IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 19, 2021, accepted February 22, 2021, date of publication February 24, 2021, date of current version March 5, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3061756

When Language Evolution Meets Multimodality: Current Status and Challenges Toward Multimodal Computational Models

PATRIZIA GRIFONI, (Member, IEEE), ARIANNA D'ULIZIA, AND FERNANDO FERRI
Consiglio Nazionale delle Ricerche – IRPPS, 00185 Rome, Italy

Corresponding author: Arianna D'Ulizia (arianna.dulizia@irpps.cnr.it)

ABSTRACT Computational models can be considered human-designed computing models inspired by the processes observed in the natural world, which allow simulating and understanding these processes. Computational modelling is notably applied to simulate the behaviour and long-term dynamics of human Language. The research effort made so far in computational modelling of language evolution considers predominantly one modality by arguing for a unimodal origin of Language. This article extends this paradigm to a new perspective that integrates into its structure and learning algorithms principles from multimodal communication. This article gives an overview of the current language evolution models. It discusses the key challenges towards multimodal language evolution modelling by envisioning a conceptual framework to design the multimodal grounding and the language learning processes, as well as their realisation through a multi-agent multimodal referential game. This framework is valuable for many researchers working on language evolution to reveal the key questions they should address and integrate for pursuing a holistic vision that combines all modalities in a multimodal language evolution model.

INDEX TERMS Natural languages, multimodality, computational modeling, agent-based modeling, language evolution.

I. INTRODUCTION

Language characterises human communication and it continuously evolves under the influence of many factors such as environmental and cultural issues. More than one-and-a-half century ago, Charles Darwin in ''The Descent of Man'' [1] observed ''curious parallels'' between the formation and development of Language and species evolution. He noticed that both languages and biological species change continuously and evolve following similar mechanisms, such as transmission from parents to children, selection, and adaptation. The questions of the origin and evolution of human Language are continuing to fascinate many disciplines, experiencing a renewed and growing interest in evolutionary theories of language change.

To investigate and simulate the behaviour and long-term dynamics of human Language, many of these language evolution theories apply computational modelling [2], [3] by spawning a great deal of literature on computational models of language evolution until now. Indeed, human Language is a complex and non-linear dynamic system [4], and computational models can be used to cope with the complexity and dynamism of its evolution. The main reason is that computational modelling allows a simulation using theoretical models and the results to be compared with empirical observations by allowing to validate the theory [5]. The computational climate change models give an example in meteorology. The earth's atmosphere and the oceanic masses constitute a complex dynamic system, which can be better understood with the help of computational modelling. Analogously, several linguistic phenomena, such as lexicon emergence, syntax acquisition, symbol grounding (i.e. how words get their meanings), emergence of compositionality (i.e. the property of systematically deriving the meaning of composite expressions from the meanings of their parts and how these parts are combined), etc., are hard to be explained and deeply understood without the use of computational language evolution models.

The major strand of these language evolution models and theories investigates language evolution under the lens of *natural Language* (i.e. human written or spoken languages as opposed to artificially constructed languages). From the historical point of view, indeed, ''*language research has focused predominantly on speech and/or text, thus ignoring*

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott.

the wealth of additional information available in face-to-face communication'' [6]. Indeed, many studies address the language evolution problem using a linguistic approach focused on speech and/or text. Further studies, mainly focused on non-human primate communication, address language evolution by analysing gestural modality. Gestural evolution studies emphasise the role of gestures, as it is supposed to be the first modality with which the human Language began. Therefore, the research effort on language evolution models made so far is based on two counterposed theories that see the origins of Language in a spoken/vocal rather than in a gestural/visual communicative system. Both these theories focus predominantly on one modality by arguing for a unimodal origin of Language.

Most recent researches in language evolution [6]–[11] start to emphasise the multimodal nature of language and, in particular, the relevance o*f multimodality* for human language evolution. Recent studies make it evident that ''*speech and gesture are part and parcel of the same system and together constitute a tightly integrated processing unit, thus underscoring the need for a multimodal approach to the study of language*'' [6].

As multimodality is progressively becoming the key to developing more realistic language evolution models, the idea of developing a multimodal language evolution model comes within reach. In this article, we discuss our work on multimodal language evolution modelling by envisioning a conceptual framework and the fundamental challenges to designing the multimodal grounding and the language learning processes, as well as their realisation through a multi-agent multimodal referential game. The choice of an agent-based approach is because it is best suited to model the behaviours and interactions of individual behavioural entities. Therefore it fits well to explain how particular language systems may emerge and evolve in a population of behavioural entities. Moreover, agent-based simulations using empirical data would make the language evolution model considerably more realistic. Not surprisingly this is the only approach applied unimodally for all the three modalities analysed in this paper, i.e. speech, gesture, and visual. The conceptual framework envisioned here is valuable for many researchers working on language evolution to reveal the key questions they should address and integrate for pursuing an integrated vision that combines all modalities in a multimodal language evolution model. First, we give an overview of the research conducted so far in computational modelling of language evolution.

II. COMPUTATIONAL MODELS OF LANGUAGE EVOLUTION SO FAR

Computational models are mathematical models ''created to simulate a set of processes observed in the natural world to gain an understanding of these processes'' [12]. In language evolution, they were built to investigate and simulate the emergence and evolution of a communication system at different linguistic levels (mainly phonological, lexical, syntactic, and semantic). Important works on computational models for language evolution were already developed in the seventies and eighties [13], [14] by following a unimodal perspective, according to which Language expressed in a single modality was studied.

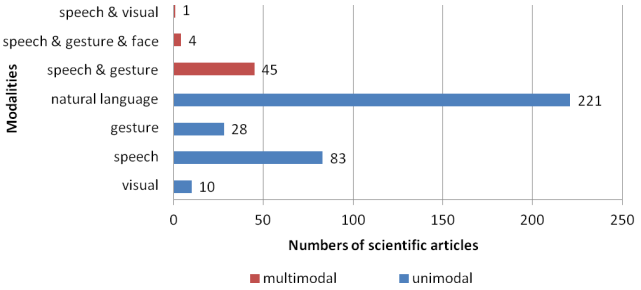

FIGURE 1. The modalities considered in the 388 examined studies.

To gain an accurate overview of the current state of the art on language evolution theories and models, we conducted a systematic search for scientific papers published from 1980 to 2019 (end of January) using two relevant search engines (Web of Science and Scopus) by including the following keywords in the search: ''language evolution'' OR ''evolution of language'' AND ''computational model*''. Only peer-reviewed articles written in English were included in the analysis.

A total of 714 articles were returned from Web of Science and Scopus by using these search keywords; respectively 462 from Web of Science and 252 from Scopus. 10 articles from Scopus were excluded because they were duplicate publications or whole conference proceedings. The resulting set of 242 articles from Scopus and the 462 from Web of Science had an overlap of 70 articles. Therefore, a total of 634 papers were examined (see the supplemental material S1 for the full list). Reading in detail the content of the papers, we further reduced this set to 388 papers by including only those that discuss language evolution theories or models (see the supplemental material S2 for the full list). For each article, we examined the modalities considered (vocal, gestural, facial, etc.) if it explicitly discusses multimodal issues of the Language, if it proposes/uses a computational model.

