Advanced Robotics

ISSN: 0169-1864 (Print) 1568-5535 (Online) Journal homepage: http://www.tandfonline.com/loi/tadr20

Symbol emergence in robotics: a survey

Tadahiro Taniguchi, Takayuki Nagai, Tomoaki Nakamura, Naoto Iwahashi, Tetsuya Ogata & Hideki Asoh

To cite this article: Tadahiro Taniguchi, Takayuki Nagai, Tomoaki Nakamura, Naoto Iwahashi, Tetsuya Ogata & Hideki Asoh (2016) Symbol emergence in robotics: a survey, Advanced Robotics, 30:11-12, 706-728, DOI: 10.1080/01691864.2016.1164622

To link to this article: https://doi.org/10.1080/01691864.2016.1164622

ADVANCED ROBOTICS, 2016
VOL. 30, NOS. 11–12, 706–728
http://dx.doi.org/10.1080/01691864.2016.1164622

SURVEY PAPER

Symbol emergence in robotics: a survey

Tadahiro Taniguchi[a], Takayuki Nagai[b], Tomoaki Nakamura[b], Naoto Iwahashi[c], Tetsuya Ogata[d] and Hideki Asoh[e]

[a]College of Information Science and Engineering, Ritsumeikan University, Kusatsu, Japan; [b]Department of Mechanical Engineering and Intelligent Systems, The University of Electro-Communications, Chofu-shi, Japan; [c]Faculty of Computer Science and Systems Engineering, Okayama Prefectural University, Okayama, Japan; [d]Department of Intermedia Art and Science, School of Fundamental Science and Engineering, Waseda University, Shinjuku, Japan; [e]Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan

ABSTRACT

Humans can learn a language through physical interaction with their environment and semiotic communication with other people. It is very important to obtain a computational understanding of how humans can form symbol systems and obtain semiotic skills through their autonomous mental development. Recently, many studies have been conducted regarding the construction of robotic systems and machine learning methods that can learn a language through embodied multimodal interaction with their environment and other systems. Understanding human?-social interactions and developing a robot that can smoothly communicate with human users in the long term require an understanding of the dynamics of symbol systems. The embodied cognition and social interaction of participants gradually alter a symbol system in a constructive manner. In this paper, we introduce a field of research called symbol emergence in robotics (SER). SER represents a constructive approach towards a symbol emergence system. The symbol emergence system is socially self-organized through both semiotic communications and physical interactions with autonomous cognitive developmental agents, i.e. humans and developmental robots. In this paper, specifically, we describe some state-of-art research topics concerning SER, such as multimodal categorization, word discovery, and double articulation analysis. They enable robots to discover words and their embodied meanings from raw sensory-motor information, including visual information, haptic information, auditory information, and acoustic speech signals, in a totally unsupervised manner. Finally, we suggest future directions for research in SER.

ARTICLE HISTORY
Received 19 September 2015
Revised 3 March 2016
Accepted 7 March 2016

KEYWORDS
Developmental robotics; language acquisition; semiotics; symbol emergence; symbol grounding

1. Introduction

The development of an intelligent robot with which people would embrace long-term interactions is one of the major challenges in the research field of robotics. Despite the rapid and remarkable progress in robotics, natural language processing, human–robot interaction, and related artificial intelligence technologies, we have not yet been able to develop such an autonomous robot. Even if an entertainment robot had sufficient capabilities for speech recognition, speech synthesis, and natural language processing, a user would find it monotonous if the robot behaved deterministically on the basis of finite hand-coded rules. To overcome this problem, a robot requires the ability for open-ended development through its physical interactions and semiotic communications. Moreover, such a robot must use symbol systems, including a language system, to communicate and collaborate with humans. To achieve actual long-term communication and collaboration between a human and a robot, the robot must be able to learn new vocabularies, understand the meanings of utterances, estimate a speaker's intention, and promote mutual understanding with people in the real world. To achieve a mutual understanding between a human and a robot, both of them must be able to infer 'what does he/she intend by the utterance?,' 'what does the word he/she uttered represent?,' and 'what should I say to make him/her understand what I want him/her to do?' by referring to many objects, events, contexts, situations, habits, and a history of interactions. When we consider human–human communication and collaboration in the real world, it is easily understood that utterances, e.g. words, phrases, and sentences, do not have one-to-one relationships with real-world phenomena. The modeling of such human-semiotic communication for use in engineering applications is important for developing a robot that can communicate and collaborate with others naturally like humans.

In general, a language system is considered as the representative of a symbol systems in semiotics. A symbol system is an important philosophical and technical

CONTACT Tadahiro Taniguchi taniguchi@ci.ritsumei.ac.jp

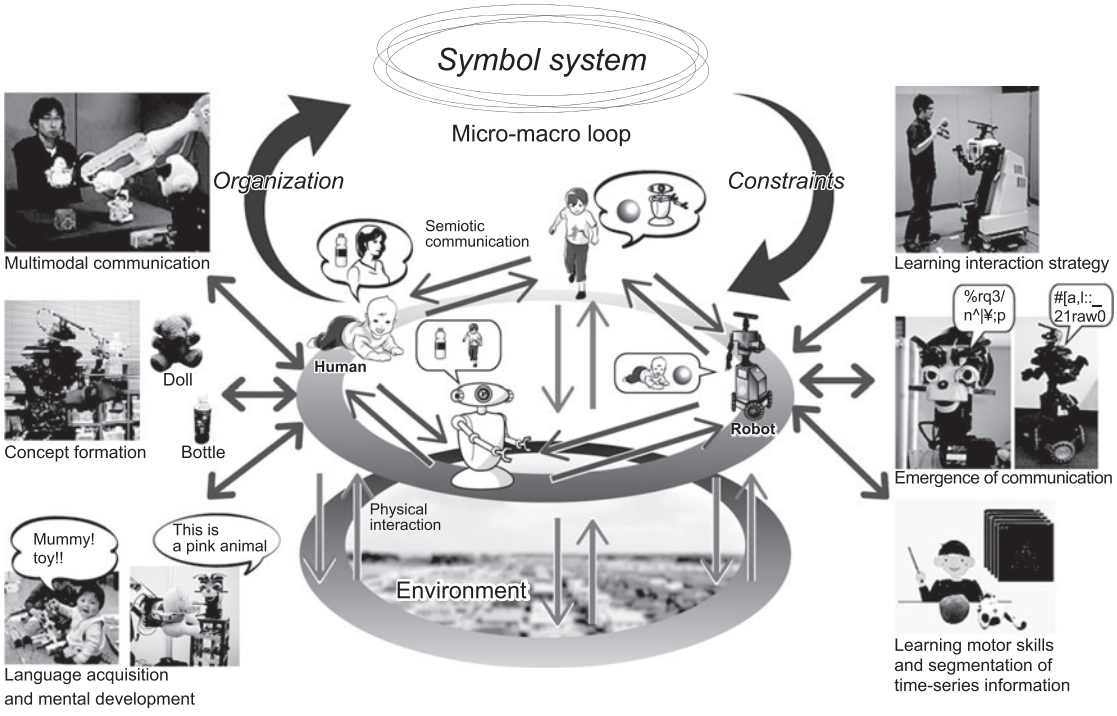


Figure 1. Symbol emergence system and research topics in SER.

keyword, not only in semiotics, artificial intelligence, and cognitive science, but also in robotics. However, the adaptability and emergent properties of symbol systems have been underestimated and even ignored in the long history of research on robotic intelligence.

