compositionality of a language should be evaluated by identifying representations that allow an explicitly compositional language to closely approximate the true underlying structure [134]. One metric for this assessment is TRE, which quantifies the discrepancy between a compositional approximation and the actual structure, using a composition function and a distance metric. A TRE value of zero indicates perfect reproduction of compositionality.

The compositional nature of a sender's language is affirmed if there exists an assignment of representations to predefined primitives (e.g., categories, concepts, or words) such that for each input, the composition of primitive representations according to the oracle's derivation precisely reproduces the sender's prediction [134]. TRE specifically measures the accuracy with which a given communication protocol can be reconstructed while adhering to the compositional structure of the derivation or embedding of the input  $e \in E$  [51].

One of the key advantages of the TRE framework is its flexibility across different settings, whether discrete or continuous. It allows for various choices of compositionality functions, distance metrics, and other parameters. However, this flexibility comes with challenges, including the requirement for an oracle-provided ground truth and the necessity of pre-trained continuous embeddings.

It is defined in a way that allows the choice of the distance metric  $\delta$  and the compositionality function  $\circ$  to be determined by the evaluator [134]. When the exact form of the compositionality function is not known a priori, it is common to define  $\circ$  with free parameters, as suggested by Andreas [134], treating these parameters as part of the learned model and optimizing them jointly with the other parameters  $\eta$ . However, care must be taken when learning the compositional function to avoid degenerate solutions [134].

Given a data sample  $k$  from the dataset  $K$  ( $k \in K$ ) and a corresponding message  $m$  from the set of all possible messages  $M$  ( $m \in M$ ), TRE requires a distance function  $\delta$  and learnable parameters  $\eta$ . Additionally, it employs a compositionality function  $\circ$  and pre-trained embeddings of ground truth, denoted by  $e \in E$ , which can be obtained using models like word2vec.

The functions involved in the TRE calculation are as follows:

- Pre-trained ground truth oracle (e.g., word2vec):  $\mathcal{E} : K \rightarrow E$
- Learned language speaker:  $\xi_S : K \rightarrow M$
- Learnable approximation function for TRE:  $\tilde{f}_\eta : E \rightarrow M$

In the discrete message setting, which is the focus here, a discrete distance metric such as  $L_1$  is typically chosen, along with a compositionality function  $\circ$  defined by a weighted linear combination [51, 134]:

$$m_1 \circ m_2 = Am_1 + Bm_2 \quad \text{with} \quad \eta = \{A, B\} \quad (28)$$

To compute the TRE, an optimized approximation function  $\tilde{f}_\eta$  is required. This function

must satisfy two key properties: embedding consistency, meaning that the learned parameters  $\eta$  are specific to an embedding, and compositionality, which ensures that the function behaves according to:

$$\tilde{f}_\eta(e_i) = \eta_i \quad \text{and} \quad \tilde{f}_\eta(\langle e_i, e_j \rangle) = \tilde{f}_\eta(e_i) \circ \tilde{f}_\eta(e_j)$$

The optimization process involves minimizing the distance between the output of the learned language speaker  $\xi_S(k_i)$  and the approximation function  $\tilde{f}_\eta(e_i)$ , based on the ground truth:

$$\eta^* = \underset{\eta}{\operatorname{argmin}} \sum_i \delta(\xi_S(k_i), \tilde{f}_\eta(e_i)) \quad \text{with} \quad \mathcal{E}(k_i) = e_i \quad (29)$$

With the optimized parameters  $\eta^*$ , TRE can be calculated at two levels: the datum level, which assesses individual instances:

$$TRE(k_i) = \delta(\xi_S(k_i), \tilde{f}_{\eta^*}(e_i)) \quad \text{with} \quad \mathcal{E}(k_i) = e_i \quad (30)$$

and the dataset level, which measures the overall communication performance across the dataset:

$$TRE(K) = \frac{1}{|K|} \sum_{k \in K} TRE(k) \quad (31)$$

### Conflict count

Conflict count, introduced by Kuciński et al. [166], is designed to quantify the extent to which the assignment of features to words in a language deviates from the word's principal meaning. This metric is particularly useful in scenarios where the language employs synonyms, as it accounts for the possibility of multiple words referring to the same concept.

The conflict count metric operates under the assumption that the number of concepts or features  $f_i$  given in a feature vector  $f$  of a sample  $k$  in the collection of samples  $K$  is equal to the message length  $|m|$ , and that there exists a one-to-one mapping between a concept  $f_i \in f$  and a word  $w \in W$ . The metric counts how frequently this one-to-one mapping is violated, with a value of 0 indicating no conflicts and, therefore, high compositionality. An advantage of this metric is its ability to accommodate redundancy in the language. However, it also has limitations, such as the assumption that the number of features or attributes equals the message length, i.e.,  $|f| = |m|$ . Additionally, because conflict count assumes the number of concepts in a derivation to be equal to the message length, it becomes undefined for languages or protocols that violate this assumption, such as those involving negation or context-sensitive constructions presented in [51].

The primary objective of conflict count is to quantify the number of times the mapping from a word  $w$  to its principal meaning  $\varphi_w$  is violated. This requires the assumption that a mapping  $\alpha$  exists from the position  $p$  of word  $w$  in message  $m$  to an individual feature in feature vector  $f$ , such that:

$$\alpha = \{1, \dots, |m|\} \rightarrow \{1, \dots, |f|\} \quad (32)$$

In this framework, the meaning of a word, denoted by  $\varphi_w$ , is determined by both the word  $w$  itself and its position  $p$  within the message. This meaning corresponds to a specific instance  $j$  of a particular feature  $i$  within the feature vector  $f$ , such that  $f_{ij} = \varphi(w, p)$ .

The process of calculating the conflict count begins by identifying the principal meaning of each word-position pair:

$$\varphi(w, p : \alpha) = \operatorname{argmax}_{f_{ij}} \operatorname{count}(w, p, f_{ij} : \alpha) \quad \forall f_{ij} \in f \quad (33)$$

using the count function:

$$\operatorname{count}(w, p, f_{ij} : \alpha) = \sum_{k \in K} |\{w \mid w \in m(k) \wedge \operatorname{pos}_m(w) = p \wedge f_{ij} \in k\}| \quad (34)$$

where  $m(k)$  is the message produced for sample  $k$  and  $\operatorname{pos}_m(w)$  computes the position of word  $w$  in message  $m$ .

Finally, the conflict count value  $\operatorname{conf}$  is determined by finding the mapping  $\alpha$  that minimizes the score:

$$\operatorname{conf} = \operatorname{argmin}_{\alpha} \sum_{w,p} \operatorname{score}(w, p : \alpha) \quad (35)$$

where the score function is defined as:

$$\operatorname{score}(w, p : \alpha) = \sum_{f_{ij} \neq \varphi(w,p)} \operatorname{count}(w, p, f_{ij} : \alpha) \quad (36)$$

### 6.4.3 Consistency

For a language to be effective, the meaning of each word must be consistent across different contexts. Inconsistent word meanings can render a language practically useless, even if the language is semantically grounded and exhibits compositional properties [127]. In dialogue settings, particularly in the absence of explicit regularization mechanisms, words often fail to maintain consistent groundings across different instances, leading to ambiguity and reduced communicative effectiveness [127]. Thus, it is crucial to carefully monitor this language characteristic in EL settings.

#### **Mutual information**

Consistency in language can be quantitatively assessed by examining the mutual information between messages and their corresponding input features. Ideally, a consistent language will exhibit a high degree of overlap between messages and features, leading to a high mutual information value, indicating strong correspondence [150].

Formally, mutual information between two random variables, say  $X$  and  $Y$ , with joint distribution  $P_{(X,Y)}$  and marginal distributions  $P_X$  and  $P_Y$ , is defined as the Kullback–Leibler divergence  $D_{KL}$  (see Eq. 11) between the joint distribution and the product of the marginals:

$$I(X; Y) = D_{KL}(P_{(X,Y)} \parallel P_X \otimes P_Y) \quad (37)$$

In the context of discrete communication, where both messages and sample features are represented as discrete variables, the mutual information between the set of messages  $M$  and the set of features  $F$  is computed using a double summation over all possible message-feature pairs:

$$I(M; F) = \sum_{m \in M} \sum_{f \in F} P_{(M,F)}(m, f) \log \left( \frac{P_{(M,F)}(m, f)}{P_M(m)P_F(f)} \right) \quad (38)$$

where  $P_{(M,F)}(m, f)$  is the joint probability of message  $m$  and feature  $f$ , and  $P_M(m)$  and  $P_F(f)$  are the marginal probabilities of  $m$  and  $f$ , respectively.

### Correlation

Various studies employ different statistical techniques to measure consistency using correlations [114, 116, 173, 176, 184, 195]. For example, consistency within a language system can be quantified by analyzing the variability of words produced for a given sample  $k$ . Specifically, given the set of all words representing  $k$ , a heatmap is generated using the mean of this set. The sharpness of the heatmap is then quantified by computing the Variance of the Laplacian (VoL). The average consistency score is obtained by dividing the VoL of the heatmap by the count of all samples considered, as introduced by Verma and Dhar [195].

Additionally, Mul et al. [176] explored the correlation between messages and actions as well as between messages and salient properties of the environment. The analysis reveals correlations by examining the conditional probability distribution of actions given the messages produced by a pretrained or fine-tuned receiver. This distribution, denoted as  $P(a | m)$ , was visualized using bin bar plots to highlight the prominent correlations [176]. Similarly, the relationship between input and messages is analyzed by examining the conditional distribution of a pretrained sender's messages given the observational input, represented as  $P(m | o)$  [176].

### Coherence

Coherence is often assessed through context independence, a metric initially proposed by Bogin et al. [33]. Context independence examines whether words within a language maintain consistent semantics across varying contexts. However, context independence may be considered restrictive, particularly in languages where synonyms are prevalent [29, 164]. The context independence metric aims to measure the alignment between words  $w \in W$  and features  $f \in F$  of the input samples by analyzing their probabilistic associations. Specifically,  $P(w | f)$  denotes the probability that a word  $w$  is used when a feature  $f$  is present, while  $P(f | w)$  represents the probability that a feature  $f$  appears when a word  $w$  is used. For each feature  $f$ , we identify the word  $w_f$  most frequently associated with it by maximizing  $P(f | w)$ :

$$w_f := \underset{w}{\operatorname{argmax}} P(f | w) \quad (39)$$

The context independence or coherence metric  $CI$  is then computed as the average product of these probabilities across all features:

$$CI(w_f, f) = \frac{1}{|F|} \sum_{f \in F} P(w_f | f) P(f | w_f) \quad (40)$$

This metric ranges from 0 to 1, with 1 indicating perfect alignment, meaning that each word retains its meaning consistently across different contexts and is thus used coherently.

### Entropy

Entropy metrics are instrumental in analyzing the variability and predictability within linguistic systems. The most fundamental use of entropy involves marginal probabilities, which capture the variability in the number of words in a language [108, 114]. More advanced applications of entropy focus on sender language entropy, which examines the conditional entropy of messages given features and vice versa [32, 118]. Specifically, low conditional entropy  $H(M | F)$  indicates that a unique message is used for a specific feature, whereas high  $H(M | F)$  reflects the generation of synonyms for the same feature [118].

Recent approaches further extend this analysis by combining conditional entropies [37, 177]. For example,  $H(M | F)$  quantifies the uncertainty remaining about messages after knowing the concepts, while  $H(F | M)$  measures the uncertainty about concepts given the messages. A negative correlation between these measures and agent performance is expected [177]. However, a notable limitation of these entropy-based methods is that they focus on complete messages rather than individual words, which can limit the evaluation of more complex languages.

For example, Ohmer et al. [177] provide the following comprehensive evaluation approach. First, the conditional entropy of messages given features  $H(M | F)$ , see Eq. 41, and  $H(F | M)$  are calculated. Additionally, the marginal entropies are calculated using Eq. 42, where  $X$  represents either messages  $M$  or features  $F$ .

$$H(M | F) = - \sum_{m \in M} \sum_{f \in F} P(f, m) \log \left( \frac{P(f, m)}{P(f)} \right) \quad (41)$$

$$H(X) = - \sum_{x \in X} P(x) \log(P(x)) \quad (42)$$

Using these entropies, *consistency*, see Eq. 43, measures how much uncertainty about the message is reduced when the feature is known, with lower values indicating more consistent message usage. *effectiveness*, on the other hand, see Eq. 44, evaluates the reduction in uncertainty about the feature when the message is known, with lower values reflecting more unique messages for individual features.

$$\text{Consistency}(F, M) = 1 - \frac{H(M | F)}{H(M)} \quad (43)$$

$$\text{Effectiveness}(F, M) = 1 - \frac{H(F | M)}{H(F)} \quad (44)$$

Finally, the normalized mutual information  $NI$  provides a combined score:

$$NI(F, M) = \frac{H(M) - H(M | F)}{0.5 \cdot (H(F) + H(M))} \quad (45)$$

A high  $NI$  score indicates a strong predictive relationship between messages and features, reflecting high consistency.

### Similarity

The Jaccard similarity coefficient is a another metric for evaluating the consistency of language usage among agents [72, 116]. It quantifies the similarity between two sets by comparing the size of their intersection to the size of their union [116]. To measure language consistency, the Jaccard similarity is computed by sampling messages for each input and averaging the similarity scores across the population [116]. This approach reflects how consistently words are used across different messages. Specifically, Jaccard similarity  $J(M_{\xi_i}, M_{\xi_j})$  is defined in Eq. 46, where  $M_{\xi_i}$  and  $M_{\xi_j}$  represent sets of messages generated by different agents based on the same input. The similarity ranges from 0 to 1, with 1 indicating complete overlap and thus perfect similarity.

In practice, Jaccard similarity helps to assess the coherence of languages emerging from agent-based systems. For instance, in referential game experiments, high perplexity (cf.

Section 6.2.2) and low Jaccard similarity have been observed, suggesting that agents assign unique but incoherent strings to object types to gain an advantage in the game without producing a consistent language [72]. However, Jaccard similarity is only applicable to scenarios where multiple agents generate messages about the same set of objects. Thus, its application is limited to cases where the goal is to compare the overlap of message sets between agents attempting to convey similar meanings.

$$J(M_{\xi_i}, M_{\xi_j}) = \frac{|M_{\xi_i} \cap M_{\xi_j}|}{|M_{\xi_i} \cup M_{\xi_j}|} = \frac{|M_{\xi_i} \cap M_{\xi_j}|}{|M_{\xi_i}| + |M_{\xi_j}| - |M_{\xi_i} \cap M_{\xi_j}|} \quad (46)$$

#### 6.4.4 Generalization

A language's ability to generalize is crucial for describing objects and concepts at different levels of complexity, allowing for effective clustering and hierarchical representation. Generalization in ELs reflects their ability to extend beyond specific training instances to novel situations. "If the emergent languages can be generalised, we then could say that these languages do capture the structure of meaning spaces" [155]. Research shows that languages capable of generalization tend to emerge only when the input is sufficiently varied [122]. In contrast, a large dictionary size often indicates a lack of generalization [108]. Human languages have evolved under the pressure of a highly complex environment, fostering their generalization capabilities [122]. However, deep learning models often exploit dataset-specific regularities rather than developing systematic solutions [46]. To address this, much research is being done on the systematic generalization abilities of ELs.

##### *Zero shot evaluation*

Zero-shot evaluation, which assesses the ability of an agent to generalize to novel stimuli [72, 168], has become a standard metric in the study of EL as illustrated in Fig. 12 in Appendix B. This evaluation is critical to understand the generalization capabilities of an agent. Zero-shot evaluation can be done in two different scenarios, one with unseen input and the other with an unseen partner.

In the unseen input scenario, models are tested on a zero-shot test set consisting of samples with feature combinations not encountered during training. Performance, such as accuracy, is reported for these unseen samples [114, 122, 127, 164, 168]. Different methods for constructing novel inputs include exposing models to objects that resemble training data but have unseen properties or entirely novel combinations of features [168]. Moreover, a more drastic approach may involve moving to entirely new input scenarios, such as testing the ability of agents to generalize across different game types [37].

The unseen partner scenario, also known as cross-play or zero-shot coordination, evaluates models by pairing agents that did not communicate during training. Again, performance is measured, typically in terms of accuracy [211, 256].

However, these approaches also have drawbacks. The unseen input scenario requires a ground truth oracle to withhold feature combinations, which is necessary to accurately define novel combinations. Meanwhile, the unseen partner setup can introduce inefficiencies by requiring additional resources to train novel communication partners for testing.

##### *Ease and transfer learning*

Ease and Transfer Learning (ETL), as proposed by Chaabouni et al. [25], evaluates how easily new listeners can adapt to an EL on distinct tasks. ETL extends the concept of ease-of-teaching [106] by assessing how effectively a deterministic language, developed by a fixed set of speakers, can be transferred to new listeners who are trained on tasks different

from the original one for which the language was optimized [25]. This metric not only gauges the language's generality but also its transferability across tasks [25].

To measure ETL, after convergence, a fixed number of speakers produce a deterministic language by selecting symbols using an argmax operation over their distributions. This language is then used to train newly initialized listeners on a new task. The training curve is tracked to observe how quickly and accurately the listeners learn the task, which may involve more challenging objectives than the former training tasks [25, 154].

## 6.5 Pragmatics

Pragmatics is a critical aspect of language that examines how context influences meaning [199]. It goes beyond the literal interpretation of words and requires the listener to infer the speaker's intentions, beliefs, and mental states, an ability known as Theory of Mind (ToM) [199]. In human interactions, this contextual reasoning is essential for predicting and understanding behavior. In the context of EL, pragmatics focuses on how effectively agents use the communication ability in their environment. Empirical studies have shown that agents may initially fail to use communication meaningfully, but, once they do communicate, they can reach a locally optimal solution to the communication problem [230]. Thus, evaluating the pragmatics of EL is essential to determining its utility and effectiveness in real-world applications.