The results of this examination show that the majority of the articles (87.1% - 338) on language evolution theories and models consider only one single modality, while only 12.9% of studies (50) consider multiple integrated modalities (see Fig. 1). Specifically, the most frequently integrated modalities are speech and gesture with 11.6% of studies (45), speech, gesture, and face with 1% (4), and speech and visual with 0.3% (1). Considering the computational models, no articles dealing with a multimodal language proposes/uses a computational model, while 116 papers (30%) dealing with a unimodal language propose/use a computational model (see the supplemental material S3 for the full list). The computational models applied in these 116 papers are shown in Fig. 2(a). We can observe that the majority of the papers (about 55%) were based on agents, followed by game-theoretic models with 19 papers (about 16.5%) and machine learning models with 16 papers (about 14%).

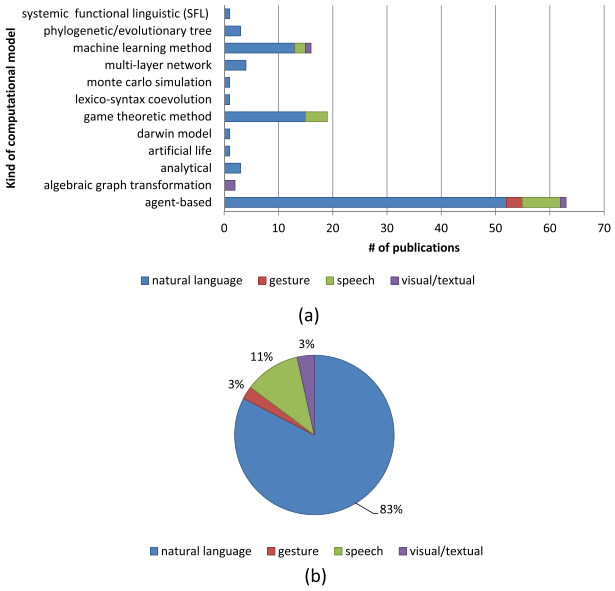

FIGURE 2. **(a) The computational models across modalities considered in the 116 examined studies. (b) The percentage of articles proposing/using a computational model across modalities.**

The remaining computational models have a scientific production that ranges from 1 to 4 articles (with an average of about 2 articles per model). Note that several papers apply more than one computational method for modeling language evolution. There was a significant difference in the number of articles proposing/using a computational model across modalities with natural Language accounting for 95 studies (82.6%), speech 13 (11.3%), visual/textual 4 (3.5%), and gesture 3 (2.6%) (see Fig. 2(b)). Natural language papers are not tied to one specific modality since natural Language can take different forms (such as speech, gesture, or text). Therefore, we have considered a computational model of natural language evolution as unimodal.

In the following sections, we discuss the main computational models developed for each of the three main modalities (speech, gesture, and visual/textual) investigated in the literature. In this discussion, we do not include computational models of natural language evolution because they do not explicitly consider the modality in which the Language is expressed. In this article, we are interested in focusing on the different modalities used. A comprehensive survey on computational models of natural language evolution can be found in Grifoni *et al.* [9] and D'Ulizia *et al.* [15].

A. COMPUTATIONAL MODELS OF SPEECH EVOLUTION

Computational models have been extensively applied to study spoken language evolution from the seventies to now [13], [16]. These models are mainly devoted to simulating the production of speech linked to s and their ancestors. A review of the different approaches to modelling speech evolution has been provided by de Boer [17], which distinguished between direct modelling techniques and computational techniques.

The former reconstructs the vocal tract of fossil hominids, mainly Neanderthals. Examples of direct modelling techniques are the reconstructed Neanderthal vocal tract [13] and the growth model [16]. The reconstructed Neanderthal vocal tract was simulated by using a fossil skull of a Neanderthal man and estimating the acoustic capabilities of his articulatory model. The growth model simulates the articulatory and acoustic vocal tract of human beings by predicting the acoustic space according to the different values of the larynx position and, consequently, inferring the maximal acoustic space used by Neanderthals. The main drawback of these techniques relies on the limited evolutionary period that can be reconstructed that is only the stage of evolution from available fossil data.

The latter uses roughly the following three computational methods to investigate speech evolution: agent-based methods, genetic algorithms, and game theory.

Agent-based methods allow the emergence of a language to be represented in a bottom-up fashion as a result of the interaction of a group of agents. These agents represent humans with different kinds of cognitive and social behavior that interact with each other in a population. Each agent is equipped with linguistic abilities for conceptualisation, production, parsing, and interpretation. Moreover, they are endowed with learning mechanisms that allow expanding their basic linguistic knowledge. Agent-based methods are naturally adopted by those who want to develop real-world applications, such as software agents or robots evolving shared communication systems [18], [19].

Genetic algorithms [20] are applied to evolve the linguistic system by considering a population of potential solutions that are recombined, crossover, and mutate. Some examples of applications of genetic algorithms in speech evolution models can be found in [21] and [22]. Warlaumont and Olney [21] use genetic algorithms integrated with a neural network model for evolving the signaler and receiver neural network connection weights. De Pauw [22] proposes a language evolution model that applies genetic algorithms to define the content and quality of the grammars that are being developed over time by imposing fitness functions on the agents' population. Their resemblance to real biological evolution is the most significant advantage of genetic algorithms [2]. A further advantage is described by Landsbergen [23] that says "the use of evolutionary theory in linguistic research also allows for a strong quantitative, mathematical view on language change". The negative aspect is the necessity of a large number of design decisions that have to be made in building a genetic algorithm: what to encode as genes, how to implement the fitness function, etc.

Game theory in speech evolution has led to the definition of a general framework, in which there are two players, the sender (or teacher) and the receiver (or learner), that apply a set of rules of the game to define how the interactions are structured and what information is exchanged. Application of game theory to the study of the evolution of spoken Language is investigated by various authors, such as Jäger [24], Nowak *et al.* [25], Mitchener [26], and Benz *et al.* [27]. Mitchener [26] proposes a communication game that is

specified by a payoff matrix whose elements represent the probabilities that a speaker (using his/her own grammar) understands the sentence spoken by the other speakers (using their own grammars). The most significant advantage of game-theoretic methods relies on the possibility to reuse the rich body of results established by game theorists. The main shortcoming of these methods is their problematic use due to conceptual confusion and empirical deficiencies, as emerged from Watumull and Hauser [28].

B. COMPUTATIONAL MODELS OF GESTURE EVOLUTION

Computational models of gesture evolution are far less studied compared to speech evolution. Moreover, the most relevant studies on gestural evolution have discussed theories of the gestural origin of the human linguistic capacity, without providing a systematisation of these theories through a computational model. Hewes [29], Rizzolatti and Arbib [30], Corballis [31], Armstrong and Wilcox [32], Tomasello [33] are the authors of some of the most relevant theories arguing that human Language began as a gestural communication system.

Some computational models for investigating the evolution of gestural movements are proposed in [34]–[36]. Gestural movements refer here to spontaneous movements of the body, including head, hands, arms, and legs. These studies reproduce the gestural behaviour of humans by using mainly agent-based simulation. They rely mainly on two theoretical strands: the mirroring mechanisms of monkeys and the emergence of sign languages in deaf communities.

The former is based on neurophysiological studies [37]–[39] that have discovered, first in monkeys and subsequently in humans, a class of nerve cells, called mirror neurons, activated by both execution and observation of motor actions (such as the gestures of grasping, breaking, or tearing). The computational models based on these neurophysiological studies investigate the evolution of imitative behaviours by applying evolutionary adaptive agents equipped with neural ''mirror'' mechanisms analogous to those found in biological systems [34], [35].