In contrast, studies that consider the adaptability of robots' internal representations and autonomous unsupervised learning processes for a language system have recently attracted attention.[1–3] One aspect of such studies is regarded as a constructive approach to human-adaptive intelligence and symbol systems. However, the majority of current approaches to symbol systems in artificial intelligence and robotics are still unable to determine the dynamics and emergent properties of human symbol systems, i.e. the symbol systems retained in our human society. A philosophical theory concerning the dynamics of human symbol systems should be established, and a more sophisticated understanding regarding human symbol systems should be obtained, in order to facilitate such studies and develop intelligent robots that are capable of long-term communication and collaboration with humans.

Based on this notion, the research field called symbol emergence in robotics (SER) has gradually emerged over the past decade, especially in Japan. We held the first organized session concerning SER in a domestic conference in Japan in 2011. SER is based on the concept of a 'symbol emergence system,' which is introduced in Section 3. The symbol emergence system presumes that a human symbol system has emergent properties, and is self-organized through physical and semiotic interactions between cognitive agents, i.e. people and developmental robots.

Figure 1 presents an abstract figure illustrating the symbol emergence system and research topics in SER. We assume that robots should be developed that semantically communicate with people and collaboratively interact with people in their environment, in order that they can be assimilated into the symbol emergence system. At the center of Figure 1, the dynamics of a symbol emergence system are schematically described. The background and concepts of symbol emergence systems are detailed in Sections 2 and 3, respectively.

SER incorporates many research topics, such as multimodal communication, concept formation, language acquisition and mental development, learning interaction strategy, emergence of communication, and learning of motor skills and the segmentation of time-series information, as illustrated in Figure 1. In SER, robots are required to learn almost everything through their sensory-motor information flow, in a bottom-up manner. A top-down design of the intelligence of the robots would deprive them of the adaptability to become an element of a symbol emergence system. In particular, the formation of concepts based on a robot's multimodal sensory-motor information and autonomous language acquisition from row speech signals are both fundamental topics in SER. SER aims to build robotic and

computational models that can describe the overall dynamics and development of language acquisition and semiotic communication on the basis of a robot's and a child's self-enclosed sensory-motor experiences. This would enable the establishment of a new theory of embodied semantics.

The remainder of this paper is organized as follows. Section 2 briefly reviews the history of 'symbol systems' in artificial intelligence, cognitive science, and robotics. This forms the background of symbol emergence systems, which is a philosophical prerequisite of SER. Section 3 describes the concept of symbol emergence systems. In the subsequent sections, we provide a survey of the existing research related to SER. In particular, we review previous studies on multimodal categorization (Section 4), word discovery (Section 5), and double articulation analysis (Section 6), which are important components of language acquisition that is a fundamental challenge in SER. In Section 7, we describe further topics that are related to SER. Section 8 concludes the paper.

2. Background

2.1. Physical symbol systems and robotics

In the research field of robotics, the term 'symbol' can be used in a variety of contexts, e.g. human–robot interaction, planning, reasoning, and communication. Historically, the physical symbol system hypothesis was proposed by Newell and Simon [4,5]. This formed the starting point of a discussion concerning symbol systems in artificial intelligence and related fields. However, this starting point was problematic. The philosophy is clearly inspired by early successes in computer science and programming languages. Many related studies that built on the physical symbol system hypothesis and/or its way of thinking have placed an emphasis on the manipulation of symbols in research on artificial intelligence. This way of thinking was inherited from the tradition of 'symbolic logic.'

In predicate logic, which is a representative of symbolic logic, predicates and variables that represent real-world phenomena are given as discrete representations in a top-down manner.[6] The fundamental assumption is that our world can be distinguished and segmented into a discrete 'symbol' system, and that the system is deterministic and static. In other words, predicate logic can describe the world so far as such assumptions are satisfied. This manner of representation is regarded as a type of 'approximation.' Almost all symbolic logic essentially shares the same assumptions. The physical symbol system hypothesis, proposed by Newell, is no exception.[4,5]

This convention has implicit effects on studies in robotics. In current robotics research, a 'symbol' tends to be regarded as a 'discrete' entity, having a 'one-to-one' relationship with a word. A symbol is regarded as a manipulatable element in the mind, i.e. a robot's memory system. For example, in [7] the authors call a type of trajectory of a humanoid's entire bodily motion, modeled by a left-to-right hidden Markov model (HMM), a 'proto-symbol.' Notions such as 'a symbol is a discrete component of a memory system in a robot,' 'a symbol is an internal representation in a robot,' and 'a symbol system is a set or a network of such components,' have spread widely throughout the artificial intelligence and robotics communities.

Figuratively speaking, symbolic logic adopts the assumptions of equilibrium and determinism for modeling an actual human symbol system. It is assumed that the human symbol system is the same for everyone, and does not change over time. This approximation has been valid for solving many problems, in the same way that linear control theory has solved many problems despite most real-world systems having nonlinear and stochastic properties, or that the theory of thermodynamics of equilibrium systems has provided fruitful results for engineering purposes.

However, these assumptions are crucially problematic, and have misled many researchers over the past four decades. This has resulted in the human symbol system being misunderstood. The blind acceptance of this approximation has meant that people have not considered several important characteristics of the human symbol system.

2.2. Physical grounding hypothesis

In the field of artificial intelligence, there have been many criticisms of the physical symbol system hypothesis and other approaches to intelligent systems based on this hypothesis. Brooks put forward a representative criticism, and insisted that sensory-motor coupling with the environment is primarily important for robots to achieve everyday tasks in our daily environment.[8,9] His famous paper 'intelligence without representation' provided a clear objection to the physical symbol system hypothesis. He proposed the physical grounding hypothesis, of which the key observation is that 'the world is its own best model.' He developed many robots based on a *subsumption architecture*, which is a reactive and decentralized robotic architecture. This behavior-based approach to robotics places an emphasis on the primal sensory-motor interaction between a robot's embodied system and its environment, and the emergence of behavior through interactions. Breazeal et al. even developed a 'social' interactive robot, using the subsumption architecture.[10,11] However, the subsumption architecture is still a framework for 'designing' a robot that behaves naturally in our

daily environment. Using such an approach, it is difficult to build an autonomous system that gradually reaches an intelligent state such that it can communicate with people using a human language system.

The field of embodied cognitive science is closely related to Brooks's approach.[12] This approach is also related to the research field of artificial life and complex systems. Metaphorically, Brooks's approach is to develop an insect-like artifact. In contrast, the traditional approach to artificial intelligence can be regarded as an attempt to develop an obstinate mathematician who cannot behave appropriately in a real-world environment.

2.3. Symbol grounding problem

The other famous criticism regarding the symbol system was expressed by Harnad. He proposed the symbol grounding problem (SGP), which is one of the most famous problems in artificial intelligence.[13] The SGP focused on the relationship between a designed symbol system and real-world phenomena. The importance of the SGP has been widely recognized over the past two and a half decades. The design of a 'symbol system' and application of it to an autonomous agent in a top-down manner inevitably leads to the SGP.