### 6.5.1 Predictability

Predictability evaluates the complexity of an environment and its effect on the need for communication. Thus, it is a central metric for the probability of emergence and the use of EL. In simple environments with limited actions, agents can often coordinate without communication [110].

#### *Behavioral divergence*

Behavioral divergence, introduced by Dubova et al. [110], posits that less diversity in actions or messages correlates with more predictable behavior, potentially reducing the need for communication. To quantify this, we calculate Behavioral Action Predictability *BAP* and Behavioral Message Predictability *BMP*. Both use the Jensen-Shannon Divergence (JSD) (see Eq. 47) which itself uses the Kullback–Leibler Divergence  $D_{KL}$  (cf. Equation 11).

$$D_{JS}(P \parallel Q) = \frac{1}{2}D_{KL}(P \parallel M) + \frac{1}{2}D_{KL}(Q \parallel M) \quad \text{with} \quad M = \frac{P+Q}{2} \quad (47)$$

BAP (see Eq. 48) and BMP (see Eq. 49) both use a uniform distribution  $Q$  for comparison. BAP further uses the distribution of actions by the agent  $P(a_\xi)$  while BMP uses the distribution of messages by the agent  $P(m_\xi)$ . Based on that, these metrics provide a robust measure of how predictable agent behaviors and messages are, with higher values indicating less predictability and greater need for beneficial communication [110].

$$BAP = D_{JS}(P(a_\xi) \parallel Q) \quad (48)$$

$$BMP = D_{JS}(P(m_\xi) \parallel Q) \quad (49)$$

### 6.5.2 Efficiency

In EL settings, efficient communication arises only when there is an opportunity cost [115]. Without such a cost, there is no drive towards brevity, which limits the effectiveness and efficiency of EL in HCI.

#### *Sparsity*

Sparsity, as proposed by Kalinowska et al. [115], measures the extent to which agents minimize their communication during task execution. This metric requires only the collection of messages exchanged per episode for computation. However, its applicability is limited to scenarios where communication is not strictly necessary for task completion, i.e., agents have the option to send no messages at all or to send messages that contain no meaningful information. A sparsity value of 0 indicates that an agent can solve the task using only a single message throughout an episode, reflecting a highly efficient communication strategy. Conversely, higher sparsity values indicate more frequent or verbose communication, which may indicate inefficiencies in the EL.

Communication sparsity *ComSpar* is mathematically defined as:

$$ComSpar = \frac{1}{n_{ep}} \cdot \sum_{M_{ep,i} \in M} -\log(|\{m \mid m \in M_{ep,i} \wedge m \neq 0\}|) \quad (50)$$

In this equation,  $M_{ep,i}$  represents the set of all messages exchanged during episode  $i$ , and  $n_{ep}$  is the total number of episodes observed. The collection  $\{m \mid m \in M_{ep,i} \wedge m \neq 0\}$  consists of all messages  $m$  of episode  $i$  that are non-zero and thus contributing.

### 6.5.3 Positive signaling

Positive signaling evaluates the alignment between an agent's observations and its communication output [29]. The goal is to ensure that the outgoing transmitted information is both relevant and observable by the agent [181].

#### *Speaker consistency*

Speaker Consistency (SC), introduced by Jaques et al. [240], measures how effectively an agent's messages reflect its state or trajectory, thereby ensuring the communication is meaningful. This is quantified using mutual information. For an agent  $\xi_i$ , the trajectory  $\tau_{\xi_i}^t$  represents the sequence of states and actions up to time step  $t$ . The message produced at time  $t$  is denoted by  $m_{\xi_i}^t$ . The mutual information  $I(m_{\xi_i}^t, \tau_{\xi_i}^t)$  between the message and trajectory is calculated as:

$$\begin{aligned} I(m_{\xi_i}^t, \tau_{\xi_i}^t) &= H(m_{\xi_i}^t) - H(m_{\xi_i}^t | \tau_{\xi_i}^t) \\ &= - \sum_{m \in M_{\xi_i}} \overline{P_{\xi_i}}(m) \log \overline{P_{\xi_i}}(m) \\ &\quad + \mathbb{E}_{\tau_{\xi_i}^t} \left[ \sum_{m \in M_{\xi_i}} P_{\xi_i}(m | \tau_{\xi_i}^t) \log P_{\xi_i}(m | \tau_{\xi_i}^t) \right] \end{aligned} \quad (51)$$

Here,  $H(m_{\xi_i}^t)$  is the entropy of the message distribution,  $H(m_{\xi_i}^t | \tau_{\xi_i}^t)$  is the conditional entropy given the trajectory,  $\overline{P_{\xi_i}}(m)$  as marginal distribution of message  $m$  over all trajectories, and  $P_{\xi_i}(m | \tau_{\xi_i}^t)$  as conditional distribution of message  $m$  given the trajectory  $\tau_{\xi_i}^t$ . This



way, the mutual information value reflects how much information the message carries about the agent's trajectory.

Lowe et al. [29] built on this concept and provided the following formula for Speaker Consistency (SC):

$$SC = \sum_{a \in A} \sum_{m \in M} P(a, m) \log \frac{P(a, m)}{P(a)P(m)} \quad (52)$$

In this equation,  $P(a, m)$  is the joint probability of action  $a$  and message  $m$ , calculated empirically by averaging their co-occurrences across episodes. In general, SC is a valuable metric for evaluating whether the EL is both informative and aligned with the behavioral patterns of the sender.

### 6.5.4 Positive listening

Positive listening evaluates the effectiveness of how a message receiver utilizes and applies incoming information [29]. However, agents should not simply process messages similarly to other observations to avoid treating them as mere directives [146]. Nevertheless, the metrics presented in this section focus on evaluating the receiver's ability to effectively integrate and use the information received, rather than evaluating the receiver's ability to do more than just follow instructions.

#### *Instantaneous coordination*

Instantaneous Coordination (IC), also referred to as listener consistency [29], was introduced by Jaques et al. [240] as a metric to evaluate how effectively an agent's message influences another agent's subsequent action. IC is computed similarly to Speaker Consistency (cf. Section 6.5.3), but differs in that it measures the mutual information between one agent's message and the other agent's next action, averaged over episodes. This metric directly captures the receiver's immediate reaction to an incoming message, making it a measure of positive listening. However, it primarily captures situations where the receiver's action is directly changed by the sender's message, without considering the broader context or long-term dependencies [29]. Accordingly, "IC can miss many positive listening relationships" [29].

Jaques et al. [240] proposed two specific measures for IC: One that quantifies the mutual information between the sender's message and the receiver's next action (see Eq. 53), and another one that measures the mutual information between the sender's current action and the receiver's next action (see Eq. 54). These measures are calculated by averaging over all trajectory steps and taking the maximum value between any two agents, focusing on short-term dependencies between consecutive timesteps.

$$IC_{m_{\xi_S} \rightarrow a_{\xi_R}} = I(m_k^t; a_j^{t+1}) \quad (53)$$

$$IC_{a_{\xi_S} \rightarrow a_{\xi_R}} = I(a_k^t; a_j^{t+1}) \quad (54)$$

A unified equation for IC is provided by Lowe et al. [29]:

$$SC = \sum_{m_{\xi_S}^t \in M_{\xi_S}} \sum_{a_{\xi_R}^{t+1} \in A_{\xi_R}} P(a_{\xi_R}^{t+1}, m_{\xi_S}^t) \log \frac{P(a_{\xi_R}^{t+1}, m_{\xi_S}^t)}{P(a_{\xi_R}^{t+1})P(m_{\xi_S}^t)} \quad (55)$$

Here,  $P(a_{\xi_R}^{t+1}, m_{\xi_S}^t)$  is the empirical joint probability of the sender's message and the receiver's subsequent action, averaged over episodes within each epoch.

### Message effect

The Message Effect (ME) metric, introduced by Bouchacourt and Baroni [202], quantifies the influence of a message sent by one agent on the subsequent actions and messages of another agent. This metric explicitly considers bidirectional communication, so in the following we use generic agents  $\xi_A$  and  $\xi_B$  instead of sender and receiver. A notable challenge of this metric is the requirement for counterfactual analysis.

Given an agent  $\xi_A$  at timestep  $t$  sending a message  $m_{\xi_A}^t$ , we define  $z_{\xi_B}^{t+1}$  as the combination of the action and message produced by agent  $\xi_B$  at the following timestep. Accordingly, the conditional distribution  $P(z_{\xi_B}^{t+1} | m_{\xi_A}^t)$  represents the response of  $\xi_B$  to the message from  $\xi_A$ . To account for counterfactuals, which encode what might have happened had  $\xi_A$  sent a different message  $\tilde{m}_{\xi_A}^t$ , we define the counterfactual distribution  $\tilde{P}(z_{\xi_B}^{t+1})$  (see Eq. 56).

The ME is then measured by the Kullback–Leibler divergence between the actual response and the counterfactual response (see Eq. 57). The computation involves sampling  $z_{\xi_B}^{t+1,k}$  from the conditional distribution for the actual message and sampling counterfactuals  $\tilde{m}_{\xi_A}^t$  to estimate  $\tilde{P}(z_{\xi_B}^{t+1,k})$  (see Eq. 58). The final ME is calculated as the average KL divergence over the collection of samples  $K$  (see Eq. 59).

$$\tilde{P}(z_{\xi_B}^{t+1}) = \sum_{\tilde{m}_{\xi_A}^t} P(z_{\xi_B}^{t+1} | \tilde{m}_{\xi_A}^t) \tilde{P}(\tilde{m}_{\xi_A}^t) \quad (56)$$

$$ME_{\xi_A \rightarrow \xi_B}^t = D_{KL}(P(z_{\xi_B}^{t+1} | m_{\xi_A}^t) \parallel \tilde{P}(z_{\xi_B}^{t+1})) \quad (57)$$

$$\tilde{P}(z_{\xi_B}^{t+1,k}) = \sum_{j=1}^J P(z_{\xi_B}^{t+1,k} | \tilde{m}_{\xi_A}^t) \tilde{P}(\tilde{m}_{\xi_A}^t) \quad (58)$$

$$ME_{\xi_A \rightarrow \xi_B}^t = \frac{1}{|K|} \sum_{k \in K} \log \frac{P(z_{\xi_B}^{t+1,k} | m_{\xi_A}^t)}{\tilde{P}(z_{\xi_B}^{t+1,k})} \quad (59)$$

### Causal influence of communication

The Causal Influence of Communication (CIC) metric, introduced independently by Jaques et al. [240] and Lowe et al. [29], provides a direct measure of positive listening by quantifying the causal effect that one agent's message has on another agent's behavior. Traditional methods of evaluating communication often fall short, as simply testing for a decrease in reward after removing the communication channel does not adequately capture the utility of communication [29].

CIC is computed using the mutual information between an agent's message and the subsequent action of the receiving agent. Unlike Instantaneous Coordination (cf. Section 6.5.4), CIC considers the probabilities  $P(a, m) = \pi_{\xi_R}(a | m) \pi_{\xi_S}(m)$  that represent changes in the action distribution of the receiver  $\xi_R$  when the message  $m$  from the sender  $\xi_S$

is altered. These probabilities are normalized within each game to accurately reflect the influence of messages on actions within the same context [29].

For multi-time-step causal influence, the CIC metric is defined as the difference between the entropy of the receiver's actions with and without communication:

$$CIC(\tau_{\xi_R}) = H(a_{\xi_R}^t | \tau_{\xi_R}) - H(a_{\xi_R}^t | \tau_{\xi_R}^{+M}) \quad (60)$$

Here,  $\tau_{\xi_R}$  denotes the standard trajectory of the receiver, comprising state-action pairs, while  $\tau_{\xi_R}^{+M}$  includes the communicated messages. The CIC is estimated by learning an approximate policy function  $\pi(\cdot | \tau_{\xi_R})$ . For more details on the multistep version, refer to Eccles et al. [230], and for the single-step version, see Jaques et al. [240].

### 6.5.5 Symmetry

Symmetry in EL refers to consistent language use across agents in settings, where agents alternate between roles such as message sender and receiver [105, 110]. Thus, symmetry ensures convergence to a common language rather than distinct dialects [110].

#### Inter-agent divergence

Inter-Agent Divergence (IAD), introduced by Dubova et al. [110, 257], quantifies the similarity in how different agents map messages to actions. Let  $a_{\xi_i}$  denote the action of agent  $\xi_i$ . The first step involves computing the marginal action distributions for each agent given a message  $m$ , represented as  $P(a_{\xi_i} | m)$ .

$$P(a_{\xi_i} | m) \quad \forall \xi_i \in \xi \wedge m \in M \quad (61)$$

The divergence between two agents,  $\xi_i$  and  $\xi_j$ , based on their responses to the same message, is then calculated using the Jensen-Shannon Divergence (JSD) as follows:

$$\begin{aligned} D_{JS}(\xi_i, \xi_j, m) &= D_{JS}(P(a_{\xi_i} | m) \| P(a_{\xi_j} | m)) \\ &= \frac{1}{2} \left[ D_{KL}(P(a_{\xi_i} | m) \| M) + D_{KL}(P(a_{\xi_j} | m) \| M) \right] \\ \text{where } M &= \frac{P(a_{\xi_i} | m) + P(a_{\xi_j} | m)}{2} \end{aligned} \quad (62)$$

Finally, the overall IAD is computed by averaging these divergences across all possible agent pairs  $(\xi_i, \xi_j) \in \xi_{\text{comb}}$  and messages  $m \in M$ :

$$IAD = \frac{1}{|\xi_{\text{comb}}|} \frac{1}{|M|} \sum_{(\xi_i, \xi_j) \in \xi_{\text{comb}}} \sum_{m \in M} D_{JS}(\xi_i, \xi_j, m) \quad (63)$$

While IAD effectively captures the consistency of inter-agent communication, it may have limitations when applied to more complex languages where message-level comparisons become difficult.

#### Within-agent divergence

Within-Agent Divergence (WAD), proposed by Dubova et al. [110, 257], measures the consistency of an agent's communication behavior when it changes roles, such as from sender to receiver. This metric captures the internal symmetry in an agent's behavior and is crucial in complex systems where agents can assume different roles within the same

environment. To compute WAD, we again first consider the action distribution  $P(a_{\xi_i} | m)$  for each agent  $\xi_i$  over a set of messages  $m \in M$ . This distribution reflects how an agent's actions are conditioned on receiving or sending a specific message.

$$P(a_{\xi_i} | m) \quad \forall \xi_i \in \xi \wedge m \in M \quad (64)$$

Given this, the Jensen-Shannon Divergence (JSD) is used to assess the divergence between an agent's behavior when acting as a sender  $\xi_{i,S}$  versus as a receiver  $\xi_{i,R}$ :

$$\begin{aligned} D_{JS}(\xi_{i,S}, \xi_{i,R}, m) &= D_{JS}(P(a_{\xi_{i,S}} | m) \parallel P(a_{\xi_{i,R}} | m)) \\ &= \frac{1}{2} [D_{KL}(P(a_{\xi_{i,S}} | m) \parallel Q) + D_{KL}(P(a_{\xi_{i,R}} | m) \parallel Q)] \\ \text{with } Q &= \frac{P(a_{\xi_{i,S}} | m) + P(a_{\xi_{i,R}} | m)}{2} \end{aligned} \quad (65)$$

Finally, the overall WAD is computed by averaging this divergence across all agents  $\xi_i \in \xi$  based on the WAD for individual agents and their messages  $m \in M_{\xi_i}$ :

$$WAD = \frac{1}{|\xi|} \sum_{\xi_i \in \xi} \frac{1}{|M_{\xi_i}|} \sum_{m \in M_{\xi_i}} D_{JS}(\xi_{i,S}, \xi_{i,R}, m) \quad (66)$$

## 6.6 Summary of the metrics

While some EL features are quantifiable by multiple metrics and have been investigated in multiple studies, others remain underexplored, as illustrated in Fig. 12 in Appendix B. Metrics such as *topographic similarity* and *zero shot evaluation*, both of which assess semantic properties, are well established and widely utilized across multiple studies. In contrast, metrics related to pragmatics, such as *speaker consistency* and *instantaneous coordination*, are fairly well established but are less frequently used. Morphology metrics, particularly *active words* and *average message length*, are more commonly used, whereas syntax remains a peripheral concern, with only two isolated metrics proposed and not adopted in subsequent research. This imbalance indicates that while semantic metrics dominate EL research, morphology and pragmatics receive moderate attention, and syntax is mostly neglected.

Furthermore, the optimality of these metrics is not straightforward. Rather than being simply minimized or maximized, their ideal values are likely to lie at a nuanced balance point that varies depending on the specific EL system and application. This uncertainty leaves the critical question of what constitutes a 'good' EL system largely unanswered. Addressing this gap will require a deeper exploration of underrepresented metrics and a more refined understanding of how to evaluate EL systems holistically.

## 7 Future work

In this section, we outline potential future directions for the research field of EL, based on our vision outlined in Sect. 7.1. We present major research opportunities, organized along key research dimensions, in Sect. 7.2.

Along with future research directions, we have summarized a list of open source code repositories in Table 10 in Appendix A that can serve as convenient starting points for

experimenting with these directions, for example, comprehensive frameworks such as the EGG toolkit [162] and BabyAI [131] are included.