The latter relies on socio-linguistic studies [40], [41] that investigate the spontaneous emergence of gestural communication systems in deaf individuals not exposed to spoken/signed languages or in deaf communities. The computational models based on this theoretical foundation simulate the conventionalisation process arising from sign use in a population of adaptive agents who adjust their signed Language in response to other agents' sign use, by using multi-agent reinforcement learning models [36].

C. COMPUTATIONAL MODELS OF VISUAL/TEXTUAL EVOLUTION

Visual/textual languages refer to languages that are expressed through a graphical notation, such as visual programming languages, visual query languages, etc.. Computational models of visual/textual evolution rely mainly on graph-based methods that allow representing linguistic units (e.g. words, sentences, etc.) as nodes in a graph and relations between them as edges. Algebraic graph transformation represents an evolution of the abstract syntax of these linguistic units. This formalism allows all kinds of transformations that range from language definition by grammars via model migration to language integration by extending the grammar rule set and/or the vocabulary [42]. Very few studies investigated the use of iterated learning [43] as a modelling approach to the evolution of graphical symbols. Iterated learning is an agent-based technique that is mainly concerned with the transmission of Language between successive generations of agents. Garrod *et al.* [43] in their study apply the game ''Pictionary'' for investigating the emergence of graphical communication systems and their evolution through a process of interactive grounding.

III. THE WAY FORWARD: A MULTIMODAL PERSPECTIVE

Current computational models of language evolution have focused predominantly on unimodal Language (speech, gesture, or text), by ignoring additional information available in multimodal communication. Small but significant researches have been conducted in the last years, highlighting the importance of abandoning traditional distinctions among modalities in language evolution research and pursuing, instead, an integrated vision that combines all modalities in a multimodal language evolution model.

Vigliocco *et al.* [6] dwelt on the distinction between language proper and communication. They argued that the majority of language studies have relied upon language proper as they produced abstract and symbolic linguistic systems that do not consider the broader context of language use (included the channels of information) characterising, on the contrary, the communication. They provided some evidence suggesting that information coming from ''non-verbal communication'', including gestures and facial expressions, is an integral part of the same linguistic system, thus underscoring the need for a multimodal approach to language study. Gillespie-Lynch *et al.* [44] reviewed several studies that support a multimodal theory of language evolution. Among them, various studies on non-human primates [45], [46] are investigated demonstrating that non-human primates have the capacity to integrate information across multiple sensory modalities and to produce a multimodal output mainly through gestures and vocalisations. Therefore, Gillespie-Lynch *et al.* [44] concluded that Language has a multimodal origin in both phylogeny and ontogeny. Slocombe *et al.* [46] analysed unimodal (i.e. either vocal, gestural, or facial) theories of language evolution in primate communication and highlighted their weaknesses. Consequently, they suggested pursuing an integrated, multimodal approach to primate communication by avoiding the traditional modality distinction and leading to the development of new theories of language evolution. Taglialatela *et al.* [45] examined the neural metabolic activity in the chimpanzee brain and observed that the Broca's region is activated by vocal signaling in conjunction with manual communicative

gestures by demonstrating a communicative behavior very similar to human beings. Their results pointed to a multimodal origin for human Language. Levinson and Holler [10] argued that the development of modern human communication derives from a sequential accumulation of layers of different communication abilities (e.g. ad hoc gesture, vocalisation, indexicality, iconicity, speech, etc.). They concluded that spoken and gestural languages have evolved together as one integrated multimodal communication system. Wacewicz and Zywiczynski [47] also support this; they argued that the rise of a multimodal approach is a natural consequence of the progression from theoretical to empirical work, as the available data are predominantly multimodal. A further study supported the redefinition of the evolutionary theories of Language starting from conceiving Language as a multimodal phenomenon has been proposed by Perniss [48]. The author provides several motivations, supported by empirical studies, which justify the multimodal nature of Language. Similarly, Waller *et al.* [11] tried to answer the question "why adopting a multimodal approach?" to the study of primate communication. They listed several advantages of a multimodal approach over the dominant, unimodal approach: (i) a more complete picture of the different functioning of the modalities; (ii) a better understanding of the characteristics of the communication deriving from an integrated (instead of isolated) analysis of the composite signal; and (iii) increase in the repertoire of signals from which it is possible to extract important patterns that would be overlooked in the case of independent repertoires of the isolated signals.

Therefore, it is clear that numerous researches pursue a multimodal approach to language evolution but none devoted to developing a computational model of multimodal language evolution. In the following section, we discuss the computational challenges for a multimodal approach to language evolution modelling that can contribute to filling the hole in our current knowledge of human language evolution.

IV. COMPUTATIONAL CHALLENGES FOR A MULTIMODAL LANGUAGE EVOLUTION MODEL

Considering more than one modality in the language evolution, modelling produces several computational challenges that lead to rethinking the current unimodal models.

We considered three reference models (one for each of the three analysed modalities) to take a picture of current unimodal models. They are Moulin-Frier's model [19] for speech, Richie *et al.*'s model [36] for gesture, and Garrod *et al.*'s model [43] for visual modality. The choice of these three unimodal models is because they rely on the same computational method (i.e. agent-based), and that allows more easy comparison and integration in a multimodal model. As shown in Fig. 2(a), indeed, the agent-based method is the only approach defined for all three modalities. From these reference models, we extracted all the elements (e.g. involved actors, background knowledge, exchanged messages, performed actions, produced output) of the models in order to depict a conceptual framework, as shown in Fig. 3(a), 3(b), and 3(c). All these models consider a sender agent that aims to communicate to a receiver agent about an object/meaning that is chosen among a set of potential objects/meanings. The set of objects represents the world that can be communicated and each object has a meaning included in the set of meanings. Each agent can have the role of sender and receiver. It has to be equipped with sensory input and control output capabilities, different for each one of the three reference models depending on the modality.

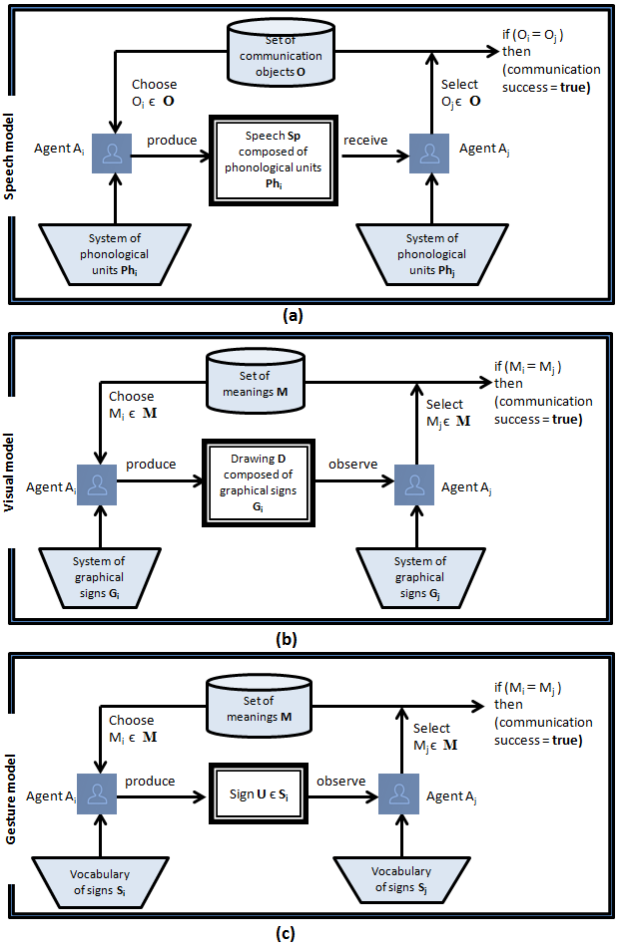

FIGURE 3. Conceptual frameworks derived from the Moulin-Frier's speech model (a), Garrod *et al.*'s visual model (b), and Richie *et al.*'s gesture model (c).