Advocates of physical symbol systems have insisted that the meaning of a symbol is syntactically determined in relation with other symbols. However, such a 'relationship' cannot reach a conclusion on what anything means. A relationship between two signifiers can never provide the relationship between a signifier and a signified object. Harnad compared this phenomenon to a 'merry-go-round.' To obtain any meaning, a word has to be grounded via sensory-motor information, or borrow meanings from other words using syntactic rules. Cangelosi et al. called these processes 'sensorimotor toil' and 'symbolic theft,' respectively.[14]

In cognitive science, physical symbol systems have also been criticized. Barsalou proposed the concept of 'perceptual symbol systems,' to place an emphasis on perceptual experiences for theories of knowledge.[15] He called the static symbolic system an 'amodal symbol system,' and pointed out its drawbacks. Although the notion of a perceptual symbol system was not completely new, Barsalou mentioned that a perceptual theory of cognition may lead to a competitive, and perhaps superior, theory. The perceptual symbol system is clearly related to the SGP.

Many interdisciplinary studies have aimed to solve the SGP.[2,16–18] Recently, Tellex et al. presented an approach to the SPG using probabilistic graphical models.[19] From a philosophical viewpoint, Taddeo et al. proposed the zero semantical commitment condition, which must be satisfied by any hypothesis seeking to solve the SGP.[20]

However, despite the long history of the SGP, a clear solution has not been found. One of the reasons for this is that the SGP itself is naively defined. The SGP is based on the physical symbol system. Therefore, the SGP itself was based on the misleading physical symbol system hypothesis. That is, the SGP is an ill-posed problem. The SGP mainly considers the problem of grounding, but almost completely ignores the fact that symbol systems change dynamically and that the meanings and rules of a symbol system are determined in a social context. Steels pointed out this problem in an ambitious and important paper, titled 'The symbol grounding problem has been solved. So what's next?.'[21] He described it as follows:

> I propose to make a distinction between *c-symbols*, the symbols of computer science, and *m-symbols*, the meaning-oriented symbols in the tradition of the arts, humanities, and social and cognitive sciences.

This distinction is crucially important for the construction of theories concerning long-term human–robot interactions. The SGP starts with c-symbols, and attempts to make them grounded. To develop an intelligent robot that people would embrace long-term interactions with, we should clearly start from m-symbols, because a human–robot interaction is composed of m-symbols from a human viewpoint. In our paper, we call refer to systems of c-symbols and m-symbols as internal representation systems and human symbol systems, respectively, according to the conventions of robotics and semiotics. For example, Weng provided a critical survey on symbolic models and emergent models in artificial intelligence.[22] In our terminology, such symbol models are concerned with internal representation systems. In contrast, the symbol emergence system we introduce in the next section considers both types of symbol, in an integrative manner.

To summarize, the blind acceptance of the physical symbol system tends to encourage the following three characteristics of the human symbol system to be overlooked:

- C1 Grounded: A symbol does not have any meaning without being grounded or interpreted.
- C2 Dynamic: There does not exist an objectively true symbol system that can be determined in top-down manner in our human society.
- C3 Social: An individual representation system and the socially shared symbol system are not same.

These characteristics of symbols have been widely accepted in semiotics, and in a broader context in humanities research.[23]

2.4. Developmental robotics

The epigenetic and/or developmental viewpoint is crucially important in creating artificial intelligent systems that can adapt to a dynamic real-world environment. The field of developmental robotics has emerged gradually over the past two decades.[24] Cangelosi et al. described developmental robotics as follows:

> Developmental robotics is an approach to the autonomous design of behavioral and cognitive capabilities in artificial agents (robots) that takes direct inspiration from the developmental principles and mechanisms observed in the natural cognitive systems of children.

The field is also referred to as 'epigenetic robotics.'[16] Asada et al. used the term 'cognitive developmental robotics.'[25] Asada et al. stated that cognitive developmental robotics places more emphasis on human/humanoid cognitive development than on related approaches. Our research field, SER, philosophically inherits many concepts and fundamental assumptions from the field of (cognitive) developmental robotics. In a manner of speaking, SER is a (crucially important) branch of developmental robotics.

Developmental robotics places an emphasis on an autonomous agent's embodied interaction with the environment and the adaptive organization of the cognitive system, including cognitive capabilities relating to language and other symbol systems.[24] However, the scope of developmental robotics is tremendously large because it involves almost all of human intelligence and its diachronic changes. Moreover, developmental robotics attaches importance to interdisciplinary communication between robotics and developmental psychology. Many efforts have been made to construct a fruitful interdisciplinary academic field.

However, these characteristics of developmental robotics have distracted attention away from a computational and constructive understanding of dynamic human symbol systems and the development of robots that achieve the overall dynamics and development of language acquisition and semiotic communication. We believe that these are central topics in robotics research for achieving long-term human–robot communication and collaboration. This is our motivation for introducing the field of SER.

2.5. Symbol emergence in robotics

The approach in SER places more emphasis on the computational understanding of symbol emergence systems and cognitive capability that enables human to communicate and collaborate using symbol systems. In addition to cognitive development, SER attempts to cover semiotic phenomena. The field of SER is an interdisciplinary field, which is not only related to robotics, artificial intelligence, development psychology, and cognitive science, but also to semiotics and linguistics.

To describe the diachronic changes in internal representation systems and human symbol systems that are caused by embodied interaction and social communication, we require mathematical models such as generative models, neural networks, and related statistical models, and also robotic models such as humanoids and mobile robots, for a productive discussion and development of the integrative theory. In cognitive science, the generative probabilistic model has recently been widely applied to represent the human cognitive system.[26] In addition to such computational models, SER places an emphasis on embodied cognition. Therefore, researchers in the field of SER use robotic models to connect computational models to the real physical world. This involves the use of state-of-art machine learning technology, including Bayesian nonparametrics and deep neural networks, to model the diachronic changes in the cognitive systems and symbol systems of human/robots.

3. Symbol emergence systems

The center of Figure 1 presents a schematic figure of a symbol emergence system that was originally introduced in [27]. SER is defined as a constructive approach towards symbol emergence systems.[28] In this section, we explain symbol emergence systems by referring to the figure.

3.1. Semiosis and umwelt

We will begin by discussing a human symbol system, i.e. the m-symbols described by Steels.[21] The preexisting interdisciplinary research field that deals with human symbol systems is called semiotics. Semiotics is concerned with everything that can be interpreted as a sign, as explained by Eco [29]. Initially, semiology was introduced by Saussure, while Peirce independently introduced semiotics.[30,31] Currently, the two fields have overlapped and merged into the academic field called semiotics. From the viewpoint of semiotics, language is a representation of general symbol systems.

In Peircean semiotics, a symbol is defined as a process having three elements. The definition has a high affinity for the bottom-up approach to cognitive systems. The first is the *sign (representamen)*, which describes the form that the sign takes. The second is the *object*, which is something that the sign refers to. The third is the *interpretant*, which, rather than an interpreter, is the sense made of the sign, and something relates a sign with an

object. The important point of the Peircean definition of a symbol is that the sign, e.g. words, visual signs, or pointing, is not a symbol itself. The interpretant, the third element of a symbol, mediates between the sign and the object. This degree of freedom allows us to take a variety of interpretations and dynamics of a symbol system into consideration. In the Peircean definition, a symbol is not a static material, but a dynamic process of interpretation. Peirce calls this process 'semiosis.'[31]

The definition of a symbol is still abstract, but the definition clearly satisfies C1, C2, and C3 from Section 2. The utterances of others are always interpreted on the basis of semiosis in *semiotic communication*.