## 7.1 Vision

Our vision for EL research is grounded in a functional perspective, aiming to achieve significant breakthroughs in human-agent interaction [24, 26, 29, 33, 35, 176, 256]. This means developing communication systems that enable HCI at the human level, addressing the purpose, cost, and value of communication with intuitive and effective interfaces [27, 36, 58, 148, 189, 233]. A key goal is to ensure that ELs are grounded in real-world contexts, allowing agents to understand and interact with human-like comprehension and vice versa [19, 21, 40, 258]. This includes creating hierarchical, compositional conceptualization capabilities that allow agents to discuss and understand novel concepts in a structured, human-relevant manner [25, 133, 182, 187]. In addition, exploring the potential for AI explainability through communication is an exciting area [21, 127, 189]. Finally, in the long term, creation and creativity through EL comparable to human capabilities would be a milestone. This would allow agents to truly communicate on a human level and enhance their ability to perceive and adapt to their environment through the use of language [103].

## 7.2 Dimensions and opportunities

The development, evaluation, and application of EL in communication systems can be systematically analyzed along several critical dimensions. Given the relative youth of the field, with the majority of research emerging within the last eight years, specific areas of focus have gained prominence, particularly in the context of semantic metrics, such as topographic similarity and zero-shot evaluation, as highlighted in Fig. 12 in Appendix B. However, Fig. 13 in Appendix B illustrates that there is no discernible chronological trend or evolution in the way the different language characteristics are addressed. This absence of a historical trajectory is likely attributable to the relatively brief history of the field and the considerable diversity of proposed approaches and methodologies. Despite this, we identified nine key dimensions that, to the best of our knowledge, represent the primary areas of focus in EL research.

### 7.2.1 Evaluation metrics

Evaluation metrics are essential for rigorously assessing the characteristics and effectiveness of ELs. As detailed in our taxonomy (cf. Section 5.4), we have identified key characteristics and their associated metrics. While some EL features are quantifiable through multiple metrics and have been examined in multiple studies, others remain underexplored, as illustrated in Fig. 12 in Appendix B. We emphasize the need to develop comprehensive and quantitative metrics that accurately capture these features, which are critical to determining the practical utility of ELs. Previous studies have similarly highlighted this need [25, 29, 46, 72, 109, 126, 179, 234, 241]. In addition, further research is needed to systematically investigate existing metrics, especially with respect to their sensitivity to variations in settings, algorithms, and agent architectures [29, 50]. It is imperative that these metrics be subjected to more rigorous investigation to ensure that they enable meaningful quantitative comparisons and support well-founded conclusions about the capabilities and

utility of ELs. Thus, we endorse more comprehensive studies, more edge case testing and, in particular, more analysis of actual human-agent interaction. We see this as a critical priority for advancing the field.

### 7.2.2 Emergent language and natural language alignment

This dimension addresses the convergence and divergence between EL and NL. A key approach to this challenge, discussed in Sect. 5.3, involves leveraging language priors to guide this alignment. Achieving robust EL-NL alignment is essential for advancing human-agent interaction. Thus, future research should explore the integration of NL-centered metrics and regularization techniques to enhance this alignment [14, 107]. However, this alignment presents a fundamental dilemma. On the one hand, agents need the autonomy to develop languages organically, tailored to their specific interactions and requirements. On the other hand, to facilitate seamless human-agent communication, these ELs must closely resemble NLs, which imposes significant constraints on their development. This tension creates what we call the Evolution-Acquisition Dilemma, where the evolutionary process fosters intrinsically motivated language emergence, while the acquisition process necessitates alignment with NL. Balancing these competing needs is a critical challenge for future research in this area.

### 7.2.3 Emergent language and large language models

The remarkable performance of LLM on various benchmarks has established them as a cornerstone of modern NLP [80] and potential foundation for more complex agents [259–262]. Despite their success, however, LLM face fundamental limitations, particularly in grounding language use in shared environments and experiences [263] as well as agency [264] and truthfulness [265]. Addressing these shortcomings may require insights from EL research. A key challenge in EL is the evolution-acquisition dilemma - the need to ground language in shared, incremental experiences, which current learning systems cannot achieve due to resource and technology constraints. While most LLM applications rely on fine-tuning [266] and scaling [85], these methods do not inherently address this challenge or the broader issues of grounding and adaptability. One promising avenue lies in the concept of agentic LLM [263, 267, 268] or cognitive language agents [269], which combine the representational strength of LLMs with the adaptive, experiential learning processes of RL. Here, looking at opportunities for EL related research, LLM might act as language priors, providing a foundation that can be iteratively refined through agentic interaction and experience [263, 270]. This approach mirrors human language acquisition, where teachers provide guidance based on shared experience. In the absence of such a teacher, agentic LLM provide a synthetic framework for combining supervised and EL paradigms, allowing agents to relax static supervised training regimes and develop more adaptive communication protocols. We argue for further exploration of cognitive language agents, focusing not only on established benchmarks but also on challenges central to EL research. Bridging these fields could open up new opportunities for developing systems that combine the scalability of LLM with the adaptability and grounding capabilities of EL.

### 7.2.4 Representation learning

EL can be viewed as a complex representation learning task, focusing on how agents encode, interpret, and construct internal representations of observations and linguistic data. While representation learning is a well-established area in artificial intelligence research, its application in the context of EL remains underexplored. This dimension is central to the analysis of meaning and language space as outlined in our framework, which is based on the semiotic cycle (cf. Figure 10). Advancing this dimension requires advanced latent space analyses to elucidate the relationships between ELs, underlying world models, and NL structures. In addition, evaluating the impact of discrete versus continuous representations is critical to refining our understanding of EL dynamics. Future research directions include developing methodologies to ensure that agent representations more accurately reflect the input they receive [16], exploring efficient representation of (multimodal) information [124], conducting in-depth analyses to uncover and mitigate influencing factors and biases in learned representations [46], and assessing the efficacy of these representations for downstream tasks [111].

### 7.2.5 Agent design

Agent design is a critical aspect in EL research, directly influencing the linguistic capabilities and adaptability of artificial agents. Prominent research directions include the investigation of advanced neural network architectures tailored for EL [25, 170, 186], the creation of architectures optimized for heterogeneous and dynamic agent populations, and the refinement of structures that enhance language emergence and linguistic properties [111, 167]. In addition, modular designs rather than monolithic ones potentially offer advantages by separating language processing from other task-specific computations. Addressing these design challenges is critical to advancing both EL research and broader artificial intelligence goals.

### 7.2.6 Setting design

The environment in which agents operate is central to shaping the EL, encompassing interaction rules, agent goals, and communication dynamics (cf. Table 3). This dimension is integral to the setting space outlined in our framework (cf. Figure 10). Important future research directions include scaling up experimental settings to include larger and more complex tasks [14, 25, 130, 146, 202, 230] with a focus on realistic perceptually grounded game environments [110, 202]. In addition, the study of the impact of populations as such [179] and the use of heterogeneous agent populations [118] are crucial areas of research. While some benchmarks have been established and utilized [131, 162], there remains a significant need for the development and widespread dissemination of comprehensive benchmarks in area of research.

### 7.2.7 Communication design

The design of the communication channel in EL systems is critical, focusing on how agents exchange and structure information through the channels available to them. This aspect is directly related to the phonetics and phonology components outlined in our taxonomy (cf. Section 5.4.1 and Sect. 5.4.2). For discrete ELs, it is essential to establish channels that support word-based communication, with considerations such as vocabulary size and variable message length being fundamental to enabling effective and scalable human-agent

interaction. Future research directions in this area include the exploration of topology-aware variable communication channels, the integration of heterogeneous channels within multi-agent systems, and the evolution of communication channels over time. Moreover, the incorporation of multimodal communication channels could provide more realistic and contextually rich stimuli, which may significantly enhance the sophistication and applicability of ELs in NL-oriented human-agent coordination [25].

### 7.2.8 Learning strategies

Learning strategies focus on how agents acquire, adapt, and refine their linguistic capabilities over time, including the development of language rules and adaptation through interactions with other agents. While MARL serves as the foundational framework, there is significant potential to enhance the learning process through strategic design choices. Future research directions include the exploration of advanced regularization techniques [107, 240], the adoption of tailored optimization strategies [25], and the integration of supervised or self-supervised learning objectives using appropriate loss designs [15, 151]. Additionally, the application of meta-learning [155], decentralized learning approaches [56], and curriculum learning methodologies [56] offer promising avenues for optimizing the EL learning process.

### 7.2.9 Human-agent interaction

The final dimension focuses on the interpretability of ELs by humans and the degree to which humans can shape their development. This aspect is critical for creating human-agent interaction systems where communication is intuitive and effective [124]. To advance this dimension, future research should prioritize the integration of human-in-the-loop feedback mechanisms to ensure that ELs are not only practical, but also comprehensible to human users [18, 24]. This will improve the usability and adoption of these systems in real-world applications. Key research directions include designing experiments that create incentives for agents to develop communication strategies more closely aligned with human language [16]. Additionally, exploring the resilience of communication protocols to deception through training with competing agents can lead to more robust and realistic interactions [34]. Exploring adaptive communication strategies to optimize the sparsity and clarity of messages based on individual or group needs within human-agent teams is another promising direction [233].

## 8 Limitations and discussion

In this section, we critically evaluate the limitations of our survey and identify areas for future improvement. Through our review, we aimed to develop a detailed taxonomy for the field of EL, focusing on its key properties (cf. Section 5), and to analyze as well as categorize quantification approaches and metrics (cf. Section 6). In addition, we curated a summary of open questions and suggestions for future research (cf. Section 7). Despite considerable efforts to establish a viable taxonomy and framework in the most systematic and unbiased manner, there are several potential limitations to our research approach and methodology.

First, while we have provided an extensive overview of 181 scientific publications in EL research, it is important to acknowledge that our search process, despite being thorough, may have overlooked significant contributions. Consequently, we do not claim completeness. However, we are very confident that our review represents a fair and well-balanced reflection of the existing body of work and the current state of the art.



Second, our review includes sources that are not peer-reviewed, such as preprints from [arXiv](#), to ensure that our work captures the most recent developments and diverse perspectives, including those that might be controversial. While we have carefully examined each paper included in this review, we cannot guarantee that every detail in non-peer-reviewed papers is entirely accurate. Consequently, we focused on concepts, findings, and metrics that are supported by multiple studies.

Third, we have introduced a taxonomy and a comprehensive metrics categorization for EL research, a field that is still in its early stages. This effort comes with inherent challenges, and while we have addressed many of these, it is important to note that our proposed framework does not represent a consensus within the wider research community. We are transparent about this limitation and encourage further discussion and validation.

Fourth, in order to maintain focus and conciseness, we have deliberately excluded ideas that lack associated metrics. As a result, some conceptual ideas from the reviewed research literature that are difficult to quantify in this early stage may not be fully explored in this survey.

Finally, we have incorporated several existing metrics into our proposed framework. While many of these metrics are well established in the field, we acknowledge that a more rigorous and critical experimental evaluation of these metrics would be beneficial. We strongly recommend that future research conduct such evaluations to further refine and validate the tools and methods used in EL research.

## 9 Conclusion

In this paper, we present a comprehensive taxonomy of (EL), an overview of applicable metrics, and a summary of open challenges and potential research directions. Additionally, we provide a list of open source code repositories of the field in Table 10 in Appendix A. Our overall goal is to create a standardised yet dynamic framework that not only facilitates progress in this area of research, but also stimulates further interest and exploration.

Section 2 introduces the foundational linguistic concepts that underpin our taxonomy. Section 5 offers a comprehensive taxonomy of EL based on the review of 181 scientific publications. Section 6 presents a unified categorization and notation for various metrics, depicted in Fig. 9, ensuring consistency and clarity. Section 7 provides a summary of current achievements and outlines research opportunities.

By providing a structured overview and systematic categorization of linguistic concepts relevant to EL we have created a common ground for research and discussion. The detailed presentation of metrics and their unified notation ensures readability and usability, making it easier for researchers to navigate related topics and identify potential research opportunities and blind spots of future publications and the research field as a whole. This survey provides a valuable perspective on the development and analysis of EL, serving as both a guide and a resource for advancing this area of study.

EL is a fascinating and promising way to achieve grounded and goal-oriented communication among agents and between humans and agents. Despite its significant progress in recent years, the field faces many open questions and requires further evaluation methods and metrics. Critical questions remain about the measurability of linguistic features, the validity of proposed metrics, their utility, and their necessity. Aligning EL with (NLP) for (HCI) presents additional opportunities and challenges. We encourage continued contributions and interdisciplinary research to address these issues and advance the field.

## Appendix A

See Tables 10, 11, 12.

**Table 10** List of code repositories for the literature reviewed

Paper	Link
[21]	<a href="https://github.com/agakshat/visualdialog-pytorch">https://github.com/agakshat/visualdialog-pytorch</a>
[134]	<a href="https://github.com/jacobandreas/tre">https://github.com/jacobandreas/tre</a>
[135]	<a href="https://github.com/facebookresearch/EGG/tree/main/egg/zoo/compo_vs_generalization_ood">https://github.com/facebookresearch/EGG/tree/main/egg/zoo/compo_vs_generalization_ood</a>
[201]	<a href="https://github.com/arski/LEW">https://github.com/arski/LEW</a>
[22]	<a href="https://github.com/proroklab/adversarial_comms">https://github.com/proroklab/adversarial_comms</a>
[33]	<a href="https://github.com/benbogin/emergence-communication-cco/">https://github.com/benbogin/emergence-communication-cco/</a>
[136]	<a href="https://github.com/brendon-boldt/filex-emergent-language">https://github.com/brendon-boldt/filex-emergent-language</a>
[137]	<a href="https://github.com/brendon-boldt/filex-emergent-language">https://github.com/brendon-boldt/filex-emergent-language</a>
[16]	<a href="https://github.com/DianeBouchacourt/SignalingGame">https://github.com/DianeBouchacourt/SignalingGame</a>
[202]	<a href="https://github.com/facebookresearch/fruit-tools-game">https://github.com/facebookresearch/fruit-tools-game</a>
[23]	<a href="https://github.com/nicofirst1/r1_werewolf">https://github.com/nicofirst1/r1_werewolf</a>
[142]	<a href="https://github.com/facebookresearch/EGG/blob/master/egg/zoo/channel/README.md">https://github.com/facebookresearch/EGG/blob/master/egg/zoo/channel/README.md</a>
[132]	<a href="https://github.com/facebookresearch/brica">https://github.com/facebookresearch/brica</a>
[122]	<a href="https://github.com/facebookresearch/EGG/blob/master/egg/zoo/compo_vs_generalization/README.md">https://github.com/facebookresearch/EGG/blob/master/egg/zoo/compo_vs_generalization/README.md</a>
[25]	<a href="https://github.com/deepmind/emergent_communication_at_scale">https://github.com/deepmind/emergent_communication_at_scale</a>
[131]	<a href="https://github.com/mila-iqia/babyai/tree/master">https://github.com/mila-iqia/babyai/tree/master</a>
[144]	<a href="https://github.com/AriChow/EL">https://github.com/AriChow/EL</a>
[105]	<a href="https://github.com/mcogswell/evolang">https://github.com/mcogswell/evolang</a>
[103]	<a href="https://github.com/flowersteam/Imagine">https://github.com/flowersteam/Imagine</a>
[109]	<a href="https://github.com/DylanCope/zero-shot-comm">https://github.com/DylanCope/zero-shot-comm</a>
[116]	<a href="https://github.com/gautierdag/cultural-evolution-engine">https://github.com/gautierdag/cultural-evolution-engine</a>
[19]	<a href="https://github.com/batra-mlp-lab/visdial-rl">https://github.com/batra-mlp-lab/visdial-rl</a>
[50]	<a href="https://github.com/Near32/ReferentialGym">https://github.com/Near32/ReferentialGym</a>
[147]	<a href="https://github.com/Near32/ReferentialGym/tree/master/zoo/referential-games%2Bst-gs">https://github.com/Near32/ReferentialGym/tree/master/zoo/referential-games%2Bst-gs</a>
[148]	<a href="https://github.com/Near32/Regym/tree/develop-ETHER/benchmark/ETHER">https://github.com/Near32/Regym/tree/develop-ETHER/benchmark/ETHER</a>
[149]	<a href="https://github.com/Near32/ReferentialGym/tree/develop/zoo/referential-games%2Bcompositionality%2Bdisentanglement">https://github.com/Near32/ReferentialGym/tree/develop/zoo/referential-games%2Bcompositionality%2Bdisentanglement</a>
[151]	<a href="https://github.com/facebookresearch/EGG/tree/main/egg/zoo/emcom_as_ssl">https://github.com/facebookresearch/EGG/tree/main/egg/zoo/emcom_as_ssl</a>
[152]	<a href="https://github.com/CLMBRs/communication-translation">https://github.com/CLMBRs/communication-translation</a>
[110]	<a href="https://github.com/blinodelka/Multiagent-Communication-Learning-in-Networks">https://github.com/blinodelka/Multiagent-Communication-Learning-in-Networks</a>
[203]	<a href="https://github.com/nyu-dl/MultimodalGame">https://github.com/nyu-dl/MultimodalGame</a>
[111]	<a href="https://github.com/jacopotagliabue/On-the-plurality-of-graphs">https://github.com/jacopotagliabue/On-the-plurality-of-graphs</a>
[242]	<a href="https://github.com/alshedivat/lola">https://github.com/alshedivat/lola</a>
[155]	<a href="https://github.com/Shawn-Guo-CN/EmergentNumerals">https://github.com/Shawn-Guo-CN/EmergentNumerals</a>
[157]	<a href="https://github.com/Shawn-Guo-CN/GameBias-EmeCom2020">https://github.com/Shawn-Guo-CN/GameBias-EmeCom2020</a>
[120]	<a href="https://github.com/uoe-agents/Expressivity-of-Emergent-Languages">https://github.com/uoe-agents/Expressivity-of-Emergent-Languages</a>
[159]	<a href="https://github.com/SonuDixit/gComm">https://github.com/SonuDixit/gComm</a>
[232]	<a href="https://github.com/Meta-optimization/emergent_communication_in_agents">https://github.com/Meta-optimization/emergent_communication_in_agents</a>
[160]	<a href="https://fringsoo.github.io/pragmatic_in2_emergent_papersite/">https://fringsoo.github.io/pragmatic_in2_emergent_papersite/</a>