Specifically, in the Moulin-Frier's model, the speaker agents A_i are equipped with a system of phonological units Ph_i that allows producing the speech Sp corresponding to the chosen communication object O_i belonging to the set of all possible communication objects O of the external world. The sender agents A_i of the Garrod *et al.*'s model [43] are equipped with a system of graphical signs G_i that the agents take to produce the drawing D that better represents the chosen meaning M_i belonging to the set of all possible meanings M. Analogously, in Richie *et al.*'s gesture model [36] the sender agents A_i are equipped with a vocabulary of gestural homesigns S_i that are used to produce the sign U corresponding to the chosen meaning M_i.

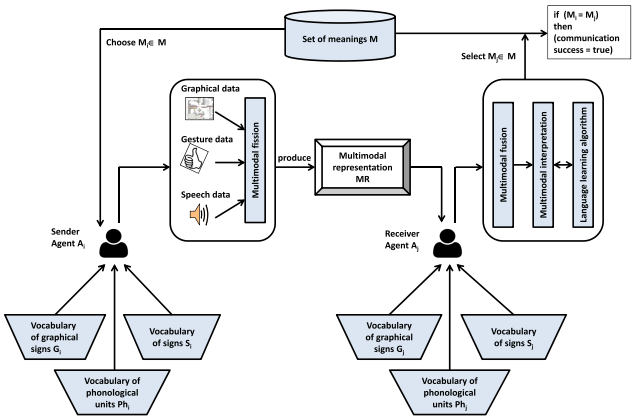

FIGURE 4. A conceptual framework for a multimodal language evolution model.

In all the three models, the receiver agents A_j listen to/observe the received spoken, drawn, or sign message and select the object O_j/meaning M_j among the set of potential objects O/meanings M that better interprets the received message according to their systems of phonological units Ph_j (in the Moulin-Frier's model), graphical signs G_j (the Garrod *et al.*'s model), or gestural homesigns S_j (in Richie *et al.*'s model). If the receiver interprets the message correctly, i.e. it extracts the sender's target object from the set of possible objects (i.e. $M_i = M_j$ or $O_i = O_j$), the game ends with a success. Otherwise, the game ends in failure.

All these models use algorithms that allow evolving and learning communication structures starting from unimodal input data. Therefore, the messages exchanged by the agents are expressed using one modality only. Moreover, in each model the agents (both the sender and the receiver) are endowed with pre-established systems of phonological units, graphical signs, or gestural homesigns. This knowledge is used to transform the chosen meaning into a meaningful message and parse the message to reconstruct its overall meaning. For this purpose, the agents apply a learning process for developing and continuously adapting their Language according to the unimodal messages received during the interactive experiences.

To further a multimodal approach to language evolution modelling, several challenges need to be addressed to integrate/change the necessary elements of the current unimodal models into a coherent multimodal conceptual framework (see Fig. 4). In the following sub-sections, we try to unravel these challenges underlying multimodal utterance processing.

A. A SEMANTIC FRAMEWORK FOR CONCEPTUALISATION

First of all, to set up the multimodal agent-based conceptual framework depicted in Fig. 4, the first challenge that needs to be addressed is the representation of the meanings of the various objects that constitute the world in which the agents act. This representation implies a formalisation of the semantics at the base of a successful (i.e. unambiguous and consistent) communication between agents. A semantic formalisation that allows assessing semantic relations between concepts is fundamental for modeling generalisation capabilities in a population of agents [70].

As well explained by Steels [71], the steps of the emergence of semantic meanings in a population of artificial agents, which are equipped as described in the three reference models depicted in Fig. 3, can be summarised in the following four: (i) *inferred*: sender and receiver agents try to come to a common understanding of an object without explicitly expressing all meanings but inferring most of them from the contextual information; (ii) *lexical*: a meaning fragment becomes associated with an object without any concern for grammar; (iii) *syntactic*: meaning is expressed using hierarchical patterns of symbols that can be combined using syntactic rules; (iv) *morphological:* meaning is expressed using complex symbols, hierarchical structure and semantic relations between symbols organised through grammatical agreement.

Most of the models of language evolution developed in the literature focused predominantly on the lexical and syntactic level, by shedding light on the emergence of phonological systems [17]–[19], lexical development [14], [36], the emergence of syntax [42], [43], and, more specifically, compositionality [62], [64], [72], [73]. The major problem of these models stays in the fact that the emergent phonological/ lexical/syntactic (compositional) systems arise only because the symbols are explicitly coupled with pre-existing, structured semantic meanings; how these meanings are originated and why they are associated with the symbols remains an open problem in these models.

To answer these questions, several approaches to conceptualisation and semantic formalisation of concepts have been put forward in the literature on natural language [68], [69], [75]. However, dealing with multimodal languages requires the definition of semantic frameworks to deal with and interpret multimodal information. In this direction, recent developments have been made to fill the gap of grounding language semantics through multimodal sensory perception [70]. These approaches are strictly linked to the fourth challenge related to the multimodal language learning process since the emergence of semantic communication systems involves a continuous learning process through interaction with the world, producing the language evolution.

Therefore, the definition of a semantic framework for representing and manipulating multimodal concepts as well as approaches for grounding multimodal language semantics must be tackled by researchers and remains an open problem for multimodal language evolution modelling.

B. MULTIMODAL MESSAGES

The messages exchanged between the agents should be multimodal. Therefore, each agent has to be endowed with the vocabularies of phonological units, graphical signs, and gestural homesigns. Moreover, the agents have to learn and infer a cross-modal mapping between the set of possible

objects/meanings and the multimodal messages allowing to describe these objects efficiently by using the symbols in the vocabularies [49]–[51].

The sender agent should be provided with speech generation, graphical sign, and gestural homesign production capabilities. Analogously, the receiver agent should be endowed with speech, graphical sign, and gestural homesign perception capabilities. Moreover, a method for fusing/combining modalities and generating multimodal representations should be defined. Based on current literature, three classes of strategies can be applied to this aim: a posteriori combination (late fusion), simultaneous learning (early fusion), and hybrid composition [52]. The a posteriori combination strategy integrates different modalities at the decision level and focuses predominantly on modeling interactions occurring among the same kind of modality rather than cross-modality interactions. The early fusion refers to integrating multimodal data at the input level. Unlike the previous one, this strategy is not very suitable to learn intra-modality interactions. Finally, hybrid composition performs fusion between the input and decision levels, by allowing to model both intra-modality and cross-modality interactions.

Considering our context in which the agents have to learn and infer a cross-modal mapping among modalities, the most suitable strategy is the hybrid composition. However, the multimodal integration in this kind of interaction context is quite complex since the composition of multiple signals is not always predictable and conventionalised. Moreover, the different signals could be not aligned, congruent, or redundant but each one conveying semantic and pragmatic information to the message. Several researchers have addressed this challenge in the literature, and the most promising solutions have been achieved by enriching the hybrid composition strategy with word-level alignment [7], [53], [54], attention mechanisms [53], [55]–[58], tensor techniques [59], [60], and multilayered approaches [61].