Peircean semiotics places a thorough emphasis on the subjective viewpoint. Uexküll, who established biosemiotics, proposed the famous notion of *umwelt*. The umwelt represents an animal's subjective world, which emerges on the basis of the animal's sensory-motor system.[32] Brooks also cited Uexküll, when he attacked the physical symbol system in his famous paper.[8] We should take the umwelts of robots and humans as a starting point. Their internal representation systems are initially formed through *physical interactions* with their *environment*, using their sensory-motor systems.

In this sense, *semiosis* is the key that connects Brooks' physical grounding hypothesis, which eliminated internal representation systems, to semiotic communication, which is required for long-term human–robot communication.

3.2. Arbitrariness and perspective of structurists

In contrast to Peirce, Saussure emphasized the synchronic structure of language. The defining notion of Saussurean semiotics is the *arbitrariness* of the sign (symbol). This embodies C2. The relationships between signs, such as labels and words, and categories are arbitrary, and the categories and segments of phenomena are also arbitrary. The arbitrarily determined categories, segments, and lexicons are retained in a language system to which many people, those who speak the same language, belong. Structuralists, the successors of Saussure, place an emphasis on arbitrariness. When we belong to a language system, i.e. a human symbol system, our cognition, interpretation, utterances, and even behaviors are affected by the symbol system. For example, we comprehend objects so as to classify them into preexisting categories that our language system retains. The symbol system provides *constraints* on our semiotic communication and physical interaction.

According to structuralists, 'things' do not exist independently of the symbol system that we use; reality is the creation of the media that seems to simply represent it. The structuralist perspective tends to reverse the precedence of language and cognition. They stress that our language, which incorporates arbitrariness, determines the order of the world.[33] This creates top-down constraints in a symbol emergence system.

3.3. Symbol emergence systems

This structurist exaggeration of unilateral determination is not accurate. Human symbol systems can be *organized* in a bottom-up manner. Genetic epistemology was proposed by Piaget [34], who is often called the father of cognitive development research. In genetic epistemology, the subjective world of humans is considered to be gradually 'constructed' through interactions with their environment. Piaget introduced a schema system, which is a self-organized cognitive system that emerges through sensory-motor interaction, and is believed to be the basis of the language system.[35]

An internal representation system is not a static system, but rather a dynamic system that is self-organized through physical interaction based on the sensory-motor system. Furthermore, a human symbol system is organized through semiotic communication on the basis of individual internal representation systems in a bottom-up manner (see *organization* in Figure 1). In semiotics, symbol systems are seldom treated as the static, closed, and stable systems that are inherited from preceding generations, but instead are regarded as constantly changing.[23] The bottom-up organization and the top-down constraints of the symbol system introduce an emergent property to the overall system. Therefore, we call a symbol system that is organized in a bottom-up manner an *emergent symbol system*.

Once a symbol system is generated in a society, people who use the symbol system must obey the rules of this system to communicate and collaborate with others. The symbol system includes phonetics, lexicons, syntax, and pragmatics as its constituents. If an agent belonging to the society does not follow the rules, i.e. the symbol system, then the agent cannot communicate its idea to others or collaborate with others. This means that the agent cannot make use of the powerful symbolic system for their further survival.

Such a bilateral relationship between an *emergent symbol system* at the macro level and a physical system consisting of communicating and collaborating agents at the micro level forms a *micro–macro loop*. Micro–macro loops are found in many complex systems, especially in living systems. This tells us that the entire system is an *emergent system*, i.e. a complex system having emergent properties. Polanyi, who introduced the notion of emergence, described it as follows [36]:

> If each higher level is to control the boundary conditions left open by the operations of the next lower level, this implies that these boundary conditions are in fact left open by the operations going on at the lower level.
>
> This stratification offered a framework for defining *emergence* as the action that produces the next higher level, first from the inanimate to the living, and then from each biotic level to the one above it.

The discussion above provides us with a novel concept, called a *symbol emergence system* (Figure 1). This symbol emergence system is also closely related to an autopoietic system.[37]

To develop a robot that naturally communicates and collaborate with humans in a long-term manner, we have to create a robotic system that behaves as an adaptive element of the symbol emergence system. To become an element of the symbol emergence system, a robot has to have the capability to form an internal representation system, acquire a language system that humans use, modulate its pragmatics, communicate with people in various contexts, and autonomously collaborate with people in a physical environment. These are the same tasks that human children are required to complete during their developmental period. To achieve a long-term development and adaptation to a symbol emergence system, a robot should learn the symbol system in an unsupervised manner and perform communication and collaboration autonomously.

In SER, which is a constructive approach to symbol emergence systems, our initial aim should be to develop an autonomous robot that can automatically acquire language like a human child, and gradually learn to communicate and collaborate with humans in our daily environment, in a bottom-up manner.

4. Multimodal categorization

From this section to Section 7, we present a brief survey of practical studies related to SER. Figure 2 illustrates the relationships between the topics described in following sections. These are the basic components required to develop a robot that can learn language autonomously and become an element of a symbol emergence system.

Embodied automatic language acquisition is one of the central issues in SER. The development of an embodied cognitive system that can automatically acquire language is different from the development of conventional natural language processing and automatic speech recognition systems. In language acquisition, a robot must learn language in a basically unsupervised manner from its embodied sensory-motor information. Owing to the remarkable progress in machine learning and robotic technologies, the range of realized automatic language acquisition methods is currently growing.

4.1. Object category formation

Before obtaining language, human children are considered to obtain object categories gradually through daily interactions with objects. Piaget insisted that the schema system self-organizes through the sensory-motor period, and that the system becomes the prerequisite for language.[35] An embodied multimodal sensory-motor experience must be a primal root for human category formation.

A category formation problem is different from a pattern recognition problem. In a pattern recognition problem, truth labels for recognition results are provided in a supervised manner. A vast number of studies have been carried out regarding the development of an accurate pattern recognizer. Recently, deep learning methods have yielded excellent results.[38,39] In contrast, category formation in a robot's umwelt must be autonomously performed in an unsupervised manner. From the viewpoint of machine learning, object categorization is regarded as a clustering problem, which is a type of unsupervised machine learning task.[40]

Historically, many studies have emphasized visual information in category formation by computational systems. However, the formation of object categories based solely on visual information is insufficient because our categories are organized on the basis of our multimodal sensory-motor experiences. The integration of multimodal information through category formation is important for a robot to predict future sensor information. By forming an object category on the basis of visual, auditory, and haptic information, a robot can infer auditory and haptic information from the recognized category, e.g. a bottle and a cymbal, from its visual information.