**Table 10** continued

Paper	Link
[162]	<a href="https://github.com/facebookresearch/EGG">https://github.com/facebookresearch/EGG</a>
[163]	<a href="https://github.com/facebookresearch/EGG/tree/master/egg/zoo/language_bottleneck">https://github.com/facebookresearch/EGG/tree/master/egg/zoo/language_bottleneck</a>
[121]	<a href="https://github.com/facebookresearch/EGG/tree/master/egg/zoo/compositional_efficiency">https://github.com/facebookresearch/EGG/tree/master/egg/zoo/compositional_efficiency</a>
[164]	<a href="https://github.com/tomekkorbak/compositional-communication-via-template-transfer">https://github.com/tomekkorbak/compositional-communication-via-template-transfer</a>
[51]	<a href="https://github.com/tomekkorbak/measuring-non-trivial-compositionality">https://github.com/tomekkorbak/measuring-non-trivial-compositionality</a>
[127]	<a href="https://github.com/batra-mlp-lab/lang-emerge">https://github.com/batra-mlp-lab/lang-emerge</a>
[206]	<a href="https://github.com/facebookresearch/translagent">https://github.com/facebookresearch/translagent</a>
[169]	<a href="https://github.com/MediaBrain-SJTU/ECISQA">https://github.com/MediaBrain-SJTU/ECISQA</a>
[170]	<a href="https://github.com/cambridgeltl/ECNMT">https://github.com/cambridgeltl/ECNMT</a>
[108]	<a href="https://github.com/pliang279/Competitive-Emergent-Communication">https://github.com/pliang279/Competitive-Emergent-Communication</a>
[112]	<a href="https://github.com/ToruOwO/marl-ae-comm">https://github.com/ToruOwO/marl-ae-comm</a>
[235]	<a href="https://github.com/olipinski/rl_werewolf">https://github.com/olipinski/rl_werewolf</a>
[171]	<a href="https://anonymous.4open.science/r/TPG-916B">https://anonymous.4open.science/r/TPG-916B</a>
[29]	<a href="https://github.com/facebookresearch/measuring-emergent-comm">https://github.com/facebookresearch/measuring-emergent-comm</a>
[113]	<a href="https://github.com/backpropper/s2p">https://github.com/backpropper/s2p</a>
[174]	<a href="https://github.com/Ddaniela13/LearningToDraw">https://github.com/Ddaniela13/LearningToDraw</a>
[37]	<a href="https://github.com/jayelm/emergent-generalization">https://github.com/jayelm/emergent-generalization</a>
[34]	<a href="https://github.com/mnoukhov/emergent-compete">https://github.com/mnoukhov/emergent-compete</a>
[177]	<a href="https://github.com/XeniaOhmer/hierarchical_reference_game">https://github.com/XeniaOhmer/hierarchical_reference_game</a>
[178]	<a href="https://github.com/XeniaOhmer/language_perception_communication_games">https://github.com/XeniaOhmer/language_perception_communication_games</a>
[207]	<a href="https://github.com/saimwani/CoMON">https://github.com/saimwani/CoMON</a>
[180]	<a href="https://github.com/asappresearch/compositional-inductive-bias">https://github.com/asappresearch/compositional-inductive-bias</a>
[181]	<a href="https://github.com/evaportelance/emergent-shape-bias">https://github.com/evaportelance/emergent-shape-bias</a>
[117]	<a href="https://github.com/Joshua-Ren/Neural_Iterated_Learning">https://github.com/Joshua-Ren/Neural_Iterated_Learning</a>
[119]	<a href="https://github.com/backpropper/cbc-emecom">https://github.com/backpropper/cbc-emecom</a>
[183]	<a href="https://github.com/MathieuRita/Population">https://github.com/MathieuRita/Population</a>
[208]	<a href="https://github.com/wilrop/communication_monfg">https://github.com/wilrop/communication_monfg</a>
[209]	<a href="https://github.com/Homagn/MultiAgentRL">https://github.com/Homagn/MultiAgentRL</a>
[226]	<a href="https://github.com/david-simoes-93/A3C3">https://github.com/david-simoes-93/A3C3</a>
[227]	<a href="https://github.com/david-simoes-93/A3C3">https://github.com/david-simoes-93/A3C3</a>
[188]	<a href="https://github.com/shanest/function-words-context">https://github.com/shanest/function-words-context</a>
[38]	<a href="https://github.com/CLMBRs/communication-translation">https://github.com/CLMBRs/communication-translation</a>
[20]	<a href="https://github.com/facebookarchive/CommNet">https://github.com/facebookarchive/CommNet</a>
[191]	<a href="https://github.com/mynlp/emecom_SignalingGame_as_betaVAE">https://github.com/mynlp/emecom_SignalingGame_as_betaVAE</a>
[192]	<a href="https://github.com/thomasaunger/babyai_sr">https://github.com/thomasaunger/babyai_sr</a>
[193]	<a href="https://github.com/i-machine-think/emergent_grammar_induction">https://github.com/i-machine-think/emergent_grammar_induction</a>
[229]	<a href="https://github.com/TonghanWang/NDQ">https://github.com/TonghanWang/NDQ</a>
[238]	<a href="https://github.com/jimmyyhwu/spatial-intention-maps">https://github.com/jimmyyhwu/spatial-intention-maps</a>
[197]	<a href="https://github.com/wildphoton/Compositional-Generalization">https://github.com/wildphoton/Compositional-Generalization</a>
[107]	<a href="https://github.com/ysmyth/ec-nl">https://github.com/ysmyth/ec-nl</a>
[244]	<a href="https://github.com/geek-ai/Magent">https://github.com/geek-ai/Magent</a>

**Table 11** Overview of language prior usage in the reviewed literature

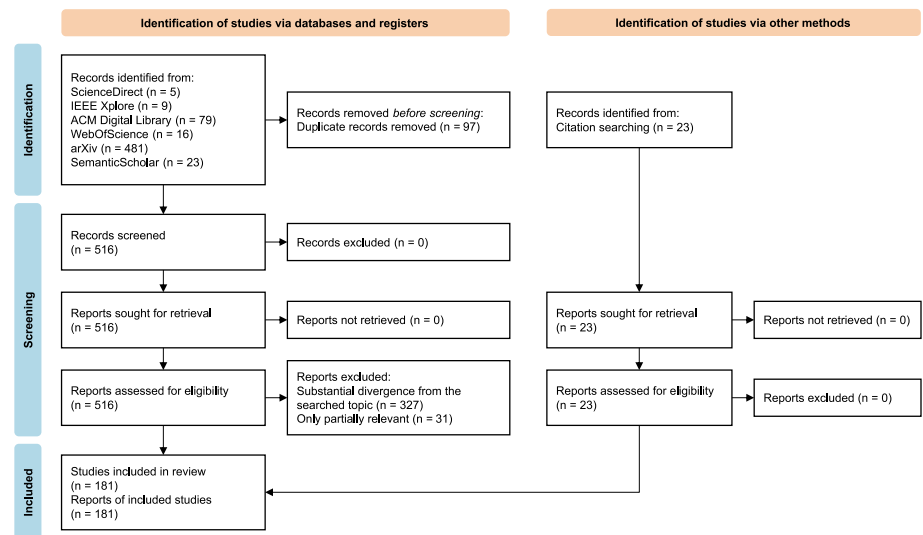
Language prior	Paper
No-Evolution	[16, 18, 20, 22, 23, 25, 27–34, 37, 46, 51, 53, 55, 72, 102, 105–112, 114–120, 122, 126, 127, 130, 134–147, 149–151, 154–158, 160–168, 170–179, 181–188, 190, 192, 193, 196–198, 200, 202–204, 207, 209–213, 215–223, 225–236, 239–241, 243]
Yes-Acquisition	[21, 26, 38, 103, 104, 113, 121, 129, 131–133, 148, 152, 159, 169, 180, 191, 194, 195, 205, 206, 214, 237, 238]
Both	[14, 15, 17, 19, 153, 189, 199]

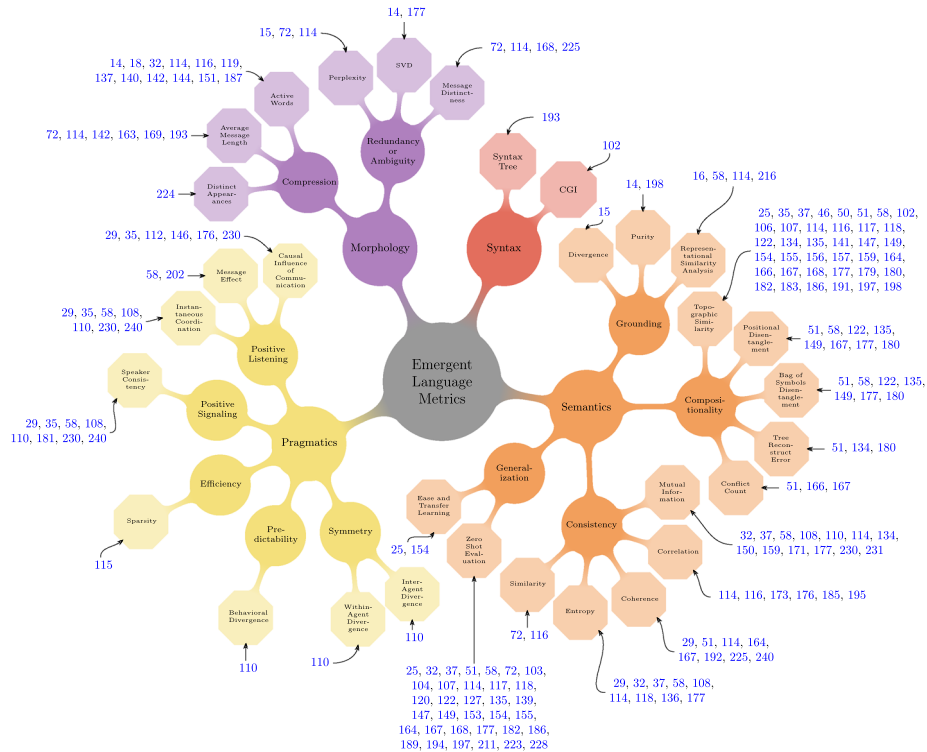
**Table 12** Overview of language characteristics discussed in the reviewed literature

Characteristic	Paper
Morphology	[14, 15, 18, 32, 50, 58, 72, 102, 106, 114, 116, 118, 119, 137, 140–142, 144, 149, 151, 163, 168, 169, 177, 187, 193, 213, 224, 225]
Syntax	[102, 193]
Semantics	[14–19, 21, 25, 26, 28, 29, 32, 33, 35, 37, 46, 50, 51, 58, 72, 102–108, 110, 114, 116–122, 127, 134–136, 138, 139, 141, 144, 146–157, 159, 160, 164–169, 171, 173–177, 179, 180, 182–187, 189, 191, 192, 194, 195, 197, 198, 202, 207, 211, 213, 216, 223–225, 228, 230, 231, 234, 240]
Pragmatics	[17, 29, 34, 35, 58, 105, 108–110, 112, 115, 124, 146, 176, 181, 186, 187, 200, 202, 207, 230, 234, 240]

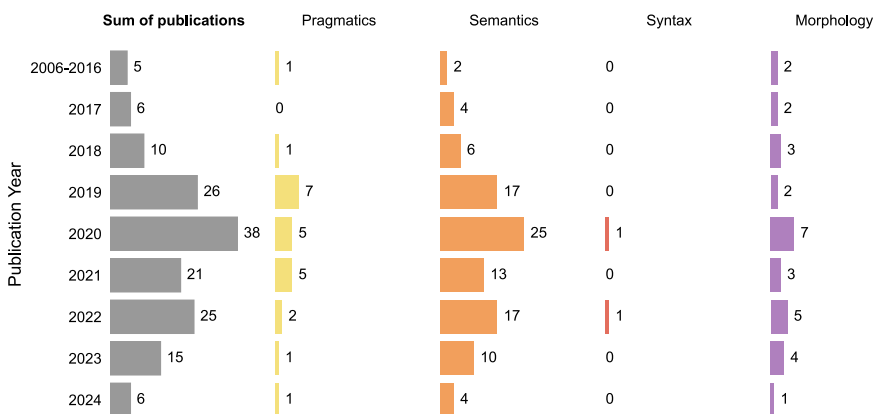
## Appendix B

See Figs. 11, 12, 13.

**Fig. 11** PRISMA 2020 flow diagram for new systematic reviews for the present survey. Adapted from [48]



**Fig. 12** Graph presenting a visual representation of the metrics identified in the surveyed literature, sorted by language characteristics. A list of references for each metric is given at each node. ● ● ● ●: Language characteristics (inner nodes) ● ● ● ●: Individual metrics (outermost nodes) (Color figure online)



**Fig. 13** Number of publications sorted by year and by language characteristics analyzed. The number of publications analyzing individual language characteristics per year is provided in the leftmost column, and the distribution across analyzed language characteristics is shown in the remaining columns

**Author Contributions** J.P., H.T. and T.M. had the idea for the article. J.P. performed the literature search and data analysis. The first draft of the manuscript was written by J.P. with the continuous support of C.W.d.P. and H.T. The first draft of the metrics section was written by A.G. All authors commented on earlier versions of the manuscript and critically revised the final manuscript.

**Funding** Open Access funding enabled and organized by Projekt DEAL. We acknowledge the funding of the internships of Arya Gopikrishnan and Gustavo Adolpho Lucas De Carvalho by the German Academic Exchange Service (DAAD) project ‘RISE Germany’.

**Data availability** Not applicable.

**Materials availability** Not applicable.

**Code availability** Not applicable.

## Declarations

**Conflict of interest** Not applicable.

**Ethics approval and consent to participate** Not applicable.

**Consent for publication** Not applicable.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Lewis, D. K. (1969). *Convention: A philosophical study* (1st ed.). Harvard University Press.
- Wagner, K., Reggia, J. A., Uriagereka, J., & Wilkinson, G. S. (2003). Progress in the simulation of emergent communication and language. *Adaptive Behavior*, 11(1), 37–69. <https://doi.org/10.1177/10597123030111003>
- Steels, L. (1997). The synthetic modeling of language origins. *Evolution of Communication*, 1(1), 1–34. <https://doi.org/10.1075/eoc.1.1.02ste>
- Nowak, M. A., & Krakauer, D. C. (1999). The evolution of language. *Proceedings of the National Academy of Sciences*, 96(14), 8028–8033. <https://doi.org/10.1073/pnas.96.14.8028>
- Kirby, S. (2002). Natural language from artificial life. *Artificial Life*, 8(2), 185–215. <https://doi.org/10.1162/106454602320184248>
- Cangelosi, A., & Parisi, D. (2002). *Simulating the Evolution of Language*. Springer London. <https://doi.org/10.1007/978-1-4471-0663-0>
- Christiansen, M. H., & Kirby, S. (2003). *Language Evolution*. Oxford University Press.
- Batali, J. (1998). Computational simulations of the emergence of grammar. In J. Hurford, C. Knight, & M. Studdert-Kennedy (Eds.), *Approaches to the Evolution of Language* (pp. 405–426). Cambridge University Press.
- Oliphant, M., & Batali, J. (1997). Learning and the emergence of coordinated communication. *Center for Research on Language Newsletter*, 11(1), 1–46.
- Steels, L. (1995). A self-organizing spatial vocabulary. *Artificial Life*, 2(3), 319–332. <https://doi.org/10.1162/artl.1995.2.3.319>
- Skyrms, B. (2002). Signals, evolution and the explanatory power of transient information. *Philosophy of Science*, 69(3), 407–428. <https://doi.org/10.1086/342451>
- Smith, K., Kirby, S., & Brighton, H. (2003). Iterated learning: A framework for the emergence of language. *Artificial Life*, 9(4), 371–386. <https://doi.org/10.1162/106454603322694825>