Word-level alignment consists of aligning each segment of a modality to corresponding segments of the other modalities. Gu *et al.* [53] applied both word-level alignment and hierarchical attention mechanisms for aligning the text and audio at the word-level and a convolutional neural network structure to combine word-level features. Dumpala *et al.* [54] explored cross-modal autoencoders for audio-visual alignment, which try to reconstruct the representations corresponding to a single modality, using the encoded representation of the available input modalities. Ferri *et al.* [7] proposed a grammatical approach in which the segments of each modality are linearised according to the kinds of cross-modality interaction and their syntactic roles. The main benefit of the word-level alignment approach is that it allows examining exactly what the model is learning at a finer resolution.

Attention mechanisms control the importance of each modality in the embedding process to handle the dynamic dependencies among them. Liang *et al.* [55] modelled cross-modal interactions using a Recurrent Multistage Fusion Network (RMFN) and attention mechanisms, which allow to automatically decompose the multimodal fusion problem into multiple recursive stages. Tsai *et al.* [56] also explored cross-modal attention mechanisms and proposed a multimodal transformer for modeling unaligned multimodal language sequences following a late fusion strategy. They merge multimodal information via a feed-forward fusion process by directly attending to low-level features in other modalities. Wang *et al.* [57] used a recurrent embedding network to model the fine-grained structure of nonverbal cues at the subword level and built multimodal-shifted word representations that dynamically capture the variations of different nonverbal contexts. Zadeh *et al.* [58] proposed a transformer model for multimodal sequential learning capable of modelling asynchronous intramodal and intermodal dynamics within one single transformer network.

Tensor techniques compute the outer product between unimodal representations from different modalities to compute a tensor representation to model inter-modality interactions. Zadeh *et al.* [59] proposed a tensor fusion network based on Cartesian-products to model unimodal, bimodal, and trimodal interactions. Liu *et al.* [60] approximated this network by proposing a low-rank model. Such methods suffer from exponentially increasing computational complexity when the dimensionality of the tensor representations increases.

Finally, Holler and Levinson [61] proposed a multimodal psycholinguistic framework to unravel the complexities of the multimodal dialogue, which is based on a multilayered approach that considers a segregation layer for filtering out the signals not relevant to the message and a binding layer for integrating multiple, asynchronous signals.

Therefore, the definition of a multimodal fusion method that enables agents to learn and infer a cross-modal mapping among modalities in order to communicate using multimodal messages must be tackled by researchers. The methods described above can help to overcome the issues of dynamicity, asynchronism, and unalignment of the elements from the different modalities.

C. MULTIMODAL REFERENTIAL GAMES

To study the emergence of linguistic phenomena in a simplified and realistic way, the three reference models depicted in Fig. 3 apply communication games among agents. This setting, where agents invent a communication protocol that lets them succeed in a given collaborative task, was introduced around the early 2000s [25] with the intent of better understanding principles guiding the evolution and emergence of natural languages. More recently, with the advent of deep learning techniques, several studies in the literature analysed the emergence of communication through cooperative multiagent referential games [62]–[66]. In these games, communities of deep neural agents cooperate to successfully solve the game via communicating about some perceptual input. Specifically, the agents (a sender and a receiver) are placed in a simple environment where they develop their Language interactively out of the need to coordinate and communicate. The sender chooses a target object (typically an image,

e.g. an apple) and communicates what it sees (e.g. a red and smooth fruit) to the receiver using its vocabulary. On the other hand, the receiver is tasked to interpret the message of the sender by figuring out what is the target object.

Havrylov and Titov [62] improved the basic version of the game by considering messages exchanged by the agents in the form of a language (i.e. variable-length sequences of discrete symbols) instead of atomic categories in order to improve the similarity to natural language. Similarly, Lazaridou *et al.* [64] considered agents that are able to learn from realistic images (e.g. raw pixel data) instead of symbolic input (e.g. attribute-based or one-hot vectors) in order to improve the similarity to the raw sensorimotor data, which humans are exposed to. Evtimova *et al.* [63] extended the basic version of the game by considering a bidirectional exchange of messages with symmetric communication abilities of the agents, as well as the capabilities of bridging different modalities (the sender is grounded in the visual modality and the receiver in textual modality). Graesser *et al.* [65], in addition to the symmetry of the communication (i.e the agents should be able to act both as sender and receiver), proposed two more properties of the referential games: externality, i.e the agents should communicate about something external to themselves; and partial observability, i.e. the environment is not all observable and the communication is essential for solving the game. Dagan *et al.* [66] introduced a language transmission bottleneck in the basic referential game, where new agents have to learn the language by playing with more experienced agents and have to overcome such bottleneck through mechanisms of cultural evolution of language and genetic evolution of agents.

All these proposed settings of the multi-agent referential game consider unimodal messages exchanged by the agents. Evtimova *et al.* [63] introduced the necessity of bridging different modalities, however, each agent is grounded with one modality only. In a multimodal approach to language evolution modelling, the challenge will be to design a setting where the collaborative task is a multimodal referential game. In other words, the referential game should be designed to cope with multimodal messages and each agent should be grounded with multiple modalities.

D. MULTIMODAL LANGUAGE LEARNING PROCESS

In the three unimodal reference models depicted in Fig. 3, as a result of the communicative interactions expected by the referential game, the agents adapt the vocabularies, the pre-established perceptive and sensory systems, as well as the Language, following a learning process. Different learning algorithms have been developed in language learning literature in multi-agent systems [62]–[65]. These learning algorithms can be grouped roughly into two main classes: evolutionary computation-based algorithms and gradient-based optimisation algorithms. The former simulates the evolution of imitative complex behaviours by applying evolutionary adaptive agents [74], while the latter applied deep learning techniques for training deep neural networks with a supervised or reinforcement learning approach. These algorithms allow the agents to learn the rules of interaction, the concepts, and the mappings between objects/meanings and unimodal messages. Specifically, Lazaridou *et al.* [64] applied a reinforcement learning approach, in which two agents take discrete actions in their environment to maximise a shared reward. The authors applied the policy gradient method to learn agents involved in a referential game and implemented the sender and receiver agents as a single-layer LSTM (Long short-term memory) using prelinguistic feed-forward encoders and decoders. Similar to Lazaridou *et al.*, Havrylov and Titov [62] also formulated the learning process as a reinforcement learning problem; unlike them, however, they used sequences of tokens rather than atomic symbols and used straight-through Gumbel-softmax estimators that allow an end-to-end differentiation. Reinforcement learning was also applied by Evtimova *et al.* [63] by using a single-layer feed-forward network for the attention-based sender and a single hidden-layer recurrent neural network for the receiver agent. Graesser *et al.* [65] used a hybrid of supervised and reinforcement learning: the former for training two predictive distributions before and after message exchange and the latter for maximising the reward.

All these proposed learning algorithms rely on a specific modality and should undergo a transformation toward the new perspective of multimodality. More than ten years ago, Kasabov [67] envisaged a principle of multiple modality learning that states that adaptive learning systems have to learn different but related information modalities. Allowing agents to conceptualise complex multimodal linguistic structures, reconstruct the correct meanings when interpreting these structures and adjust the conceptualisation and interpretation processes according to the communicative interactions remains an open problem for multimodal language evolution modelling. For this purpose, specific multimodal language learning mechanisms must be developed and optimised.