4.2. Computational models for multimodal object categorization

Recently, various computational models for multimodal object categorization have been proposed.[3,17,18,41–51] For example, Sinapov et al. proposed a graph-based multimodal categorization method, which allows a robot to recognize new objects on the basis of similarities to a set of familiar objects.[42] They also made a robot perform 10 different behaviors; obtain visual, auditory, and haptic information; and explore 100 different objects, classifying them into 20 object categories.[18] However, their multimodal categorization is performed in a supervised manner. Celikkanat et al. modeled the context

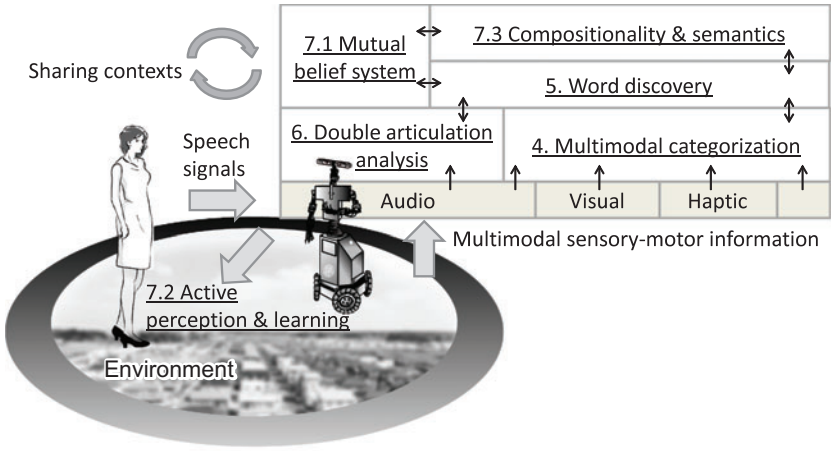


Figure 2. Mutual relationships of components corresponding to research topics introduced between Sections 4 and 7. Each arrow indicates the dependency of each component and information flow among the components.

in terms of a set of concepts, allowing many-to-many relationships between objects and contexts using latent Dirichlet allocation (LDA), inspired by the notion of situated concepts that was introduced by Yeh and Barsalou [41,52]. Mangin used a nonnegative matrix factorization algorithm to learn a dictionary of components from multimodal time series data.[53] They also showed that a type of concept characterized by cross-modal association can emerge using their proposed method, multimodal concept acquisition with non-negative matrix factorization.[54] Natale et al. have demonstrated that a robot can recognize objects with the help of a self-organizing map (SOM), using proprioceptive data extracted from the robot's hand as it grasps an object.[43] Lallee et al. proposed multi-modal convergence maps on the basis of SOMs. This method can integrate visual, motor, and language modalities.[55] Invalid et al. proposed a cognitive architecture, and developed a child-like robot that can automatically learn object categories through active exploration.[56]

A series of studies on multimodal categorization using multimodal latent Dirichlet allocation (MLDA) and its extensions led to some theoretically sophisticated statistical models.[3,17,44–48,57] Nakamura et al. have extended LDA, which was first proposed by Blei et al. for document-word clustering, to a model that can treat multimodal information.[3,58,59] Concrete illustrations of the graphical models of LDA and MLDA and its extended models are presented in Figure 3. MLDA incorporates several emission distributions for an object, i.e. a document in LDA. The object categories of objects in multimodal categorization in MLDA correspond to topics in document-word clustering in LDA. The authors developed a robotic system that can obtain visual, audio, and haptic information by interacting with objects to show the effectiveness of the multimodal categorization methods. An overview of the robot is presented in Figure 4. The robot can grasp an object and observe it from various viewpoints. The robot has cameras, microphones, arms, and hands with pressure sensors. The robot obtains visual information by taking pictures of a target object from many directions, by rotating the object with its hand. The robot also obtains haptic information by grasping the target object several times, and audio information by shaking the target object. Feature vectors are extracted from the observed information of each modality, and the feature vectors of each modality are transformed into bag-of-features representations using the K-means method, i.e. vector quantization. The bag-of-features representations are passed to MLDA, and the clustering procedure is performed. They demonstrated that a robot can categorize a large number of objects in a home environment into categories that are similar to human categorization results.[3] Araki et al. developed online MLDA, and performed an experiment on multimodal category acquisition in a fully autonomous manner in a home environment.[60] The result indicates that teacher signals provided by human participants are not necessarily required for a cognitive system to form human-like object categories. This suggests that the human symbol system is not completely arbitrary, but has a certain rationality brought by the latent structure embedded in sensory-motor information, as admitted by Saussure [30].

4.3. Estimating latent structure in multimodal categories

Although MLDA is able to form multimodal categories, it cannot adaptively estimate the number of object categories. It is unlikely that the number of categories contained in a cognitive system is determined in advance. The Bayesian nonparametric approach provides a reasonable solution to the problem. Nakamura et al.

extended the hierarchical Dirichlet process (HDP), which was a nonparametric Bayesian clustering method proposed by Blei et al. for document-word clustering, to MHDP, which can treat multimodal information and automatically estimate the number of categories from the observed multimodal information.[47,61,62]

Bayesian nonparametrics is a branch of the Bayesian approach. In general, the number of hidden variables in the Bayesian nonparametric approach can be automatically estimated using the infinite dimensional prior distribution, for example, using Dirichlet or Beta processes, and a feasible inference procedure.[61,63] The mathematical framework is very important for constructing a computational model relating to symbol emergence systems. Nakamura et al. demonstrated that a robot can also estimate the number of object categories, giving similar results to human categorization results.[47]

MLDA and MHDP can easily be extended to treat 'words.' LDA and HDP were originally applied to document and word clustering methods.[58,61] By adding an observation variable to MLDA's graphical model to represent words, MLDA is able to cluster multimodal information and words simultaneously. As a result, a robot can estimate the label of a category in an unsupervised manner.

More complex latent structures of multimodal categories can be estimated. Ando et al. proposed hierarchical MLDA, by extending hierarchical LDA for hierarchical multimodal categorization.[44,65] This method enabled a robot to form a hierarchical structure of object concepts from multimodal sensory-motor information, e.g. 'plastic bottle' is a subcategory of 'water container.'

Nakamura et al. proposed the bag of MLDA, bag of MHDP (BoMHDP), and infinite mixture of MHDPs methods, which can perform various types of categorization with different perspectives.[46,57,64] These methods emphasize some modalities, and organize categories based on information regarding the modalities that are focused on. By adopting these methods, their robot was able to form concepts about not only 'objects,' but also 'attributes,' e.g. soft, hard, green, red, or yellow, from multimodal information.