13. Foerster, J. N., Assael, Y. M., Freitas, N. D., & Whiteson, S. (2017). Learning to communicate with deep multi-agent reinforcement learning. In D. D. Lee, U. Luxburg, R. Garnett, M. Sugiyama, & I. Guyon (Eds.), *Advances in neural information processing systems* (29th ed., pp. 2145–2153). Curran Associates Inc.
14. Lazaridou, A., Peysakhovich, A., Baroni, M. (2017). Multi-agent cooperation and the emergence of (natural) language. In: OpenReview.net (ed.) 5th international conference on learning representations. <https://openreview.net/forum?id=Hk8N3ScIq>
15. Havrylov, S., & Titov, I. (2017). Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In U. Luxburg, I. Guyon, S. Bengio, H. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (30th ed., pp. 2146–2156). Curran Associates Inc.
16. Bouchacourt, D., Baroni, M. (2018). How agents see things: On visual representations in an emergent language game. In: Association for Computational Linguistics (ed.) Proceedings of the 2018 conference on empirical methods in natural language processing, pp. 981–985. <https://doi.org/10.18653/v1/D18-1119>
17. Cao, K., Lazaridou, A., Lanctot, M., Leibo, J.Z., Tuyls, K., Clark, S. (2018). Emergent communication through negotiation. In: OpenReview.net (ed.) 6th international conference on learning representations: Conference track proceedings. <https://openreview.net/forum?id=Hk6WhagRW>
18. Mordatch, I., Abbeel, P. (2018) Emergence of grounded compositional language in multi-agent populations. In: Association for the advancement of artificial intelligence (ed.) Proceedings of the thirty-second AAAI conference on artificial intelligence and thirtieth innovative applications of artificial intelligence conference and eighth AAAI symposium on educational advances in artificial intelligence, pp. 1495–1502. AAAI Press. <https://cdn.aaai.org/ojs/11492/11492-13-15020-1-2-20201228.pdf>
19. Das, A., Kottur, S., Moura, J.M.F., Lee, S., Batra, D. (2017). Learning cooperative visual dialog agents with deep reinforcement learning. In: 2017 IEEE International conference on computer vision (ICCV), pp. 2970–2979. IEEE. <https://doi.org/10.1109/ICCV.2017.321> . <http://arxiv.org/pdf/1703.06585v2>
20. Sukhbaatar, S., Szlam, A., & Fergus, R. (2017). Learning multiagent communication with backpropagation. In D. D. Lee, U. Luxburg, R. Garnett, M. Sugiyama, & I. Guyon (Eds.), *Advances in neural information processing systems* (29th ed., pp. 2252–2260). Curran Associates Inc.
21. Agarwal, A., Gurumurthy, S., Sharma, V., Lewis, M., Sycara, K. (2019). Community regularization of visually-grounded dialog. In: International foundation for autonomous agents and multiagent systems (ed.) proceedings of the 18th international conference on autonomous agents and multiagent systems. ACM digital library, pp. 1042–1050. International Foundation for Autonomous Agents and Multiagent Systems. <https://www.ifaamas.org/Proceedings/aamas2019/pdfs/p1042.pdf>
22. Blumenkamp, J., Prorok, A. (2020). The emergence of adversarial communication in multi-agent reinforcement learning. In: PMLR (ed.) 4th Conference on robot learning. Proceedings of machine learning research, pp. 1394–1414. <https://proceedings.mlr.press/v155/blumenkamp21a.html>
23. Brandizzi, N., Grossi, D., Iocchi, L. (2021). Rlupus: Cooperation through emergent communication in the werewolf social deduction game. In: 13th Adaptive and learning agents workshop at AAMAS 2021. <http://arxiv.org/pdf/2106.05018v2>
24. Brandizzi, N., Iocchi, L. (2022). Emergent communication in human-machine games. In: 5th Workshop on emergent communication at ICLR 2022. <https://openreview.net/forum?id=rqLgeQWCXZ9>
25. Chaabouni, R., Strub, F., Altché, F., Tarassov, E., Tallec, C., Davoodi, E., Mathewson, K.W., Tieleman, O., Lazaridou, A., Piot, B. (2022). Emergent communication at scale. In: OpenReview.net (ed.) 10th International conference on learning representations. <https://openreview.net/forum?id=AUGBfDIV9rL>
26. Gupta, A., Lanctot, M., & Lazaridou, A. (2021). Dynamic population-based meta-learning for multi-agent communication with natural language. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, & J. Wortman Vaughan (Eds.), *Advances in neural information processing systems (NeurIPS 2021)*, (34th ed.). Curran Associates Inc.
27. Karten, S., Agrawal, S., Tucker, M., Hughes, D., Lewis, M., Shah, J., Sycara, K. (2022). The enforcers: Consistent sparse-discrete methods for constraining informative emergent communication. <http://arxiv.org/pdf/2201.07452v1>
28. Lo, Y.L., Sengupta, B., Lo Long, Y. (2022). Learning to ground decentralized multi-agent communication with contrastive learning. In: 5th Workshop on emergent communication at ICLR 2022. <https://openreview.net/forum?id=rLceWXCmZc>
29. Lowe, R., Foerster, J., Boureau, Y.-L., Pineau, J., Dauphin, Y. (2019). On the pitfalls of measuring emergent communication. In: International foundation for autonomous agents and multiagent systems (ed.) Proceedings of the 18th international conference on autonomous agents and multiagent systems. ACM digital library, pp. 693–701. International foundation for autonomous agents and multiagent systems. <https://www.ifaamas.org/Proceedings/aamas2019/pdfs/p693.pdf>

30. Vanneste, S., Vanneste, A., Mets, K., Schepper, T.D., Anwar, A., Mercelis, S., Latré, S., Hellinckx, P. (2022). Learning to communicate using counterfactual reasoning. In: 14th Workshop on adaptive and learning agents at AAMAS 2022. [https://ala2022.github.io/papers/ALA2022\\_paper\\_17.pdf](https://ala2022.github.io/papers/ALA2022_paper_17.pdf)
31. Verma, S. (2021). Towards sample efficient learners in population based referential games through action advising: Extended abstract. In: Proceedings of the 20th international conference on autonomous agents and multiagent systems. AAMAS '21. International foundation for autonomous agents and multiagent systems. <https://www.ifaamas.org/Proceedings/aamas2021/pdfs/p1689.pdf>
32. Yu, D., Mu, J., Goodman, N. (2022). Emergent covert signaling in adversarial reference games. In: 5th Workshop on emergent communication at ICLR 2022. <https://openreview.net/forum?id=H-eMQbR7Z5>
33. Bogin, B., Geva, M., Berant, J. (2018). Emergence of communication in an interactive world with consistent speakers. In: 2nd Workshop on emergent communication at NeurIPS 2018. <http://arxiv.org/pdf/1809.00549v2>
34. Noukhovitch, M., LaCroix, T., Lazaridou, A., Courville, A. (2021). Emergent communication under competition. In: Proceedings of the 20th international conference on autonomous agents and multiagent systems. AAMAS '21. International foundation for autonomous agents and multiagent systems. <https://www.ifaamas.org/Proceedings/aamas2021/pdfs/p974.pdf>
35. Lazaridou, A., Baroni, M. (2020). Emergent multi-agent communication in the deep learning era. <http://arxiv.org/pdf/2006.02419v2>
36. Galke, L., Ram, Y., Raviv, L. (2022). Emergent communication for understanding human language evolution: What's missing? In: 5th Workshop on emergent communication at ICLR 2022. <https://openreview.net/forum?id=rqUGZQ-0XZ5>
37. Mu, J., Goodman, N. (2021). Emergent communication of generalizations. In Neural Information Processing Systems Foundation (ed.) Advances in neural information processing systems. Advances in neural information processing systems, 34, pp 17994–18007. Curran Associates Inc. <https://papers.nips.cc/paper/2021/file/9597353e41e6957b5e7aa79214fcb256-Paper.pdf>
38. Steinert-Threlkeld, S., Zhou, X., Liu, Z., Downey, C.M. (2022). Emergent communication fine-tuning (ec-ft) for pretrained language models. In: 5th Workshop on emergent communication at ICLR 2022. <https://openreview.net/forum?id=SUqrM7WR7W5>
39. Bender, E.M., Koller, A. (2020). Climbing towards nlu: On meaning, form, and understanding in the age of data. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th annual meeting of the association for computational linguistics, pp. 5185–5198. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.463>
40. Lemon, O. (2022). Conversational grounding in emergent communication—data and divergence. In: 5th Workshop on emergent communication at ICLR 2022. <https://openreview.net/forum?id=BbG-m-0Xbq>
41. Browning, J., Lecun, Y. (2022). AI and the limits of language: An artificial intelligence system trained on words and sentences alone will never approximate human understanding., Online. <https://www.noemamag.com/ai-and-the-limits-of-language/>
42. Manning, C. D., & Schütze, H. (2005). *Foundations of Statistical Natural Language Processing* (8th ed.). MIT Press.
43. Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10), 1872–1897. <https://doi.org/10.1007/s11431-020-1647-3>
44. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A. (2020). Transformers: State-of-the-art natural language processing. In: Liu, Q., Schlangen, D. (eds.) Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations, pp. 38–45. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-demos.6>
45. van Eecke, P., Beuls, K. (2021). Re-conceptualising the language game paradigm in the framework of multi-agent reinforcement learning. In: Association for the Advancement of Artificial (ed.) COMARL AAAI 2020-2021—Challenges and opportunities for multi-agent reinforcement learning, AAAI Spring Symposium Series. <https://arxiv.org/pdf/2004.04722>
46. Keresztury, B., Bruni, E. (2020). Compositional properties of emergent languages in deep learning. <http://arxiv.org/pdf/2001.08618v1>
47. Hernandez-Leal, P., Kartal, B., & Taylor, M. E. (2019). A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 33(6), 750–797. <https://doi.org/10.1007/s10458-019-09421-1>
48. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson,



- A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The prisma 2020 statement: An updated guideline for reporting systematic reviews. *BMJ (Clinical Research Edition)*, 372, 71. <https://doi.org/10.1136/bmj.n71>
49. Lipowska, D., & Lipowski, A. (2022). Emergence and evolution of language in multi-agent systems. *Lingua*, 272, 103331. <https://doi.org/10.1016/j.lingua.2022.103331>
50. Denamganaï, K., Walker, J.A. (2020). Referentialgym: A nomenclature and framework for language emergence & grounding in (visual) referential games. In: 4th Workshop on emergent communication at NeurIPS 2020. <http://arxiv.org/pdf/2012.09486v1>
51. Korbak, T., Zubek, J., Rączaszek-Leonardi, J. (2020). Measuring non-trivial compositionality in emergent communication. In: 4th Workshop on emergent communication at NeurIPS 2020. <http://arxiv.org/pdf/2010.15058v2>
52. LaCroix, T. (2019). Biology and compositionality: Empirical considerations for emergent-communication protocols. In: 3rd workshop on emergent communication at NeurIPS 2019. <http://arxiv.org/pdf/1911.11668v2>
53. Mihai, D., Hare, J. (2021). The emergence of visual semantics through communication games. <http://arxiv.org/pdf/2101.10253v1>
54. Galke, L., Raviv, L. (2024). Emergent communication and learning pressures in language models: a language evolution perspective. <http://arxiv.org/pdf/2403.14427v1>
55. Vanneste, A., Vanneste, S., Mets, K., Schepper, T.D., Mercelis, S., Latré, S., Hellinckx, P. (2022). An analysis of discretization methods for communication learning with multi-agent reinforcement learning. In: 14th Workshop on adaptive and learning agents at AAMAS 2022. <http://arxiv.org/pdf/2204.05669v1>
56. Moulin-Frier, C., Oudeyer, P.-Y. (2021). Multi-agent reinforcement learning as a computational tool for language evolution research: Historical context and future challenges. In: Association for the Advancement of Artificial (ed.) COMARL AAAI 2020-2021—Challenges and opportunities for multi-agent reinforcement learning, AAAI Spring Symposium Series. <https://arxiv.org/pdf/2002.08878>
57. Fernando, C., Zenkova, D., Nikolov, S., Osindero, S. (2020). From language games to drawing games. <http://arxiv.org/pdf/2010.02820v2>
58. Brandizzi, N. (2023). Toward more human-like ai communication: A review of emergent communication research. *IEEE Access*, 11, 142317–142340. <https://doi.org/10.1109/ACCESS.2023.3339656>
59. Carston, R. (2009). The explicit/implicit distinction in pragmatics and the limits of explicit communication. *International Review of Pragmatics*, 1(1), 35–62. <https://doi.org/10.1163/187731009X455839>
60. Watzlawick, P., Bavelas, J. B., & Jackson, D. D. (1967). *Pragmatics of human communication: A study of interactional patterns, pathologies, and paradoxes*. Norton.
61. Andersen, P. A. (1991). When one cannot not communicate: A challenge to motley's traditional communication postulates. *Communication Studies*, 42(4), 309–325. <https://doi.org/10.1080/10510979109368346>
62. Antos, G., Ventola, E., & Weber, T. (2008). *Handbook of interpersonal communication*. De Gruyter Mouton. <https://doi.org/10.1515/9783110211399>
63. Witt, P. (2016). *Communication and Learning*. De Gruyter Mouton. <https://doi.org/10.1515/9781501502446>
64. Hartley, P. (1993). *Interpersonal Communication* (1st ed.). Routledge.
65. Jones, R. G. (2018). *Communication in the real world* (2nd ed.). Flat World Knowledge.
66. Bossert, W. H., & Wilson, E. O. (1963). The analysis of olfactory communication among animals. *Journal of Theoretical Biology*, 5(3), 443–469. [https://doi.org/10.1016/0022-5193\(63\)90089-4](https://doi.org/10.1016/0022-5193(63)90089-4)
67. Sales, G., & Pye, D. (1974). *Ultrasonic Communication by Animals*. Springer. <https://doi.org/10.1007/978-94-011-6901-1>
68. Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nature Neuroscience*, 12(6), 718–724. <https://doi.org/10.1038/nn.2331>
69. Tronick, E. Z. (1989). Emotions and emotional communication in infants. *The American Psychologist*, 44(2), 112–119. <https://doi.org/10.1037/0003-066X.44.2.112>
70. Grosse, G., Behne, T., Carpenter, M., & Tomasello, M. (2010). Infants communicate in order to be understood. *Developmental Psychology*, 46(6), 1710–1722. <https://doi.org/10.1037/a0020727>
71. Stokoe, W. C. (1980). Sign language structure. *Annual Review of Anthropology*, 9(1), 365–390. <https://doi.org/10.1146/annurev.an.09.100180.002053>
72. Choi, E., Lazaridou, A., Freitas, N.d. (2018) Compositional oververter communication learning from raw visual input. In: OpenReview.net (ed.) 6th International conference on learning representations: Conference track proceedings. <https://openreview.net/forum?id=rknt2Be0->
73. Austin, J. L. (1975). *How to do Things with Words: The William James lectures delivered at Harvard University in 1955* (2nd ed.). Clarendon Press.

74. Clark, H. H. (1996). *Using language* (1st ed.). Cambridge University Press.
75. Wittgenstein, L. (1989). *Philosophical investigations* (3rd ed.). Blackwell.
76. Adler, R. B. (2012). *Interplay: The process of interpersonal communication* (3rd ed.). Oxford University Press.
77. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Iegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D. (2020) Language models are few-shot learners. In: Neural Information Processing Systems Foundation (ed.) *Advances in neural information processing systems* 33. *Advances in neural information processing systems*, vol. 33, pp. 1877–1901. Curran Associates Inc. <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfc4967418fbf-b8ac142f64a-Paper.pdf>
78. Lauriola, I., Lavelli, A., & Aioli, F. (2022). An introduction to deep learning in natural language processing: Models, techniques, and tools. *Neurocomputing*, 470, 443–456. <https://doi.org/10.1016/j.neucom.2021.05.103>
79. Khurana, D., Koli, A., Khatter, K., & Singh, S. (2022). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*. <https://doi.org/10.1007/s11042-022-13428-4>
80. Lappin, S. (2023). Assessing the strengths and weaknesses of large language models. *Journal of Logic, Language and Information*. <https://doi.org/10.1007/s10849-023-09409-x>
81. Lazaridou, A., Kuncoro, A., Gribovskaya, E., Agrawal, D., Liska, A., Terzi, T., Gimenez, M., Massond Autume, C., Kocisky, T., Ruder, S., Yogatama, D., Cao, K., Young, S., & Blunsom, P. (2021). Mind the gap: Assessing temporal generalization in neural language models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, & J. Wortman Vaughan (Eds.), *Advances in neural information processing systems (NeurIPS)* (pp. 29348–29363). Curran Associates Inc.
82. Merrill, W., Goldberg, Y., Schwartz, R., & Smith, N. A. (2021). Provable limitations of acquiring meaning from ungrounded form: What will future language models understand? *Transactions of the Association for Computational Linguistics*, 9, 1047–1060.
83. Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du Zhengxiao, Zhang, C., Shen, S., Zhang, T., Su, Y., Sun, H., Huang, M., Dong, Y., Tang, J. (2024). Agentbench: Evaluating llms as agents. In: OpenReview.net (ed.) 12th International conference on learning representations. <https://openreview.net/forum?id=zAdUB0aCTQ>
84. Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., Farajtabar, M. (2024). GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. <http://arxiv.org/pdf/2410.05229v1>
85. Zhou, L., Schellaert, W., Martínez-Plumed, F., Moros-Daval, Y., Ferri, C., & Hernández-Orallo, J. (2024). Larger and more instructable language models become less reliable. *Nature*, 634(8032), 61–68. <https://doi.org/10.1038/s41586-024-07930-y>
86. Grupen, N.A., Lee, D.D., Selman, B. (2022). Multi-agent curricula and emergent implicit signaling. In: Proceedings of the 21st International conference on autonomous agents and multiagent systems (AAMAS 2022). <https://www.ifaamas.org/Proceedings/aamas2022/pdfs/p553.pdf>
87. Dor, D., Knight, C., Lewis, J. (2014). The Social Origins of Language, In: Oxford linguistics, Oxford University Press, 1, 19.
88. Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, 13(4), 707–727. <https://doi.org/10.1017/S0140525X00081061>
89. Hauser, M. D., Yang, C., Berwick, R. C., Tattersall, I., Ryan, M. J., Watumull, J., Chomsky, N., & Lewontin, R. C. (2014). The mystery of language evolution. *Frontiers in Psychology*, 5, 401. <https://doi.org/10.3389/fpsyg.2014.00401>
90. Hock, H. H., & Joseph, B. D. (2019). *Language History, Language Change, and Language Relationship*. De Gruyter Mouton. <https://doi.org/10.1515/9783110613285>
91. Chomsky, N. (1986). *Knowledge of language: its nature, origin, and use*. Convergence. Praeger.
92. Ney, J. W. (1989). Knowledge of language: Its nature, origin, and use. *Language Sciences*, 11(4), 409–423. [https://doi.org/10.1016/0388-0001\(89\)90029-6](https://doi.org/10.1016/0388-0001(89)90029-6)
93. Tomasello, M. (2010). *Origins of Human Communication* A Bradford book (1st ed.). MIT Press.
94. Lakoff, G., & Johnson, M. (2003). *Metaphors we live by: With a new afterword*. University of Chicago Press.
95. Locke, J. L. (1997). A theory of neurolinguistic development. *Brain and Language*, 58(2), 265–326. <https://doi.org/10.1006/brln.1997.1791>
96. Tomasello, M. (2009). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