V. CONCLUSION

Over the past half-century, several language evolution theories investigating the behaviour and long-term dynamics of human Language have been developed by spawning a great deal of literature on computational models of language evolution. They focus predominantly on one modality by arguing for a unimodal investigation of language evolution. This study has sought to illustrate how these unimodal language evolution models are conceived and the key computational challenges necessary to overcome the limitations of the traditional modality distinction-based approaches by pursuing an integrated vision that combines all modalities in a multimodal approach.

Several studies in the literature believed and demonstrated that integrating multimodal information into language evolution models improves the understanding of multimodal grounding and language learning processes. Developing multimodal communicative structures, multimodal multi-agents referential games, and multimodal language learning

TABLE 1. Computational challenges for a multimodal language evolution model and possible solutions.

Open challenges	Description	Possible Solutions
A semantic framework for conceptualisation	To define a framework for representing and manipulating meanings as well as mechanisms for grounding multimodal language semantics	To adapt the mechanisms for conceptualisation and semantic formalisation of concepts [68-69] and for grounding language semantics through multimodal sensory perception [70].
Multimodal messages	The agents have to learn and infer a cross-modal mapping among modalities	Hybrid composition [52] integrated with: 1) Word-level alignment [53-54][7] 2) Attention mechanisms [53][55-58] 3) Tensor techniques [59-60] 4) Multilayered approaches [61]
Multimodal referential games	To design a setting where the agents perform collaborative tasks in a multimodal referential game, which is suitable to cope with multimodal messages and each agent is grounded with multiple modalities	To adapt cooperative multi-agent referential games by using: 1) variable-length sequences of discrete symbols [62] 2) Raw pixel data [64] 3) Symmetric communication abilities [63] 4) language transmission bottleneck [66]
Multimodal language learning process	To conceptualise complex multimodal linguistic structures and to reconstruct the correct meanings when interpreting these structures, but also to adjust the conceptualisation and interpretation processes according to the communicative interactions	To adapt language learning algorithms in multi-agent systems: 1) evolutionary computation-based algorithms [21-22] 2) gradient-based optimisation algorithms: supervised or reinforcement learning [62-65]

mechanisms could contribute to better characterise how human beings have learned to react to multimodal stimuli by developing and evolving their linguistic capabilities. A summary of the open challenges toward a multimodal approach to language evolution modelling is provided in Table 1, along with a brief description of the challenge and the possible solutions envisaged in the paper.

Conceiving multimodal language evolution models has important implications in many domains. To date, we have focused on the evolutionary linguistic field, however, these kinds of models can be applied in question answering systems, mechatronics, automotive systems, robotic systems, and, in general, in all domains that need model-driven mechanisms supporting language evolution.

REFERENCES

[1] C. Darwin, *The Expression of the Emotions in Man and Animals*. London, U.K.: J. Murray, 1872.
[2] B. de Boer, "Computer modelling as a tool for understanding language evolution," in *Evolutionary Epistemology, Language and Culture*. Dordrecht, The Netherlands: Springer, 2006, pp. 381–406, doi: 10.1007/1-4020-3395-8_17.
[3] M. H. Christiansen and S. Kirby, "Language evolution: Consensus and controversies," *Trends Cogn. Sci.*, vol. 7, no. 7, pp. 300–307, Jul. 2003.
[4] L. Steels, "The synthetic modelling of language origins," *Evol. Commun. J*, vol. 1, no. 1, pp. 1–34, 1999.
[5] P. Vogt, "Modeling interactions between language evolution and demography," *Hum. Biol.*, vol. 81, nos. 2–3, pp. 237–258, Apr. 2009, doi: 10.3378/027.081.0307.
[6] G. Vigliocco, P. Perniss, and D. Vinson, "Language as a multimodal phenomenon: Implications for language learning, processing and evolution," *Phil. Trans. Roy. Soc. B, Biol. Sci.*, vol. 369, no. 1651, Sep. 2014, Art. no. 20130292.
[7] F. Ferri, A. D'Ulizia, and P. Grifoni, "A grammar inference approach for language self-adaptation and evolution in digital ecosystems," *J. Intell. Inf. Syst.*, vol. 53, no. 3, pp. 409–430, Dec. 2019, doi: 10.1007/s10844-019-00566-9.
[8] F. Ferri, A. D'Ulizia, and P. Grifoni, "Computational models of language evolution: Challenges and future perspectives," *J. Universal Comput. Sci.*, vol. 24, no. 10, pp. 1345–1377, 2018.
[9] P. Grifoni, A. D'Ulizia, and F. Ferri, "Computational methods and grammars in language evolution: A survey," *Artif. Intell. Rev.*, vol. 45, no. 3, pp. 369–403, Mar. 2016, doi: 10.1007/s10462-015-9449-3.
[10] S. C. Levinson and J. Holler, "The origin of human multi-modal communication," *Phil. Trans. Roy. Soc. B, Biol. Sci.*, vol. 369, no. 1651, 2014, Art. no. 20130302.
[11] B. Waller, K. Liebal, A. M. Burrows, and K. Slocombe, "How can a multimodal approach to primate communication help us understand the evolution of communication?" *Evol. Psychol.*, vol. 11, no. 3, pp. 538–549, 2013.
[12] L. L. Hill, S. J. Crosier, T. R. Smith, and M. Goodchild, "A content standard for computational models," *D-Lib Mag.*, vol. 7, no. 6, pp. 1082–9873, Jun. 2001, doi: 10.1045/june2001-hill.
[13] P. Lieberman and E. S. Crelin, "On the speech of Neanderthal man," *Linguistic Inquiry*, vol. 2, no. 2, pp. 203–222, 1971.
[14] J. R. Hurford, "Biological evolution of the Saussurean sign as a component of the language acquisition device," *Lingua*, vol. 77, no. 2, pp. 187–222, Feb. 1989.
[15] A. D'Ulizia, F. Ferri, and P. Grifoni, "A survey on modeling language evolution in the new millennium," *New Gener. Comput.*, vol. 38, no. 1, pp. 97–124, Mar. 2020, doi: 10.1007/s00354-019-00079-7.
[16] L.-J. Boë, J.-L. Heim, K. Honda, and S. Maeda, "The potential Neandertal vowel space was as large as that of modern humans," *J. Phonetics*, vol. 30, no. 3, pp. 465–484, Jul. 2002, doi: 10.1006/jpho.2002.0170.
[17] B. de Boer, "Evolving sound systems," in *Simulating the Evolution of Language*, A. Cangelosi and D. Parisi, Eds. Berlin, Germany: Springer-Verlag, 2002, pp. 79–97.
[18] B. de Boer and W. Zuidema, "Multi-agent simulations of the evolution of combinatorial phonology," *Adapt. Behav.*, vol. 18, no. 2, pp. 141–154, Apr. 2010, doi: 10.1177/1059712309345789.
[19] C. Moulin-Frier, J. Diard, J.-L. Schwartz, and P. Bessière, "COSMO ('communicating about objects using sensory–motor operations'): A Bayesian modeling framework for studying speech communication and the emergence of phonological systems," *J. Phonetics*, vol. 53, pp. 5–41, Nov. 2015, doi: 10.1016/j.wocn.2015.06.001.
[20] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA, USA: Addison-Wesley, 1998.
[21] A. S. Warlaumont and A. M. Olney, "Evolution of reflexive signals using a realistic vocal tract model," *Adapt. Behav.*, vol. 23, no. 4, pp. 183–205, Aug. 2015, doi: 10.1177/1059712315585941.
[22] G. De Pauw, "Evolutionary computing as a tool for grammar development," in *Proc. GECCO*, in Lecture Notes in Computer Science, Chicago, IL, USA, vol. 2723. Berlin, Germany: Springer-Verlag, Jul. 2003, pp. 549–560.
[23] F. Landsbergen, "Cultural evolutionary modeling of patterns in language change: Exercises in evolutionary linguistics," Ph.D. dissertation, LOT, Netherlands Graduate School Linguistics, Utrecht, The Netherlands, 2009.
[24] G. Jäger, "Evolutionary game theory and typology: A case study," *Language*, vol. 83, no. 1, pp. 74–109, 2007, doi: 10.1353/lan.2007.0020.