Compared with related methods for multimodal categorization, the MHDP-based approach is sophisticated from the viewpoint of Bayesian modeling. Its mathematical soundness and theoretical consistency help us to build new methods based on it, e.g. active perception.[66]

4.4. Multimodal representation learning using neural networks

Another method of integrating multimodal information involves approaches that utilize neural networks. Ngiam et al. applied deep networks for learning features over multiple modalities. By integrating visual, i.e. lip motions, and auditory information, they developed a robust speech classification system. They also demonstrated that the system exhibits the McGurk effect, which is an audio-visual perception phenomenon, in the similar manner to humans.[67] Noda et al. integrated auditory, visual, and motor information using a deep neural network.[68] In their experiment, a robot was able to recall upper bodily motion from visual and audio information, and retrieve image information from sound and joint angle inputs. Heinrich et al. extended the multiple timescale recurrent neural network (MTRNN), and obtained multimodal MTRNN, integrating visual, auditory, and motor information. Recently, neural networks with a deep architecture have been the subject of attention. Le et al. demonstrated that large-scale unsupervised learning using a deep neural network could be used to construct high-level features automatically from image data.[69] Bridging representation learning using neural networks and multimodal categorization using generative models represent important elements in this field.[70]

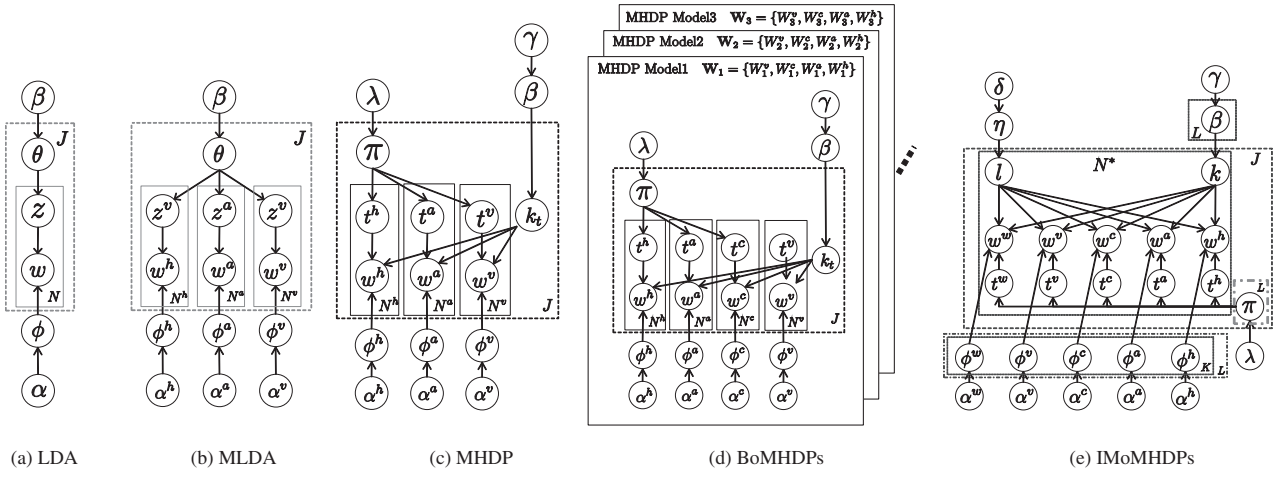
5. Word discovery

In order to obtain new vocabularies, a robot must discover words from continuous speech signals automatically. This section describes computational studies relating to word discovery.

5.1. Word discovery by human children

In language acquisition, word discovery, i.e. word segmentation, is an important task for children. A word is an elemental pattern of a linguistic sign. A phoneme is an element that is acoustically but not semantically distinguishable. Discovering words from continuous speech signals is a fundamental task that children must solve in order to acquire language. Unlike an automatic speech recognition system, children must learn a language model, i.e. a word inventory and transitional information about the words. In an acoustic model, i.e. an organized memory regarding phonemes, this must be done from speech signals in an unsupervised manner.

What types of cue can be used by children to discover words from continuous speech signals? Three representative cues for word segmentation are listed by Saffran et al. [71]. These are *distributional*, *co-occurrence*, and *prosodic* cues. Distributional cues concern the statistical relationships between neighboring speech sounds. These can be modeled as n-gram statistics to some extent, once each phoneme is recognized correctly. Co-occurrence cues concern entities detected in the environment by children.



(a) LDA (b) MLDA (c) MHDP (d) BoMHDPs (e) IMoMHDPs

Figure 3. Graphical models for object categorization. From left to right, the graphical models show categorization for (a) LDA [58], (b) MLDA [3], (c) MHDP [47], (d) BoMHDPs [64], and (e) IMoMHDPs [57].

For example, if a child hears two sentences while he/she is looking at an apple, then these sentences are likely to contain overlapping words, such as 'apple.' Prosodic cues relate more to superficial acoustic information, such as stressed syllables, post-utterance pauses, and acoustically distinctive final syllables.

All of the above cues are believed to contribute to word discovery in an integrative manner. Among these, Saffran emphasized distributional cues. She reported that word segmentation from fluent speech could be accomplished by eight-month-old infants using only distributional cues.[72] By the age of seven months, infants are reported to use distributional cues.[73]

5.2. Word discovery by robotic systems

In SER, the autonomous discovery of words by robots is one of the first challenges that should be solved. Over the past two decades, many types of unsupervised machine learning methods for word discovery (segmentation) have been proposed.[74–82] Conventionally, Brent proposed the use of model-based dynamic programming for finding word boundaries in a natural-language text whose word boundaries are deleted.[74] Venkataraman proposed a statistical model to improve Brent's algorithm.[75]

In contrast with such text-based approaches, Roy et al. developed a computational model and a robotic system that autonomously discovers words from a raw multimodal sensory input.[51] The experimental results of Roy et al. demonstrated the development of a cognitive robot that can acquire a lexicon from raw sensor data without human transcription or labeling. Although not perfect, the results were encouraging. Their results showed that it is possible to develop cognitive models that can process raw sensor data and acquire a lexicon, without the need for human transcription or labeling.

Contemporaneously, Iwahashi et al. independently proposed a sophisticated probabilistic method that enables a robot to acquire linguistic knowledge, including speech units, lexicons, grammar, and interpretation, through human–robot embodied communication, in an unsupervised manner.[83] This integrated speech, visual, and behavioral information within a probabilistic framework. This line of research was built upon by Iwahashi [84]. The learning process was carried out online, incrementally, actively, and in an unsupervised manner. On the basis of this work, Iwahashi et al. developed an integrated online machine learning system called LCore, which combined speech, visual, and tactile information obtained through interactions, and enabled robots to learn beliefs regarding speech units, words, the concepts of objects, motions, grammar, and pragmatic and communicative capabilities.[50] These pioneering studies clearly demonstrated the possibility of the SER approach.

5.3. Nonparametric Bayesian word segmentation

In word discovery and segmentation tasks, the efficient management of a word inventory through the learning process is a fundamental computational problem. Although a robot can only memorize a finite number of words, there are potentially an infinite number of words in our society, i.e. in the human symbol system. The selection of an appropriate set of words constitutes a type of model selection problem, and usually involves a very large computational cost. Recently, Bayesian nonparametrics have provided a sophisticated theoretical solution to this problem.[26,61,63] A nonparametric Bayesian language model, e.g. a hierarchical Pitman–Yor process language model, can assign an adequate probability to an