97. Lakoff, G. (1990). *Women, fire, and dangerous things: What categories reveal about the mind*. The University of Chicago Press.
98. Bleys, J. (2015). Language strategies for the domain of colour. Language Science Press. <https://doi.org/10.17169/langsci.b51.104>
99. Thomas, J. J., & Cook, K. A. (2005). *Illuminating the path: The research and development agenda for visual analytics*. IEEE Computer Society.
100. Chandler, D. (2007). *Semiotics: The basics*. London, England: Routledge. <https://doi.org/10.4324/9780203014936>
101. Brinton, L. J., & Brinton, D. M. (2010). *The linguistic structure of modern English*. John Benjamins Publishing Company. <https://doi.org/10.1075/z.156>
102. Ueda, R., Ishii, T., Washio, K., Miyao, Y. (2022). Categorical grammar induction as a compositionality measure for emergent languages in signaling games. In: 5th Workshop on emergent communication at ICLR 2022. <https://openreview.net/forum?id=Sbg7b0Q-5>
103. Colas, C., Karch, T., Lair, N., Dussoux, J.-M., Moulin-Frier, C., Dominey, P.F., Oudeyer, P.-Y. (2020). Language as a cognitive tool to imagine goals in curiosity-driven exploration. In: Neural Information Processing Systems Foundation (ed.) Advances in neural information processing systems 33. Advances in neural information processing systems. Curran Associates Inc. [https://papers.neurips.cc/paper\\_files/paper/2020/file/274e6fcf4a583de4a81c6376f17673e7-Paper.pdf](https://papers.neurips.cc/paper_files/paper/2020/file/274e6fcf4a583de4a81c6376f17673e7-Paper.pdf)
104. Qiu, S., Xie, S., Fan, L., Gao, T., Zhu, S.-C., Zhu, Y. (2022). Emergent graphical conventions in a visual communication game. In: Neural Information Processing Systems Foundation (ed.) Advances in neural information processing systems 35. Advances in neural information processing systems. Curran Associates Inc. [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/550ff553efc2c58410f277c667d12786-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/550ff553efc2c58410f277c667d12786-Paper-Conference.pdf)
105. Cogswell, M., Lu, J., Lee, S., Parikh, D., Batra, D. (2019). Emergence of compositional language with deep generational transmission. <http://arxiv.org/pdf/1904.09067v2>
106. Li, F., Bowling, M. (2019). Ease-of-teaching and language structure from emergent communication. In: Neural Information Processing Systems Foundation (ed.) Advances in neural information processing systems 32. Advances in neural information processing systems. Curran Associates Inc. <https://papers.nips.cc/paper/2019/file/b0cf188d74589db9b23d5d277238a929-Paper.pdf>
107. Yao, S., Yu, M., Zhang, Y., Narasimhan, K., Tenenbaum, J.B., Gan, C. (2022). Linking emergent and natural languages via corpus transfer. In: OpenReview.net (ed.) 10th international conference on learning representations. <https://openreview.net/forum?id=49A1Y6tRhaq>
108. Liang, P.P., Chen, J., Salakhutdinov, R., Morency, L.-P., Kottur, S. (2020). On emergent communication in competitive multi-agent teams. In: Proceedings of the 19th International conference on autonomous agents and multiagent systems. AAMAS '20, pp. 735–743. International Foundation for Autonomous Agents and Multiagent Systems. <https://www.ifaamas.org/Proceedings/aamas2020/pdfs/p735.pdf>
109. Cope, D., Schoots, N. (2020). Learning to communicate with strangers via channel randomisation methods. In: 4th Workshop on emergent communication at NeurIPS 2020. <http://arxiv.org/pdf/2104.09557v1>
110. Dubova, M., Moskvichev, A., Goldstone, R. (2020). Reinforcement communication learning in different social network structures. In: 1st Workshop on language in reinforcement learning at ICML 2020. <http://arxiv.org/pdf/2007.09820v1>
111. Fitzgerald, N., Tagliabue, J. (2020). On the plurality of graphs. In: Ntereason @ ECAI 2020. <http://arxiv.org/pdf/2008.00920v1>
112. Lin, T., Huh, M., Stauffer, C., Lim, S.-N., Isola, P. (2021). Learning to ground multi-agent communication with autoencoders. In: Neural Information Processing Systems Foundation (ed.) Advances in neural information processing systems 34. Advances in neural information processing systems, pp. 15230–15242. Curran Associates Inc. <https://papers.nips.cc/paper/2021/file/80fee67c8a4c4989bf8a580b4bbb0cd2-Paper.pdf>
113. Lowe, R., Gupta, A., Foerster, J., Kiela, D., Pineau, J. (2020). On the interaction between supervision and self-play in emergent communication. In: OpenReview.net (ed.) 8th International conference on learning representations. <https://openreview.net/forum?id=rJxGLlBtWtH>
114. Luna, D.R., Ponti, E.M., Hupkes, D., Bruni, E. (2020). Internal and external pressures on language emergence: least effort, object constancy and frequency. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.findings-emnlp.397>
115. Kalinowska, A., Davoodi, E., Strub, F., Mathewson, K., Murphey, T., Pilarski, P. (2022). Situated communication: A solution to over-communication between artificial agents. In: 5th Workshop on emergent communication at ICLR 2022. <https://openreview.net/forum?id=HLqzzQWA7Z9>

116. Dagan, G., Hupkes, D., Bruni, E. (2021). Co-evolution of language and agents in referential games. In: Association for Computational Linguistics (ed.) Proceedings of the 16th conference of the European chapter of the association for computational linguistics: Main Volume. <https://doi.org/10.18653/v1/2021.eacl-main.260>
117. Ren, Y., Guo, S., Labeau, M., Cohen, S.B., Kirby, S. (2020). Compositional languages emerge in a neural iterated learning model. In: OpenReview.net (ed.) 8th International conference on learning representations. <https://openreview.net/forum?id=HkePNpVKPB>
118. Rita, M., Strub, F., Grill, J.-B., Pietquin, O., Dupoux, E. (2022). On the role of population heterogeneity in emergent communication. In: OpenReview.net (ed.) 10th International conference on learning representations. <https://openreview.net/forum?id=5Qkd7-bZfl>
119. Resnick, C., Gupta, A., Foerster, J., Dai, A.M., Cho, K. (2020). Capacity, bandwidth, and compositionality in emergent language learning. In: Proceedings of the 19th International conference on autonomous agents and multiagent systems. AAMAS '20, pp. 1125–1133. International foundation for autonomous agents and multiagent systems. <https://www.ifaamas.org/Proceedings/aamas2020/pdfs/p1125.pdf>
120. Guo, S., Ren, Y., Mathewson, K., Kirby, S., Albrecht, S.V., Smith, K. (2022). Expressivity of emergent language is a trade-off between contextual complexity and unpredictability. In: OpenReview.net (ed.) 10th International conference on learning representations. [https://openreview.net/forum?id=WxuE\\_JWxjkW](https://openreview.net/forum?id=WxuE_JWxjkW)
121. Kharitonov, E., Baroni, M. (2020). Emergent language generalization and acquisition speed are not tied to compositionality. In: Proceedings of the third BlackboxNLP workshop on analyzing and interpreting neural networks for NLP, pp. 11–15. <https://aclanthology.org/2020.blackboxnlp-1.2.pdf>
122. Chaabouni, R., Kharitonov, E., Bouchacourt, D., Dupoux, E., Baroni, M. (2020). Compositionality and generalization in emergent languages. In: Association for Computational Linguistics (ed.) Proceedings of the 58th annual meeting of the association for computational linguistics, pp. 4427–4442. <https://doi.org/10.18653/v1/2020.acl-main.407>
123. Suglia, A., Konstas, I., & Lemon, O. (2024). Visually grounded language learning: A review of language games, datasets, tasks, and models. *Journal of Artificial Intelligence Research*, 79, 173–239. <https://doi.org/10.1613/jair.1.15185>
124. Zhu, C., Dastani, M., Wang, S. (2024). A survey of multi-agent deep reinforcement learning with communication. In: Proceedings of the 23rd International conference on autonomous agents and multiagent systems. AAMAS '24, pp. 2845–2847. International Foundation for Autonomous Agents and Multiagent Systems. <https://doi.org/10.1007/s10458-023-09633-6>
125. Boldt, B., Mortensen, D. (2022). Recommendations for systematic research on emergent language. <http://arxiv.org/pdf/2206.11302v1>
126. Chen, R., Guo, S. (2023). Emergent semantic communications for mobile augmented reality: Basic ideas and opportunities. <http://arxiv.org/pdf/2308.07342v1>
127. Kottur, S., Moura, J., Lee, S., Batra, D. (2017). Natural language does not emerge ‘naturally’ in multi-agent dialog. In: Palmer, M., Hwa, R., Riedel, S. (eds.) Proceedings of the 2017 conference on empirical methods in natural language processing, pp. 2962–2967. Association for Computational Linguistics. <https://doi.org/10.18653/v1/D17-1321>
128. Perkins, H. (2021). TexRel: A green family of datasets for emergent communications on relations. <http://arxiv.org/pdf/2105.12804v1>
129. Buck, C., Bulian, J., Ciaramita, M., Gajewski, W., Gesmundo, A., Hounsby, N., Wang, W. (2017). Analyzing language learned by an active question answering agent. In: 1st Workshop on emergent communication at NeurIPS 2017. <http://arxiv.org/pdf/1801.07537v1>
130. Harding Graesser, L., Cho, K., Kiela, D. (2019). Emergent linguistic phenomena in multi-agent communication games. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp. 3698–3708. Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1384>
131. Chevalier-Boisvert, M., Bahdanau, D., Lahlou, S., Willems, L., Saharia, C., Nguyen, T.H., Bengio, Y. (2019). Babyai: A platform to study the sample efficiency of grounded language learning. In: OpenReview.net (ed.) 7th International conference on learning representations. <https://openreview.net/forum?id=rJeXCo0cYX>
132. Chaabouni, R., Kharitonov, E., Lazaric, A., Dupoux, E., Baroni, M. (2019). Word-order biases in deep-agent emergent communication. In: Association for Computational Linguistics (ed.) Proceedings of the 57th annual meeting of the association for computational linguistics, pp. 5166–5175. <https://doi.org/10.18653/v1/P19-1509>

133. Woodward, M., Finn, C., Hausman, K. (2020). Learning to interactively learn and assist. In: Proceedings of the 34th AAAI conference on artificial intelligence (AAAI-20), pp. 2535–2543. <https://doi.org/10.1609/aaai.v34i03.5636>
134. Andreas, J. (2019). Measuring compositionality in representation learning. In: OpenReview.net (ed.) 7th International conference on learning representations. <https://openreview.net/forum?id=HJz05o0qK7>
135. Auersperger, M., Pecina, P. (2022). Defending compositionality in emergent languages. In: Ippolito, D., Li, L.H., Pacheco, M.L., Chen, D., Xue, N. (eds.) Proceedings of the 2022 Conference of the North American chapter of the association for computational linguistics: Human language technologies: Student research workshop, pp. 285–291. Association for Computational Linguistics, Hybrid+ Online. <https://doi.org/10.18653/v1/2022.naacl-srw.35>
136. Boldt, B., Mortensen, D. (2022). Modeling emergent lexicon formation with a self-reinforcing stochastic process. In: 5th Workshop on emergent communication at ICLR 2022. <https://openreview.net/forum?id=BdbVIXbRXbq>
137. Boldt, B., Mortensen, D. (2022). Mathematically modeling the lexicon entropy of emergent language. <http://arxiv.org/pdf/2211.15783v2>
138. Bosc, T. (2022). Varying meaning complexity to explain and measure compositionality. In: 5th Workshop on emergent communication at ICLR 2022. <https://openreview.net/forum?id=BnGzfmZ07bq>
139. Bullard, K., Kiela, D., Pineau, J., Foerster, J. (2021). Quasi-equivalence discovery for zero-shot emergent communication. <http://arxiv.org/pdf/2103.08067v1>
140. Carmeli, B., Meir, R., Belinkov, Y. (2023). Emergent quantized communication. In: Williams, B.K., Chen, Y., Neville, J. (eds.) Thirty-Seventh AAAI Conference on artificial intelligence thirty-fifth conference on innovative applications of artificial intelligence thirteenth symposium on educational advances in artificial intelligence, February 7–14, 2023, Washington DC, USA. Proceedings of the AAAI conference on artificial intelligence, vol. 37, pp. 11533–11541. Association for the Advancement of Artificial Intelligence = AAAI. <https://doi.org/10.1609/aaai.v37i10.26363>
141. Carmeli, B., Belinkov, Y., Meir, R. (2024). Concept-best-matching: Evaluating compositionality in emergent communication. <http://arxiv.org/pdf/2403.14705v1>
142. Chaabouni, R., Kharitonov, E., Dupoux, E., Baroni, M. (2019). Anti-efficient encoding in emergent communication. In: Neural Information Processing Systems Foundation (ed.) Advances in neural information processing systems, 39. Advances in neural information processing systems. Curran Associates Inc. <https://proceedings.neurips.cc/paper/2019/file/31ca0ca71184bbdb3de7b20a51e88e90-Paper.pdf>
143. Chowdhury, A., Santamaria-Pang, A., Kubricht, J.R., Qiu, J., Tu, P. (2020). Symbolic semantic segmentation and interpretation of COVID-19 Lung infections in chest CT volumes based on emergent languages. <http://arxiv.org/pdf/2008.09866v1>
144. Chowdhury, A., Santamaria-Pang, A., Kubricht, J.R., Tu, P. (2021). Emergent symbolic language based deep medical image classification. In: 2021 IEEE 18th International symposium on biomedical imaging (ISBI), pp. 689–692. IEEE. <https://doi.org/10.1109/ISBI48211.2021.9434073>
145. Chowdhury, A., Kubricht, J.R., Sood, A., Tu, P., Santamaria-Pang, A. (2020). Escell: Emergent symbolic cellular language. In: 2020 IEEE 17th International symposium on biomedical imaging (ISBI), pp. 1604–1607. IEEE. <https://doi.org/10.1109/ISBI45749.2020.9098343>
146. Cowen-Rivers, A.I., Naradowsky, J. (2019). Emergent communication with world models. In: 3rd Workshop on emergent communication at NeurIPS 2019. <http://arxiv.org/pdf/2002.09604v1>
147. Denamganai, K., Walker, J.A. (2020). On (emergent) systematic generalisation and compositionality in visual referential games with straight-through gumbel-softmax estimator. In: 4th Workshop on emergent communication at NeurIPS 2020. <http://arxiv.org/pdf/2012.10776v1>
148. Denamganai, K., Hernandez, D., Vardal, O., Missaoui, S., Walker, J.A. (2023). ETHER: Aligning emergent communication for hindsight experience replay. <http://arxiv.org/pdf/2307.15494v2>
149. Denamganai, K., Missaoui, S., Walker, J.A. (2023). Visual referential games further the emergence of disentangled representations. <http://arxiv.org/pdf/2304.14511v1>
150. Dessi, R., Bouchacourt, D., Crepaldi, D., Baroni, M. (2019). Focus on what's informative and ignore what's not: Communication strategies in a referential game. In: 3rd Workshop on emergent communication at NeurIPS 2019. <http://arxiv.org/pdf/1911.01892v1>
151. Dessi, R., Kharitonov, E., Baroni, M. (2021). Interpretable agent communication from scratch (with a generic visual processor emerging on the side). In: Neural Information Processing Systems Foundation (ed.) Advances in neural information processing systems 34. Advances in neural information processing systems, pp. 26937–26949. Curran Associates Inc. <https://papers.nips.cc/paper/2021/file/e250ce59336b505ed411d455abaa30b4d-Paper.pdf>
152. Downey, C.M., Zhou, X., Liu, Z., Steinert-Threlkeld, S. (2023). Learning to translate by learning to communicate. In: Ataman, D. (ed.) Proceedings of the 3rd Workshop on multi-lingual representation