[25] M. A. Nowak, J. B. Plotkin, and D. C. Krakauer, "The evolutionary language game," *J. Theor. Biol.*, vol. 200, no. 2, pp. 147–162, 1999.

[26] W. G. Mitchener, "Game dynamics with learning and evolution of universal grammar," *Bull. Math. Biol.*, vol. 69, no. 3, pp. 1093–1118, Mar. 2007, doi: 10.1007/s11538-006-9165-x.

[27] A. Benz, C. Ebert, G. Jäger, and R. van Rooij, "Language, games, and evolution: An introduction," in *Language, Games, and Evolution* (Lecture Notes in Computer Science), vol. 6207. Berlin, Germany: Springer, 2011.

[28] J. Watumull and M. D. Hauser, "Conceptual and empirical problems with game theoretic approaches to language evolution," *Frontiers Psychol.*, vol. 5, p. 226, Mar. 2014, doi: 10.3389/fpsyg.2014.00226.

[29] G. W. Hewes, R. J. Andrew, L. Carini, H. Choe, R. A. Gardner, A. Kortlandt, G. S. Krantz, G. McBride, F. Nottebohm, J. Pfeiffer, D. G. Rumbaugh, H. D. Steklis, M. J. Raliegh, R. Stopa, A. Suzuki, S. L. Washburn, and R. W. Wescott, "Primate communication and the gestural origin of language [and comments and reply]," *Current Anthropol.*, vol. 14, nos. 1–2, pp. 5–24, Feb. 1973, doi: 10.1086/201401.

[30] G. Rizzolatti and M. A. Arbib, "Language within our grasp," *Trends Neurosci.*, vol. 21, no. 5, pp. 188–194, May 1998, doi: 10.1016/S0166-2236(98)01260-0.

[31] M. C. Corballis, *From Hand to Mouth: The Origins of Language*. Princeton, NJ, USA: Princeton Univ. Press, 2002.

[32] D. F. Armstrong and S. E. Wilcox, *The Gestural Origin of Language*. New York, NY, USA: Oxford Univ. Press, 2007.

[33] M. Tomasello, *Origins of Human Communication*. Cambridge, MA, USA: MIT Press, 2008.

[34] E. Borenstein and E. Ruppin, "The evolution of imitation and mirror neurons in adaptive agents," *Cogn. Syst. Res.*, vol. 6, no. 3, pp. 229–242, Sep. 2005, doi: 10.1016/j.cogsys.2004.11.004.

[35] E. Spaak, "From imitation to action understanding: On the evolution of mirror neurons," M.S. thesis, Donders Inst. Brain, Cognition Behav., Centre Cognition, Radboud Univ., Nijmegen, The Netherlands, 2008.

[36] R. Richie, C. Yang, and M. Coppola, "Modeling the emergence of lexicons in homesign systems," *Topics Cogn. Sci.*, vol. 6, no. 1, pp. 183–195, Jan. 2014, doi: 10.1111/tops.12076.

[37] G. di Pellegrino, L. Fadiga, L. Fogassi, V. Gallese, and G. Rizzolatti, "Understanding motor events: A neurophysiological study," *Exp. Brain Res.*, vol. 91, no. 1, pp. 176–180, Oct. 1992, doi: 10.1007/BF00230027.

[38] L. Fogassi and P. F. Ferrari, "Mirror neurons and the evolution of embodied language," *Current Directions Psychol. Sci.*, vol. 16, no. 3, pp. 136–141, Jun. 2007.

[39] L. Bonini and P. F. Ferrari, "Evolution of mirror systems: A simple mechanism for complex cognitive functions," *Ann. New York Acad. Sci.*, vol. 1225, no. 1, pp. 166–175, Apr. 2011, doi: 10.1111/j.1749-6632.2011.06002.x.

[40] I. Meir, W. Sandler, C. Padden, and M. Aronoff, "Emerging sign languages," in *Oxford Handbook of Deaf Studies, Language, and Education*, vol. 2, M. Marschark and P. Spencer, Eds. Oxford, U.K.: Oxford Univ. Press, 2010, pp. 267–279.

[41] D. Brentari and M. Coppola, "What sign language creation teaches us about language," *WIREs Cogn. Sci.*, vol. 4, no. 2, pp. 201–212, Mar. 2013, doi: 10.1002/wcs.1212.

[42] B. Braatz and C. Brandt, "A framework for families of domain-specific modelling languages," *Softw. Syst. Model.*, vol. 13, no. 1, pp. 109–132, Feb. 2014, doi: 10.1007/s10270-012-0271-y.

[43] S. Garrod, N. Fay, S. Rogers, B. Walker, and N. Swoboda, "Can iterated learning explain the emergence of graphical symbols?" *Exp. Semiotics*, vol. 11, no. 1, pp. 33–50, Mar. 2010, doi: 10.1075/is.11.1.04gar.

[44] K. Gillespie-Lynch, P. M. Greenfield, H. Lyn, and S. Savage-Rumbaugh, "Gestural and symbolic development among apes and humans: Support for a multimodal theory of language evolution," *Frontiers Psychol.*, vol. 5, p. 1228, Oct. 2014.

[45] J. P. Taglialatela, J. L. Russell, J. A. Schaeffer, and W. D. Hopkins, "Chimpanzee vocal signaling points to a multimodal origin of human language," *PLoS ONE*, vol. 6, no. 4, Apr. 2011, Art. no. e18852.

[46] K. E. Slocombe, B. M. Waller, and K. Liebal, "The language void: The need for multimodality in primate communication research," *Animal Behav.*, vol. 81, no. 5, pp. 919–924, May 2011, doi: 10.1016/j.anbehav.2011.02.002.

[47] S. Wacewicz and P. Zywiczynski, "The multimodal origins of linguistic communication," *Lang. Commun.*, vol. 54, pp. 1–8, May 2017.

[48] P. Perniss, "Why we should study multimodal language," *Frontiers Psychol.*, vol. 9, p. 1109, Jun. 2018.

[49] F. Ferri, A. D'Ulizia, and P. Grifoni, "Multimodal language specification for human adaptive mechatronics," 2017, *arXiv:1703.05616*. [Online]. Available: http://arxiv.org/abs/1703.05616

[50] A. D'Andrea, A. D'Ulizia, F. Ferri, and P. Grifoni, "EMAG: An extended multimodal attribute grammar for behavioural features," *Digit. Scholarship Hum.*, vol. 32, no. 2, pp. 251–275, 2015.

[51] A. D'Ulizia, F. Ferri, and P. Grifoni, "Toward the development of an integrative framework for multimodal dialogue processing," in *Proc. OTM Workshops*, in Lecture Notes in Computer Science, vol. 5333. Berlin, Germany: Springer, 2008, pp. 509–518.