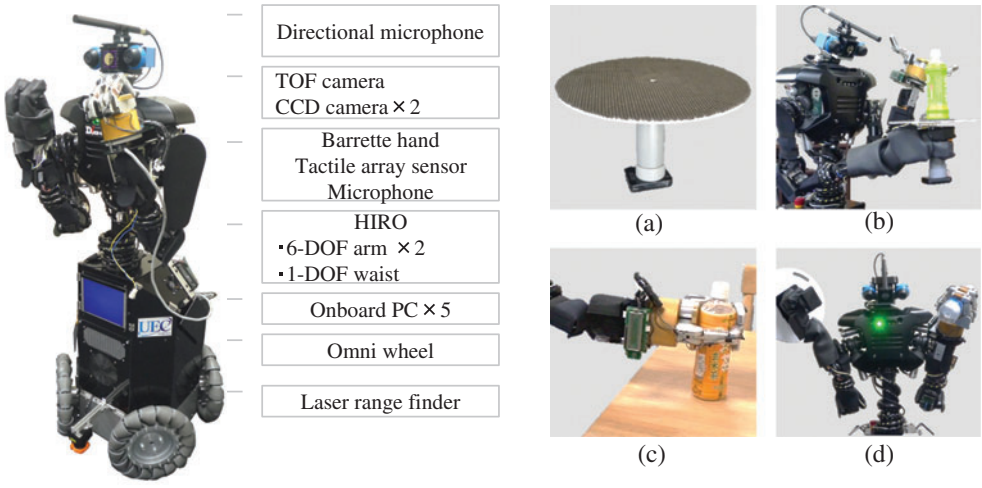


Figure 4. The robot used in multimodal categorization experiments. (Left) the robot has a variety of sensors and actuators to move around, behave, and obtain sensory information automatically. (Right) (a) the robot placed a target object on a turntable, and (b) the robot looks at the object while turning the small table from many defections. (c) The robot grasps the target object several times to obtain haptic information. (d) the robot shakes the target object to obtain auditory information. The robot itself conducts all of the behaviors to obtain multimodal sensory information.

infinite number of possible words using a fully Bayesian framework.[85] On the basis of this framework, word segmentation methods can be developed that assume that there exist an infinite number of possible words. Goldwater proposed an HDP-based word segmentation method.[76,77] Mochihashi et al. proposed a nested Pitman–Yor language model (NPYLM), in which a letter n-gram model based on a hierarchical Pitman–Yor language model is embedded into the word n-gram model.[78] An efficient blocked Gibbs sampler, which employs the forward filtering backward sampling procedure, was also introduced in that study. These methods have made it possible to discover words from transcribed phoneme sequences or text data without any recognition errors.

However, in practice phoneme recognition errors are inevitable, especially during the language acquisition phase, because phoneme recognition error rates are usually quite high without a word dictionary, i.e. an appropriate language model. In order to overcome this problem, several extensions have been proposed. Neubig et al. extended the word segmentation methods of Mochihashi et al. to analyze phoneme lattices, which constitute a way of expressing noisy speech recognition results.[86] Heymann et al. modified the algorithm of Neubig et al., and proposed a suboptimal algorithm.[87,88] Elsner et al. developed a learning method that jointly performs word segmentation and learns an explicit model of phonetic variation.[89] Hinoshita et al. solved a similar problem using MTRNN.[90]

Recently, several advanced machine learning methods have attempted to learn words from acoustic data without a preexisting language model and acoustic model.[91–94] However, this problem remains a challenging one.

5.4. Integrative word discovery by robots

It is difficult to discover words from uttered sentences using only distributed cues. Researchers have developed robotic systems and computational models in which co-occurrence cues aid the discovery of words using distributional cues. Taguchi et al. proposed a method for the unsupervised learning of place-names, using information pairs that consist of spoken utterances and a mobile robot's estimated current location, without any prior linguistic knowledge other than a phoneme acoustic model.[95] They optimized a word list using a model selection method based on a description length criterion. Araki et al. proposed an integrative computational model that involved MLDA and NPYLM.[17] Through MLDA, a robot can detect co-occurrence cues in the environment and use that information to increase word segmentation performance. It was demonstrated that the iterative learning process comprising MLDA and NPYLM increased the word segmentation accuracy. However, they reported that the accuracy decreases as the phoneme recognition error rate increases.[17] This implies that phoneme recognition errors and word segmentation errors should be mitigated simultaneously. Nakamura et al. developed an integrated statistical model for word segmentation, speech recognition, and multimodal categorization, in order to overcome this problem. The robot in their experiment simultaneously formed object categories and learned related words from continuous speech

signals and continuous visual, auditory, and haptic information, i.e. sensory-motor information, through an iterative learning process.[48]

Various word discovery methods exist that enable robots to obtain words and relationships between words and objects. However, situations in which robots can discover words remain limited. To simulate the robust word discovery process by human children in real-world environment, further studies will be required.

6. Double articulation analysis

Time series data, e.g. speech signals, human bodily motion data, and driving behaviors, are continuous. However, we can segment continuous speech signals to acquire phonemes and words, and can separate continuous human physical movements to imitate others. Segmentation of time-series data constitutes an important topic in SER. In this section, we introduce the research topic regarding double articulation analysis. That is, an unsupervised learning task for finding meaningful segments from continuous time-series data that latently have a double articulation structure.

6.1. Double articulation structure

Double articulation is an important property of human language systems. Chandler described double articulation as follows, in a textbook on semiotics [23]:

> One of the most powerful 'design features' of language is called double articulation (or 'duality of patterning'). Double articulation enables a semiotic code to form an infinite number of meaningful combinations using a small number of low-level units which by themselves are meaningless (e.g. phonemes in speech or graphemes in writing). The infinite use of finite elements is a feature that about media, in general, has been referred to as 'semiotic economy.'

Our speech signals and some semiotic time-series data are considered to have a double articulation structure. This means that a sentence can be decomposed into words, and a word can be decomposed into letters or phonemes. Automatic speech recognition systems usually presume the property of double articulation in speech signals. A continuous speech signal is first segmented into phonemes, such as 'a,' 'e,' 'i,' and 's.' Then, the phonemes are chunked into words, such as 'can,' 'dog,' and 'pen.' A phoneme cannot usually act as a sign for an object in the sense of Peircean semiosis, but a word constitutes a sign, i.e. it has certain meaning. Usually, the relationship between speech signals and phonemes is stored in a phonetic and/or acoustic model, and the relationship between phonemes and words is stored in a language model. Therefore, direct word discovery from speech signals can be regarded as the analyzing of a latent double articulation structure embedded in speech signals in an unsupervised manner.

In addition to speech signals, other time-series data generated by humans may have a double articulation structure. If such time-series data exists, then the analysis of such data would contribute to our understanding of symbol emergence systems. For this purpose, several researchers have developed a computational model that can automatically analyze double articulation structures. [93,94,96–99]

6.2. Segmentation of human bodily motion

Human bodily motion is a candidate for doubly articulated time-series data. Inamura et al. proposed the use of a left-to-right HMM to recognize and reconstruct human bodily motion.[7] Essentially, left-to-right HMMs are often used to model words in automatic speech recognition systems. These computational models implicitly bridge speech recognition and human motion modeling.

Many previous studies exist concerning motion segmentation. However, the definition of a unit of motion in many studies has been unclear for a long time. Roughly speaking, some researchers have focused on physically elemental segments,[100–104] and others on semantically elemental segments.[105–107] For example, when we semantically segment a baseball player's motion, 'pitching' definitely becomes a candidate for a unit of motion. However, pitching consists of several low-level segmental motions from the viewpoint of physical dynamics. If we segment the pitching motion according to the criterion that an elemental motion has linear dynamics, then the pitching motion will be segmented into several elemental motions, e.g. 'raising a knee' and 'swinging an arm.' However, these elemental motions seem to be meaningless, and have no special names. The two-layer hierarchical structure is similar to that existing in speech signals. We call a short-term motion that corresponds to phoneme a *segment*, and a long-term motion that corresponds to word a *chunk*. A chunk corresponds to a sequence of segments.