- learning (MRL), pp. 218–238. Association for Computational Linguistics. <https://aclanthology.org/2023.mrl-1.17>
153. Elof, K., Pretorius, A., Räsänen, O., Engelbrecht, H.A., Kamper, H. (2023). Towards learning to speak and hear through multi-agent communication over a continuous acoustic channel. <http://arxiv.org/pdf/2111.02827v2>
  154. Feng, Y., An, B., & Lu, Z. (2024). Learning multi-object positional relationships via emergent communication. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16), 17371–17379. <https://doi.org/10.1609/aaai.v38i16.29685>
  155. Guo, S. (2019). Emergence of numeric concepts in multi-agent autonomous communication. In: Master thesis, University of Edinburgh. <http://arxiv.org/pdf/1911.01098v1>
  156. Guo, S., Ren, Y., Havrylov, S., Frank, S., Titov, I., Smith, K. (2020). The emergence of compositional languages for numeric concepts through iterated learning in neural agents. In: Proceedings of the 13th International conference on evolution in language (EvoLang XIII). [https://brussels.evolang.org/proceedings/papers/EvoLang13\\_paper\\_82.pdf](https://brussels.evolang.org/proceedings/papers/EvoLang13_paper_82.pdf)
  157. Guo, S., Ren, Y., Slowik, A., Mathewson, K. (2020). Inductive bias and language expressivity in emergent communication. In: 4th Workshop on emergent communication at NeurIPS 2020. <http://arxiv.org/pdf/2012.02875v1>
  158. Hagiwara, Y., Furukawa, K., Taniguchi, A., & Taniguchi, T. (2022). Multiagent multimodal categorization for symbol emergence: Emergent communication via interpersonal cross-modal inference. *Advanced Robotics*, 36(5–6), 239–260. <https://doi.org/10.1080/01691864.2022.2029721>
  159. Hazra, R., Dixit, S., Sen, S. (2020). Infinite use of finite means: Zero-Shot generalization using compositional emergent protocols. <http://arxiv.org/pdf/2012.05011v2>
  160. Kang, Y., Wang, T., Melo, G.D. (2020). Incorporating pragmatic reasoning communication into emergent language. In: Neural Information Processing Systems Foundation (ed.) Advances in neural information processing systems 33. Advances in neural information processing systems. Curran Associates Inc. [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/7520fa31d14f45add6d61e52df5a03ff-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/7520fa31d14f45add6d61e52df5a03ff-Paper.pdf)
  161. Karten, S., Kailas, S., Li, H., Sycara, K. (2023) On the role of emergent communication for social learning in multi-agent reinforcement learning. <http://arxiv.org/pdf/2302.14276v1>
  162. Kharitonov, E., Chaabouni, R., Bouchacourt, D., Baroni, M. (2019). Egg: a toolkit for research on emergence of language in games. In: Association for Computational Linguistics (ed.) Proceedings of the 2019 Conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP): System Demonstrations, pp. 55–60. <https://doi.org/10.18653/v1/D19-3010>
  163. Kharitonov, E., Chaabouni, R., Bouchacourt, D., Baroni, M. (2020). Entropy minimization in emergent languages. In: Proceedings of the 37th International conference on machine learning (ICML 2020). <https://proceedings.mlr.press/v119/kharitonov20a/kharitonov20a.pdf>
  164. Korbak, T., Zubek, J., Kuciński, Ł., Miłoś, P., Rączaszek-Leonardi, J. (2019). Developmentally motivated emergence of compositional communication via template transfer. In: 3rd Workshop on emergent communication at NeurIPS 2019. <http://arxiv.org/pdf/1910.06079v1>
  165. Kubricht, J.R., Yang, Z., Qiu, J., Tu, P.H. (2023). Grounded language acquisition from object and action imagery. <http://arxiv.org/pdf/2309.06335v1>
  166. Kuciński, Ł., Kołodziej, P., Miłoś, P. (2020). Emergence of compositional language in communication through noisy channel. In: 1st Workshop on language in reinforcement learning at ICML 2020. [https://openreview.net/forum?id=ZbXISL\\_xwtA](https://openreview.net/forum?id=ZbXISL_xwtA)
  167. Kuciński, Ł., Korbak, T., Kołodziej, P., Miłoś, P. (2021). Catalytic role of noise and necessity of inductive biases in the emergence of compositional communication. In: Neural Information Processing Systems Foundation (ed.) Advances in neural information processing systems 34. Advances in neural information processing systems, pp. 23075–23088. Curran Associates Inc. <https://papers.nips.cc/paper/2021/file/c2839bed26321da8b466c80a032e4714-Paper.pdf>
  168. Lazaridou, A., Hermann, K.M., Tuyls, K., Clark, S. (2018). Emergence of linguistic communication from referential games with symbolic and pixel input. In: OpenReview.net (ed.) 6th International conference on learning representations: conference track proceedings. <https://openreview.net/forum?id=HJGv1Z-AW>
  169. Lei, Z., Zhang, Y., Xiong, Y., Chen, S. (2023). Emergent communication in interactive sketch question answering. In: Neural Information Processing Systems Foundation (ed.) Advances in neural information processing systems 36. Advances in neural information processing systems. [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/746cflbc2337700f7f0c35c7b02638cc-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/746cflbc2337700f7f0c35c7b02638cc-Paper-Conference.pdf)
  170. Li, Y., Ponti, E.M., Vulić, I., Korhonen, A. (2020). Emergent communication pretraining for few-shot machine translation. In: International Committee on Computational Linguistics (ed.) Proceedings of the

- 28th international conference on computational linguistics. <https://doi.org/10.18653/v1/2020.coling-main.416>
171. Lipinski, O., Sobey, A.J., Cerutti, F., Norman, T.J. (2024). Speaking your language: Spatial relationships in interpretable emergent communication. In: Neural Information Processing Systems Foundation (ed.) Advances in neural information processing systems 37. Advances in neural information processing systems. <https://openreview.net/forum?id=vIP8IWmZIN>
  172. Lobos-Tsunekawa, K., Srinivasan, A., Spranger, M. (2022). MA-Dreamer: Coordination and communication through shared imagination. <http://arxiv.org/pdf/2204.04687v1>
  173. Mihai, D., Hare, J. (2019). Avoiding hashing and encouraging visual semantics in referential emergent language games. In: 3rd Workshop on emergent communication at NeurIPS 2019. <http://arxiv.org/pdf/1911.05546v1>
  174. Mihai, D., Hare, J. (2021). Learning to draw: Emergent communication through sketching. In: Neural Information Processing Systems Foundation (ed.) Advances in neural information processing systems 34. Advances in neural information processing systems, pp. 7153–7166. Curran Associates Inc. <https://papers.nips.cc/paper/2021/file/39d0a8908fbc6c18039ea8227f827023-Paper.pdf>
  175. Mu, Y., Yao, S., Ding, M., Luo, P., Gan, C. (2023). Ec2 : Emergent communication for embodied control. In: 2023 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp. 6704–6714. IEEE. <https://doi.org/10.1109/CVPR52729.2023.00648>
  176. Mul, M., Bouchacourt, D., Bruni, E. (2019). Mastering emergent language: learning to guide in simulated navigation. <http://arxiv.org/pdf/1908.05135v1>
  177. Ohmer, X., Duda, M., Bruni, E.: Emergence of hierarchical reference systems in multi-agent communication. In: Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T.K., Santus, E., Bond, F., Na, S.-H. (eds.) (2022). Proceedings of the 29th International conference on computational linguistics. International Committee on Computational Linguistics. <https://aclanthology.org/2022.coling-1.501/>
  178. Ohmer, X., Marino, M., Franke, M., & König, P. (2022). Mutual influence between language and perception in multi-agent communication games. *PLoS Computational Biology*, 18(10), 1010658. <https://doi.org/10.1371/journal.pcbi.1010658>
  179. Ossenkopf, M., Luck, K.S., Mathewson, K.W. (2022). Which language evolves between heterogeneous agents? - communicating movement instructions with widely different time scopes. In: 5th Workshop on emergent communication at ICLR 2022. <https://openreview.net/forum?id=BnfgM7-0mW5>
  180. Perkins, H. (2021). Neural networks can understand compositional functions that humans do not, in the context of emergent communication. <http://arxiv.org/pdf/2103.04180v1>
  181. Portelance, E., Frank, M.C., Jurafsky, D., Sordoni, A., Laroché, R. (2021). The emergence of the shape bias results from communicative efficiency. In: Bisazza, A., Abend, O. (eds.) Proceedings of the 25th conference on computational natural language learning. Association for Computational Linguistics, Online. <https://doi.org/10.18653/v1/2021.conll-1.48>
  182. Ri, R., Ueda, R., Naradowsky, J. (2023). Emergent communication with attention. In: Goldwater, M., Anggoro, F.K., Hayes, B.K., Ong, D.C. (eds.) Proceedings of the annual meeting of the cognitive science society. <http://arxiv.org/pdf/2305.10920v1>
  183. Rita, M., Tallec, C., Michel, P., Grill, J.-B., Pietquin, O., Dupoux, E., Strub, F. (2022). Emergent communication: Generalization and overfitting in lewis games. In: Neural Information Processing Systems Foundation (ed.) Advances in neural information processing systems 35. Advances in neural information processing systems. Curran Associates Inc. [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/093b08a7ad6e6dd8d34b9cc86bb5f07c-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/093b08a7ad6e6dd8d34b9cc86bb5f07c-Paper-Conference.pdf)
  184. Santamaria-Pang, A., Kubricht, J.R., Devaraj, C., Chowdhury, A., Tu, P. (2019). Towards semantic action analysis via emergent language. In: 2019 IEEE International conference on artificial intelligence and virtual reality (AIVR), pp. 224–2244. IEEE. <https://doi.org/10.1109/AIVR46125.2019.00047>
  185. Santamaria-Pang, A., Kubricht, J.R., Chowdhury, A., Bhushan, C., Tu, P. (2020). Towards emergent language symbolic semantic segmentation and model interpretability. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K. (eds.) Medical image computing and computer assisted intervention — MICCAI 2020. Image processing, computer vision, pattern recognition, and graphics. Springer International Publishing and Imprint. [https://doi.org/10.1007/978-3-030-59710-8\\_32](https://doi.org/10.1007/978-3-030-59710-8_32)
  186. Slowik, A., Gupta, A., Hamilton, W.L., Jamnik, M., Holden, S.B., Pal, C. (2021). Exploring structural inductive biases in emergent communication. In: Cognitive Science Society (ed.) 43rd Annual meeting of the cognitive science society (CogSci 2021). Curran Associates Inc. <http://arxiv.org/pdf/2002.01335v1>
  187. Slowik, A., Gupta, A., Hamilton, W.L., Jamnik, M., Holden, S.B. (2020) Towards graph representation learning in emergent communication. In: Association for the Advancement of Artificial (ed.) Workshop

- on reinforcement learning in games at AAAI-20. [https://rlg.mlanctot.info/papers/AAAI20-RLG\\_paper\\_27.pdf](https://rlg.mlanctot.info/papers/AAAI20-RLG_paper_27.pdf)
188. Steinert-Threlkeld, S. (2018). Paying attention to function words. In: 2nd Workshop on emergent communication at NeurIPS 2018. <http://arxiv.org/pdf/1909.11060v1>
  189. Tucker, M., Li, H., Agrawal, S., Hughes, D., Sycara, K., Lewis, M., Shah, J. (2021). Emergent discrete communication in semantic spaces. In: Neural Information Processing Systems Foundation (ed.) Advances in neural information processing systems 34. Advances in neural information processing systems. Curran Associates Inc. [https://proceedings.neurips.cc/paper\\_files/paper/2021/file/5812f92450ccaf17275500841c70924a-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2021/file/5812f92450ccaf17275500841c70924a-Paper.pdf)
  190. Tucker, M., Shah, J., Levy, R., Zaslavsky, N. (2022). Towards human-agent communication via the information bottleneck principle. In: Robotics Science and Systems (ed.) Social intelligence in humans and robots workshop at RSS 2022. <http://arxiv.org/pdf/2207.00088v1>
  191. Ueda, R., Taniguchi, T. (2024). Lewis's signaling game as beta-vae for natural word lengths and segments. In: OpenReview.net (ed.) 12th International conference on learning representations. <https://openreview.net/forum?id=HC0msxE3sf>
  192. Unger, T.A., Bruni, E. (2020). Generalizing emergent communication. <http://arxiv.org/pdf/2001.01772v3>
  193. van der Wal, O., Boer, S., Bruni, E., Hupkes, D. (2020). The grammar of emergent languages. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on empirical methods in natural language processing (EMNLP), pp. 3339–3359. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.270>
  194. Vani, A., Schwarzer, M., Lu, Y., Dhekane, E., Courville, A. (2021). Iterated learning for emergent systematicity in vqa. In: OpenReview.net (ed.) 9th International conference on learning representations. [https://openreview.net/forum?id=PD\\_oMxH8IIF](https://openreview.net/forum?id=PD_oMxH8IIF)
  195. Verma, S., Dhar, J. (2020). Emergence of writing systems through multi-agent cooperation. In: Proceedings of the 34th AAAI Conference on artificial intelligence (AAAI-20), pp. 13941–13942. <https://doi.org/10.1609/aaai.v34i10.7243>
  196. Villanger, J.I.F., Bojesen, T.A. (2024). An inductive bias for emergent communication in a continuous setting. In: Lutchyn, T., Ramirez Rivera, A., Ricaud, B. (eds.) Proceedings of the 5th northern lights deep learning conference (NLDL). Proceedings of Machine Learning Research, vol. 233, pp. 235–243. PMLR, Online. <https://proceedings.mlr.press/v233/villanger24a.html>
  197. Xu, Z., Niethammer, M., Raffel, C. (2022). Compositional generalization in unsupervised compositional representation learning: A study on disentanglement and emergent language. In: Neural Information Processing Systems Foundation (ed.) Advances in neural information processing systems 35. Advances in neural information processing systems. Curran Associates Inc. [https://papers.neurips.cc/paper\\_files/paper/2022/file/9f9ecbf4062842df17ec3f4ea3ad7f54-Paper-Conference.pdf](https://papers.neurips.cc/paper_files/paper/2022/file/9f9ecbf4062842df17ec3f4ea3ad7f54-Paper-Conference.pdf)
  198. Yu, H., Shen, W., Huang, L., Yuan, C. (2023). Manipulating multi-agent navigation task via emergent communications. In: 2023 IEEE 9th International conference on cloud computing and intelligent systems (CCIS), pp. 351–355. IEEE. <https://doi.org/10.1109/CCIS59572.2023.10262852>
  199. Yuan, L., Fu, Z., Shen, J., Xu, L., Shen, J., Zhu, S.-C. (2019). Emergence of pragmatics from referential game between theory of mind agents. In: 3rd Workshop on emergent communication at NeurIPS 2019 (2019). <http://arxiv.org/pdf/2001.07752v2>
  200. Ampatzis, C., Tuci, E., Trianni, V., & Dorigo, M. (2008). Evolution of signaling in a multi-robot system: Categorization and communication. *Adaptive Behavior*, 16(1), 5–26. <https://doi.org/10.1177/1059712307087282>
  201. Bachwerk, M., Vogel, C. (2011). Establishing linguistic conventions in task-oriented primeval dialogue. In: Esposito, A., Vinciarelli, A., Vicsi, K., Pelachaud, C., Nijholt, A. (eds.) Analysis of verbal and nonverbal communication and enactment. The Processing Issues. Lecture Notes in Computer Science, vol. 6800, pp. 48–55. Springer Berlin Heidelberg. [https://doi.org/10.1007/978-3-642-25775-9\\_4](https://doi.org/10.1007/978-3-642-25775-9_4)
  202. Bouchacourt, D., Baroni, M. (2019). Miss tools and mr fruit: Emergent communication in agents learning about object affordances. In: Association for Computational Linguistics (ed.) Proceedings of the 57th Annual meeting of the association for computational linguistics, pp. 3909–3918. <https://doi.org/10.18653/v1/P19-1380>
  203. Evtimova, K., Drozdov, A., Kiela, D., Cho, K. (2018). Emergent communication in a multi-modal, multi-step referential game. In: OpenReview.net (ed.) 6th International conference on learning representations: Conference track proceedings. <https://openreview.net/forum?id=rJGZq6g0->
  204. Hagiwara, Y., Kobayashi, H., Taniguchi, A., & Taniguchi, T. (2019). Symbol emergence as an inter-personal multimodal categorization. *Frontiers in Robotics and AI*, 6, 134. <https://doi.org/10.3389/frobt.2019.00134>