[52] A. D'Ulizia, "Exploring multimodal input fusion strategies," in *Multimodal Human Computer Interaction and Pervasive Services*. Hershey, PA, USA: IGI Global, 2009, pp. 34–57.

[53] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proc. Conf. Assoc. Comput. Linguistics, Meeting*, 2018, p. 2225.

[54] S. H. Dumpala, I. Sheikh, R. Chakraborty, and S. K. Kopparapu, "Audio-visual fusion for sentiment classification using cross-modal autoencoder," in *Proc. NIPS*, 2019, pp. 1–4.

[55] P. Pu Liang, Z. Liu, A. Zadeh, and L.-P. Morency, "Multimodal language analysis with recurrent multistage fusion," 2018, *arXiv:1808.03920*. [Online]. Available: http://arxiv.org/abs/1808.03920

[56] Y.-H.-H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, p. 6558.

[57] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L. P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 7216–7223.

[58] A. Zadeh, C. Mao, K. Shi, Y. Zhang, P. Pu Liang, S. Poria, and L.-P. Morency, "Factorized multimodal transformer for multimodal sequential learning," 2019, *arXiv:1911.09826*. [Online]. Available: http://arxiv.org/abs/1911.09826

[59] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2017, pp. 1–12.

[60] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," 2018, *arXiv:1806.00064*. [Online]. Available: http://arxiv.org/abs/1806.00064

[61] J. Holler and S. C. Levinson, "Multimodal language processing in human communication," *Trends Cogn. Sci.*, vol. 23, no. 8, pp. 639–652, Aug. 2019, doi: 10.1016/j.tics.2019.05.006.

[62] S. Havrylov and I. Titov, "Emergence of language with multi-agent games: Learning to communicate with sequences of symbols," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2149–2159.

[63] K. Evtimova, A. Drozdov, D. Kiela, and K. Cho, "Emergent language in a multi-modal, multi-step referential game," *CoRR*, vol. abs/1705.10369, pp. 1–13, May 2017.

[64] A. Lazaridou, K. M. Hermann, K. Tuyls, and S. Clark, "Emergence of linguistic communication from referential games with symbolic and pixel input," 2018, *arXiv:1804.03984*. [Online]. Available: http://arxiv.org/abs/1804.03984

[65] L. Graesser, K. Cho, and D. Kiela, "Emergent linguistic phenomena in multi-agent communication games," 2019, *arXiv:1901.08706*. [Online]. Available: http://arxiv.org/abs/1901.08706

[66] G. Dagan, D. Hupkes, and E. Bruni, "Co-evolution of language and agents in referential games," 2020, *arXiv:2001.03361*. [Online]. Available: http://arxiv.org/abs/2001.03361

[67] N. Kasabov, "Integrative connectionist learning systems inspired by nature: Current models, future trends and challenges," *Natural Comput.*, vol. 8, no. 2, pp. 199–218, Jun. 2009, doi: 10.1007/s11047-008-9066-z.

[68] Y. Duan and C. Cruz, "Formalising semantic of natural language through conceptualization from existence," *Int. J. Innov., Manage. Technol.*, vol. 2, no. 1, p. 37, 2011.

[69] Y. Duan, "Towards a periodic table of conceptualization and formalization on state, style, structure, pattern, framework, architecture, service and so on," in *Proc. SNPD*, Jul. 2019, pp. 133–138.

[70] L. Beinborn, T. Botschen, and I. Gurevych, "Multimodal grounding for language processing," in *Proc. COLING*, 2018, pp. 2325–2339.

[71] L. Steels, "Agent-based models for the emergence and evolution of grammar," *Phil. Trans. Roy. Soc. B, Biol. Sci.*, vol. 371, no. 1701, Aug. 2016, Art. no. 20150447.

[72] H. Brighton, "Compositional syntax from cultural transmission," *Artif. Life*, vol. 8, no. 1, pp. 25–54, Jan. 2002.
[73] I. Mordatch and P. Abbeel, "Emergence of grounded compositional language in multi-agent populations," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2018, vol. 32, no. 1, pp. 1–8.
[74] M. C. Caschera, A. D'Ulizia, F. Ferri, and P. Grifoni, "Towards evolutionary multimodal interaction," in *Proc. OTM Confederated Int. Conf.* Berlin, Germany: Springer, Sep. 2012, pp. 608–616.
[75] Y. Duan, X. Sun, H. Che, C. Cao, Z. Li, and X. Yang, "Modeling data, information and knowledge for security protection of hybrid IoT and edge resources," *IEEE Access*, vol. 7, pp. 99161–99176, 2019, doi: 10.1109/ACCESS.2019.2931365.

PATRIZIA GRIFONI (Member, IEEE) received the degree in electronic engineering from the University of Rome La Sapienza, in 1990.

Since 1990, she has been a Senior Researcher with the Institute of Research on Population and Social Policies of the National Research Council of Italy. From 1993 to 2000, she has been a Contract Professor of image processing with the University of Macerata. She is currently a Professor of software engineering and object oriented programming with the Telematic International Uninettuno University. She is the author of more than 250 papers on journal, books, and conferences. She was the scientific responsible for CNR of many European and National projects. She is leading and participating in several national and international research projects. She has experience in organising several international conferences and scientific workshops, such as RRI-SIS2017, MONET, MIMIC, M2AS, and MAPS. Her main research interests include social informatics, social computing, human–machine interaction, multimodal interaction, sketch-based interfaces, multimedia applications, user modeling, data and knowledge bases, ontologies, machine learning, statistical databases, spatial information, geographic information systems, responsible research and innovation, ethics, risk management, digital ecosystems, web applications, Web 2.0, Internet of the future, and social networks.

Ms. Grifoni is a member of some Program Committees of International workshops and conferences. She is a member of the editorial board of some international journals. She is involved in the editorial board of some international journals. She has been a guest editor for several special issues of various international journals.

ARIANNA D'ULIZIA graduated in computer science engineering from the La Sapienza University, Rome, in 2005. She received the Ph.D. degree in computer science and automation from Roma Tre University, Rome, in 2009.

She is currently a Researcher with the Italian National Research Council Institute of Research on Population and Social Policies, Rome. Since 2005, she has participated in 19 European and national projects. She is the author of more than 70 papers in international journals, conferences, and books. Her research interests include human–computer interaction, multimodal interaction, visual languages, geographical query languages, social computing, risk governance, knowledge management, and innovation.

Dr. D'ulizia was a program chair for several international workshops and conferences. She is involved in the editorial board of several international journals. She has been a guest editor for several special issues of various international journals. Since 2016, she has been a Scientific Reviewer for several projects. She is part of many programming committees for national and international conferences and workshops.

FERNANDO FERRI graduated in electronic engineering, in 1990. He received the Ph.D. degree in medical informatics from La Sapienza University, Rome, in 1993.

From 1990 to 2001, he was a Researcher. From 1993 to 2000, he was a Lecturer of processing systems with the University of Macerata. Since 2001, he has been a Director of research with the Italian National Research Council. He is the author of more than 250 papers in international journals, books, and conference proceedings. He has coordinated and participated in several national and international research projects. He has organised several international events (scientific conferences and workshops). His main research interests include social informatics, social computing, data and knowledge bases, human–machine interaction, user–machine natural interaction, user modeling, visual interaction, sketch-based interfaces, geographic information systems, risk management, and medical informatics.

Dr. Ferri has been a guest editor for several special issues of various international journals.