205. Kolb, B., Lang, L., Bartsch, H., Gansekoele, A., Koopmanschap, R., Romor, L., Speck, D., Mul, M., Bruni, E. (2019). Learning to request guidance in emergent language. In: Association for Computational Linguistics (ed.) Proceedings of the beyond vision and language: In tegrating real-world knowledge (LANTERN), pp. 41–50. <https://doi.org/10.18653/v1/D19-6407>
206. Lee, J., Cho, K., Weston, J., Kiela, D. (2018). Emergent translation in multi-agent communication. In: OpenReview.net (ed.) 6th International conference on learning representations: Conference track proceedings. <https://openreview.net/forum?id=H1vEXAxA->
207. Patel, S., Wani, S., Jain, U., Schwing, A., Lazebnik, S., Savva, M., Chang, A.X. (2021) Interpretation of emergent communication in heterogeneous collaborative embodied agents. In: 2021 IEEE/CVF International conference on computer vision. IEEE. <https://doi.org/10.1109/ICCV48922.2021.01565>
208. Röpke, W., Roijers, D. M., Nowé, A., & Rafdulescu, R. (2022). Preference communication in multi-objective normal-form games. *Neural Computing and Applications*. <https://doi.org/10.1007/s00521-022-07533-6>
209. Saha, H., Venkataraman, V., Speranzon, A., Sarkar, S. (2019). A perspective on multi-agent communication for information fusion. In: Neural Information Processing Systems Foundation (ed.) 3rd Workshop on visually grounded interaction and language at NeurIPS 2019. <http://arxiv.org/pdf/1911.03743v1>
210. Yuan, L., Fu, Z., Zhou, L., Yang, K., Zhu, S.-C. (2019). Emergence of theory of mind collaboration in multiagent systems. In: 3rd Workshop on emergent communication at NeurIPS 2019. <http://arxiv.org/pdf/2110.00121v1>
211. Bullard, K., Meier, F., Kiela, D., Pineau, J., Foerster, J. (2020). Exploring Zero-Shot Emergent Communication in Embodied Multi-Agent Populations. <http://arxiv.org/pdf/2010.15896v2>
212. Fitzgerald, N. (2019). To populate is to regulate. In: 3rd Workshop on emergent communication at NeurIPS 2019. <http://arxiv.org/pdf/1911.04362v1>
213. Kajić, I., Aygün, E., Precup, D. (2020). Learning to cooperate: Emergent communication in multi-agent navigation. In: Cognitive Science Society (ed.) 42nd Annual meeting of the cognitive science society (CogSci 2020). Curran Associates Inc. <http://arxiv.org/pdf/2004.01097v2>
214. Nevens, J., van Eecke, P., Beuls, K. (2019). A practical guide to studying emergent communication through grounded language games. In: Society for the study of artificial intelligence and the simulation of behaviour (AISB) annual convention 2019. <http://arxiv.org/pdf/2004.09218v1>
215. Sirota, J., Bulitko, V., Brown, M.R.G., Hernandez, S.P. (2019). Evolving recurrent neural networks for emergent communication. In: Proceedings of the genetic and evolutionary computation conference companion. GECCO '19, pp. 189–190. Association for Computing Machinery. <https://doi.org/10.1145/3319619.3321957>
216. Tieleman, O., Lazaridou, A., Mourad, S., Blundell, C., Precup, D. (2019). Shaping representations through communication: community size effect in artificial learning systems. In: Neural Information Processing Systems Foundation (ed.) 3rd Workshop on visually grounded interaction and language at NeurIPS 2019. <http://arxiv.org/pdf/1912.06208v1>
217. Gupta, S., Dukkupati, A. (2020). Winning an election: On emergent strategic communication in multi-agent networks. In: Proceedings of the 19th International conference on autonomous agents and multiagent systems. AAMAS '20, pp. 1861–1863. International Foundation for Autonomous Agents and Multiagent Systems. <https://www.ifaamas.org/Proceedings/aamas2020/pdfs/p1861.pdf>
218. Thomas, J.D., Santos-Rodríguez, R., Piechocki, R., Anca, M. (2021). Multi-lingual agents through multi-headed neural networks. In: 2nd Workshop on cooperative AI at NeurIPS 2021. <http://arxiv.org/pdf/2111.11129v1>
219. Baronchelli, A., Dall'Asta, L., Barrat, A., & Loreto, V. (2006). Strategies for fast convergence in semiotic dynamics. In L. M. Rocha (Ed.), *Artificial life X. A Bradford book* (pp. 480–485). MIT Press.
220. Botoko Ekila, J. (2024). Emergence of linguistic conventions in multi-agent systems through situated communicative interactions. In: Proceedings of the 23rd international conference on autonomous agents and multiagent systems. AAMAS '24, pp. 2725–2727. International Foundation for Autonomous Agents and Multiagent Systems. <https://doi.org/10.5555/3635637.3663267>
221. Botoko Ekila, J., Nevens, J., Verheyen, L., Beuls, K., van Eecke, P. (2024). 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 multiagent systems. AAMAS '24, pp. 2168–2170. International Foundation for Autonomous Agents and Multiagent Systems. <https://www.ifaamas.org/Proceedings/aamas2024/pdfs/p2168.pdf>
222. Das, A., Gervet, T., Romoff, J., Batra, D., Parikh, D., Rabbat, M., Pineau, J. (2019). Tarmac: Targeted multi-agent communication. In: Proceedings of the 36th international conference on machine learning (ICML 2019). <https://proceedings.mlr.press/v97/das19a/das19a.pdf>

223. Hildreth, D., & Guy, S. J. (2019). Coordinating multi-agent navigation by learning communication. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*. <https://doi.org/10.1145/3340261>
224. Loreto, V., Gravino, P., Servedio, V. D. P., & Tria, F. (2016). On the emergence of syntactic structures: Quantifying and modelling duality of patterning. *Topics in Cognitive Science*, 8(2), 469–480. <https://doi.org/10.1111/tops.12193>. 1602.03661v1.
225. Lorkiewicz, W., Kowalczyk, R., Katarzyniak, R., Vo, Q.B. (2011). On topic selection strategies in multi-agent naming game. In: The 10th International conference on autonomous agents and multiagent systems—Volume 2. AAMAS '11, pp. 499–506. International Foundation for Autonomous Agents and Multiagent Systems. <https://doi.org/10.5555/2031678.2031688>
226. Simoes, D., Lau, N., & Reis, L. P. (2020). Multi-agent actor centralized-critic with communication. *Neurocomputing*, 390, 40–56. <https://doi.org/10.1016/j.neucom.2020.01.079>
227. Simoes, D., Lau, N., & Reis, L. P. (2020). Exploring communication protocols and centralized critics in multi-agent deep learning. *Integrated Computer-aided Engineering*, 27(4), 333–351. <https://doi.org/10.3233/ICA-200631>
228. Taylor, J., Nisioti, E., Moulin-Frier, C. (2022). Socially supervised representation learning: the role of subjectivity in learning efficient representations. In: Proceedings of the 21st international conference on autonomous agents and multiagent systems (AAMAS 2022). <https://www.ifaamas.org/Proceedings/aamas2022/pdfs/p1274.pdf>
229. Wang, T., Wang, J., Zheng, C., Zhang, C. (2020) Learning nearly decomposable value functions via communication minimization. In: OpenReview.net (ed.) 8th International conference on learning representations. <https://openreview.net/forum?id=HJx-3grYDB>
230. Eccles, T., Bachrach, Y., Lever, G., Lazaridou, A., Graepel, T. (2019). Biases for emergent communication in multi-agent reinforcement learning. In: Neural Information Processing Systems Foundation (ed.) Advances in neural information processing systems 32. Advances in neural information processing systems. Curran Associates Inc. [https://proceedings.neurips.cc/paper\\_files/paper/2019/file/fe5e7cb609bdeb6d62449d61849c38b0-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2019/file/fe5e7cb609bdeb6d62449d61849c38b0-Paper.pdf)
231. Gupta, S., Hazra, R., Dukkupati, A. (2020). Networked multi-agent reinforcement learning with emergent communication. In: Proceedings of the 19th international conference on autonomous agents and multiagent systems. AAMAS '20, pp. 1858–1860. International Foundation for Autonomous Agents and Multiagent Systems. <https://aamas.csc.liv.ac.uk/Proceedings/aamas2020/pdfs/p1858.pdf>
232. Jimenez Romero, C., Yegenoglu, A., Pérez Martin, A., Diaz-Pier, S., & Morrison, A. (2023). Emergent communication enhances foraging behavior in evolved swarms controlled by spiking neural networks. *Swarm Intelligence*. <https://doi.org/10.1007/s11721-023-00231-6>
233. Karten, S., Tucker, M., Li, H., Kailas, S., Lewis, M., & Sycara, K. (2023). Interpretable learned emergent communication for human-agent teams. *IEEE Transactions on Cognitive and Developmental Systems*, 15(4), 1801–1811. <https://doi.org/10.1109/TCDS.2023.3236599>
234. Li, S., Zhou, Y., Allen, R., Kochenderfer, M.J. (2022). Learning emergent discrete message communication for cooperative reinforcement learning. In: 2022 International conference on robotics and automation (ICRA), pp. 5511–5517. IEEE. <https://doi.org/10.1109/ICRA46639.2022.9812285>
235. Lipinski, O., Sobey, A.J., Cerutti, F., Norman, T.J. (2022). Emergent password signalling in the game of werewolf. In: 5th Workshop on emergent communication at ICLR 2022. <https://openreview.net/forum?id=B4xM-Qb0mbq>
236. Pesce, E., & Montana, G. (2020). Improving coordination in small-scale multi-agent deep reinforcement learning through memory-driven communication. *Machine Learning*, 109(9–10), 1727–1747. <https://doi.org/10.1007/s10994-019-05864-5>
237. Resnick, C., Kulikov, I., Cho, K., Weston, J. (2017). Vehicle communication strategies for simulated highway driving. In: 1st Workshop on emergent communication at NeurIPS 2017. <http://arxiv.org/pdf/1804.07178v2>
238. Wu, J., Sun, X., Zeng, A., Song, S., Rusinkiewicz, S., Funkhouser, T. (2021). Spatial intention maps for multi-agent mobile manipulation. In: IEEE international conference on robotics and automation. <https://doi.org/10.1109/ICRA48506.2021.9561359>
239. Yuan, L., Chen, F., Zhang, Z., & Yu, Y. (2024). Communication-robust multi-agent learning by adaptable auxiliary multi-agent adversary generation. *Frontiers of Computer Science*. <https://doi.org/10.1007/s11704-023-2733-5>
240. Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P.A., Strouse, D.J., Leibo, J.Z., Freitas, N. d.: Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In: Proceedings of the 36th international conference on machine learning (ICML 2019). <https://proceedings.mlr.press/v97/jaques19a/jaques19a.pdf>

241. Abdel-Aziz, M. K., Elbamby, M. S., Samarakoon, S., & Bennis, M. (2024). Cooperative multi-agent learning for navigation via structured state abstraction. *IEEE Transactions on Communications*. <https://doi.org/10.1109/TCOMM.2024.3365520>
242. Foerster, J., Chen, R.Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., Mordatch, I. (2018). Learning with opponent-learning awareness. In: Proceedings of the 17th international conference on autonomous agents and multiagent systems. AAMAS '18, pp. 122–130. International Foundation for Autonomous Agents and Multiagent Systems. <https://ifaamas.org/Proceedings/aamas2018/pdfs/p122.pdf>
243. Nakamura, T., Taniguchi, A., Taniguchi, T. (2023). Control as probabilistic inference as an emergent communication mechanism in multi-agent reinforcement learning. <http://arxiv.org/pdf/2307.05004v1>
244. Zheng, L., Yang, J., Cai, H., Zhang, W., Wang, J., Yu, Y. (2018) Magent: A many-agent reinforcement learning platform for artificial collective intelligence. In: Association for the Advancement of Artificial Intelligence (ed.) Proceedings of the thirty-second AAAI conference on artificial intelligence and thirtieth innovative applications of artificial intelligence conference and eighth AAAI symposium on educational advances in artificial intelligence. AAAI Press. <https://doi.org/10.1609/aaai.v32i1.11371>
245. Lee, J., Cho, K., Kiela, D. (2019) Countering language drift via visual grounding. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp. 4385–4395. Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1447>
246. Clark, H. H. (2012). *Using Language*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511620539>
247. Hockett, C. F. (1960). The origin of speech. *Scientific American*, 203, 89–96.
248. Bard, N., Foerster, J. N., Chandar, S., Burch, N., Lanctot, M., Song, H. F., Parisotto, E., Dumoulin, V., Moitra, S., Hughes, E., Dunning, I., Mourad, S., Larochelle, H., Bellemare, M. G., & Bowling, M. (2020). The hanabi challenge: A new frontier for ai research. *Artificial Intelligence*, 280(1–2), 103216. <https://doi.org/10.1016/j.artint.2019.103216>
249. Lipowska, D., & Lipowski, A. (2018). Emergence of linguistic conventions in multi-agent reinforcement learning. *PloS One*. <https://doi.org/10.1371/journal.pone.0208095>
250. Park, H. H., Zhang, K. J., Haley, C., Steimel, K., Liu, H., & Schwartz, L. (2021). Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9, 261–276. [https://doi.org/10.1162/tac1\\_a\\_00365](https://doi.org/10.1162/tac1_a_00365)
251. Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 4. <https://doi.org/10.3389/neuro.06.004.2008>
252. Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1), 72. <https://doi.org/10.2307/1412159>
253. Brighton, H., & Kirby, S. (2006). Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artificial Life*, 12(2), 229–242. <https://doi.org/10.1162/artl.2006.12.2.229>
254. Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System Technical Journal*, 29 (2), 147–160. <https://doi.org/10.1002/j.1538-7305.1950.tb00463.x>
255. Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10, 707.
256. Hu, H., Lerer, A., Peysakhovich, A., Foerster, J. (2020). "other-play" for zero-shot coordination. In: Proceedings of the 37th international conference on machine learning (ICML 2020). <https://proceedings.mlr.press/v119/hu20a/hu20a.pdf>
257. Dubova, M., Moskvichev, A. (eds.): (2020). Effects of supervision, population size, and self-play on multi-agent reinforcement learning to communicate. Artificial life conference proceedings, vol. ALIFE 2020: The 2020 conference on artificial life. [https://doi.org/10.1162/isal\\_a\\_00328](https://doi.org/10.1162/isal_a_00328)
258. Hermann, K.M., Hill, F., Green, S., Wang, F., Faulkner, R., Soyer, H., Szepesvari, D., Czarnecki, W.M., Jaderberg, M., Teplyashin, D., Wainwright, M., Apps, C., Hassabis, D., Blunsom, P. (2017). Grounded language learning in a simulated 3D world. <http://arxiv.org/pdf/1706.06551v2>
259. Wang, Z., Cai, S., Chen, G., Liu, A., Ma, X., Liang, Y. (2023). Describe, explain, plan and select: Interactive planning with llms enables open-world multi-task agents. In: Neural Information Processing Systems Foundation (ed.) Advances in neural information processing systems 36. Advances in neural information processing systems. <https://openreview.net/forum?id=KtvPdGb31Z>
260. Wang, Z., Cai, S., Liu, A., Jin, Y., Hou, J., Zhang, B., Lin, H., He, Z., Zheng, Z., Yang, Y., Ma, X., Liang, Y. (2023). JARVIS-1: Open-world multi-task agents with memory-augmented multimodal language models. <http://arxiv.org/pdf/2311.05997v3>
261. Wenqi Zhang, Ke Tang, Hai Wu, Mengna Wang, Yongliang Shen, Guiyang Hou, Zeqi Tan, Peng Li, Yueting Zhuang, Weiming Lu (2024) Agent-pro: Learning to evolve via policy-level reflection and

- optimization. In: Association for Computational Linguistics (ed.) Proceedings of the 62nd annual meeting of the association for computational linguistics. Annual Meeting of the Association for Computational Linguistics, pp. 5348–5375 <https://doi.org/10.18653/V1/2024.ACL-LONG.292>
262. Zhao, A., Huang, D., Xu, Q., Lin, M., Liu, Y.-J., & Huang, G. (2024). Expel: Llm agents are experiential learners. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17), 19632–19642. <https://doi.org/10.1609/aaai.v38i17.29936>
  263. Nottingham, K., Ammanabrolu, P., Suhr, A., Choi, Y., Hajishirzi, H., Singh, S., Fox, R. (2023). Do embodied agents dream of pixelated sheep?: Embodied decision making using language guided world modelling. In: Workshop on reincarnating reinforcement learning at ICLR 2023. <http://arxiv.org/pdf/2301.12050v2>
  264. Sharma, A., Rao, Sudha and Brockett, Chris and Malhotra, Akanksha and Jojic, Nebojsa and Dolan, Bill (2024) Investigating agency of llms in human-ai collaboration tasks. In: Graham, Y., Purver, M. (eds.) Proceedings of the 18th conference of the european chapter of the association for computational linguistics (Volume 1: Long Papers), pp. 1968–1987. Association for Computational Linguistics. <https://aclanthology.org/2024.eacl-long.119>
  265. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*. <https://doi.org/10.1007/s11704-024-40231-1>
  266. Malladi, S., Gao, T., Nichani, E., Damian, A., Lee, J.D., Chen, D., Arora, S. (2023) Fine-tuning language models with just forward passes. In: Neural Information Processing Systems Foundation (ed.) Advances in neural information processing systems 36. Advances in neural information processing systems. <https://openreview.net/forum?id=Vota6rFhBQ>
  267. Chen, W., Su, Y., Zuo, J., Yang, C., Yuan, C., Chan, C.-M., Yu, H., Lu, Y., Hung, Y.-H., Qian, C., Qin, Y., Cong, X., Xie, R., Liu, Z., Sun, M., Zhou, J. (2024). Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In: OpenReview.net (ed.) 12th international conference on learning representations. <https://openreview.net/forum?id=EHg5GDnyq1>
  268. Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N.V., Wiest, O., Zhang, X. (2024). Large language model based multi-agents: A survey of progress and challenges. In: Kate Larson (ed.) Proceedings of the thirty-third international joint conference on artificial intelligence, pp. 8048–8057. <https://doi.org/10.24963/ijcai.2024/890>
  269. Theodore, S., Shunyu, Y., Narasimhan, K., Griffiths, T. (2024). Cognitive architectures for language agents. Transactions on Machine Learning Research.
  270. Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. In: Neural Information Processing Systems Foundation (ed.) Advances in neural information processing systems 36. Advances in neural information processing systems. [https://papers.nips.cc/paper\\_files/paper/2023/file/1b44b878bb782e6954cd888628510e90-Paper-Conference.pdf](https://papers.nips.cc/paper_files/paper/2023/file/1b44b878bb782e6954cd888628510e90-Paper-Conference.pdf)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Authors and Affiliations

Jannik Peters<sup>1</sup>  · Constantin Waubert de Puiseau<sup>1</sup>  · Hasan Tercan<sup>1</sup>  ·  
Arya Gopikrishnan<sup>2</sup>  · Gustavo Adolpho Lucas de Carvalho<sup>3</sup>  ·  
Christian Bitter<sup>1</sup>  · Tobias Meisen<sup>1</sup> 

✉ Jannik Peters  
jpeters@uni-wuppertal.de

Constantin Waubert de Puiseau  
waubert@uni-wuppertal.de

Hasan Tercan  
tercan@uni-wuppertal.de

Arya Gopikrishnan  
ag3974@drexel.edu

Gustavo Adolpho Lucas de Carvalho  
lucasdec@usc.edu

Christian Bitter  
bitter@uni-wuppertal.de

Tobias Meisen  
meisen@uni-wuppertal.de

<sup>1</sup> Institute of Technologies and Management of the Digital Transformation, University of Wuppertal, Rainer-Gruenter-Str. 21, 42119 Wuppertal, NRW, Germany

<sup>2</sup> College of Engineering, Drexel University, 3141 Chestnut St, Philadelphia, PA 19104, USA

<sup>3</sup> Department of Computer Science, University of Southern California, 941 Bloom Walk, Los Angeles, CA 90089, USA