Takano et al. developed a large-scale database of human whole-body motion, and modeled the motion data using a large number of HMMs.[97,108] They roughly clustered the given motions, and constructed many left-to-right HMMs, corresponding to meaningful motions. They hierarchically clustered the HMMs again, and obtained a large motion database. An online incremental learning method was also provided by Kulić et al. [109]. They implicitly assumed double articulation in human bodily motion. Based on this database, they developed a machine translation system that can

translate continuous human motion into a sentence, and vice versa.[110,111]

Taniguchi et al. proposed a double articulation analyzer (DAA) by combining a sticky hierarchical Dirichlet process-HMM (sticky HDP-HMM) and NPYLM.[96,112] The DAA explicitly assumes double articulation, and infers the latent letters, i.e. the segment or phoneme, and the latent words, i.e. the words or segments, in an unsupervised manner. These two nonparametric Bayesian models, sticky HDP-HMM and NPYLM, were sequentially applied to the target data, and a two-layered hierarchical structure was inferred in the DAA. They then applied the DAA to human motion data to extract unit motions from unsegmented human motion data.

6.3. Modeling driving behavior data

Data on driving behavior represents another candidate for doubly articulated time-series data. A meaningful chunk of driver behavior seems to consist of sequences of simple segments. Figure 5 presents an example. In this figure, when a driver 'turns right at an intersection,' he/she 'steps on the brake,' 'turns the steering wheel to the right,' 'steps on the accelerator,' 'turns the steering wheel back to center,' and 'steps on the accelerator' again, as a sequence of physically elemental driving behaviors.

The DAA has been utilized for various applications, such as segmentation,[98] prediction,[113,114] data mining,[115] topic modeling,[116,117] and video summarization.[118] Through experiments performed in a series of studies, it has become clear that driving behavior data has a double articulation structure. For example, a prediction method based on the DAA has outperformed conventional methods in a driving behavior prediction task.[119] This implies that the assumption of double articulation is appropriate. Taniguchi et al. called the latent states corresponding to semantically elemental driving behaviors *driving words*, and those corresponding to physically elemental driving behaviors *driving letters*. Recently, the DAA was applied to large-scale driving behavior data, and its effectiveness for driving behavior analysis has been verified.[120]

6.4. Predicting long-term sensory-motor information

A third candidate for doubly articulated time-series data is provided by sensory-motor information flow. Unlike speech signals, sensory-motor information flow data comprises sensor information as well as motor information. Modeling sensory-motor time-series data for a robot means estimating the forward dynamics of the system that the robot confronts. Many modular learning architectures, e.g. modular selection and identification for control (MOSAIC), hierarchical, attentive, multiple models for execution and recognition (HAMMER), and the dual schemata model, have been proposed for modeling switching forward dynamics.[121–125] Most of these involve forward-inverse models in the learning system. Methods for imitation learning, reinforcement learning (RL), and the emergence of communication have been proposed on the basis of such modular learning architectures.[126–129]

In dynamic environments, the forward dynamics of a robot change intermittently. Such contextual information tends to have a certain structure. Tani et al. proposed the use of a hierarchical mixture of RNNs [99] In their experiment, a longer context was coded into the activations of the context nodes of the RNNs at a higher level, and a shorter context was coded into those at a lower level. In other words, their model captured the double articulation structure in the environmental dynamics. Hierarchical MOSAIC and HAMMER architectures have also been proposed as computational models that could capture such hierarchical structures.[122,130] MTRNN and recurrent neural networks with parametric bias (RNNPB) are candidates for modeling contextually changing forward dynamics.[2,131–133]

6.5. Direct word discovery from speech signals

As mentioned above, the double articulation analysis is deeply related to direct word discovery from acoustic speech signals. Several studies have recently been carried out in relation to this difficult problem. A direct application of the DAA, as proposed by Taniguchi and Nagasaka [96], is one possible approach. However, poor results are expected to be obtained. In this approach, the DAA simply applies the two nonparametric Bayesian methods sequentially. These are not integrated into a single generative model. Therefore, if there are many recognition or categorization errors in the result of the first segmentation process using sticky HDP-HMM, then the performance of the subsequent unsupervised chunking by NPYLM deteriorates.

To overcome this problem, Taniguchi et al. proposed a nonparametric Bayesian double articulation analyzer (NPB-DAA), which is a two-layered generative model.[94] The generative model represents a complete data generation process with a double articulation structure. An efficient blocked Gibbs sampler was also derived in the same study. They demonstrated that NPB-DAA could automatically find a word list from vowel speech signals directly and completely, in an unsupervised manner. Additional approaches have recently been proposed.[91,93]

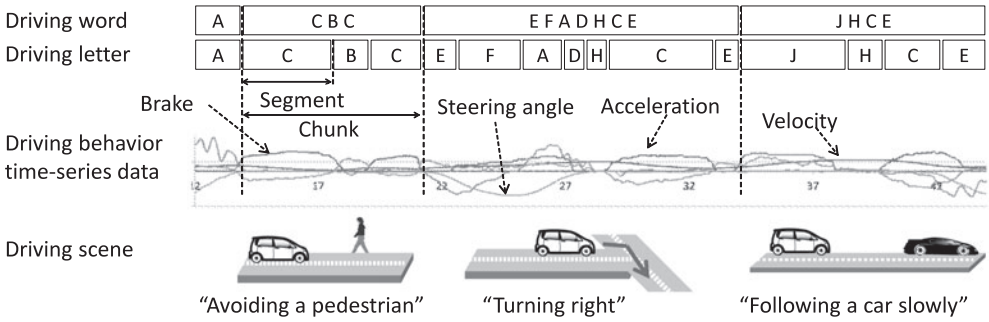


Figure 5. Double articulation in driving behavior. Observed driving behavior time-series data are segmented into chunks and segments by a DAA. A driving letter and a driving word correspond to a segment and a chunk, respectively. It is considered that each driving word corresponds to each meaningful driving scene.

It is interesting that superficially different time-series data generated from human behavior can be analyzed using almost the same computational model. In addition, the characteristic of double articulation is satisfied. That is, the elements in the first layer, i.e. the phonemes and segments, are meaningless, and the elements in the second layer, i.e. words and chunks, are meaningful. This suggests that different examples of such time-series data potentially share the same computational processes in our brain. In addition, we hypothesize that these are profoundly related to the nature of our symbol emergence system.

7. Further topics

In order to construct a computational model that describes an entire symbol emergence system, and to develop a robot that can communicate and collaborate with humans in a long-term manner, many other challenges exist in the field of SER that must be considered and overcome. In this section, we describe some of these.

7.1. Mutual belief system

An utterance is not usually interpreted 'as it is' by a person to whom it is told. A mutual belief system always affects a person's interpretation. Roy provides a coffee scenario as an example in his survey paper.[134] Imagine a situation in which a cup of coffee is served to a customer by a waitperson, and the customer says, 'This coffee is cold.' In this case, the referential meaning of the utterance is the fact that the temperature of the coffee is low in the sense of thermodynamics, but the functional meaning is 'please get me a hotter coffee.' The speech act conveys a meaning interpreted by referring to the physical situation shared by the communication partners. This means that the mutual belief system is important in generating natural sentences and interpreting uttered sentences. Roy emphasized this aspect of language for solving the SGP.

A pioneering study that develops a constructive model involving a mutual belief system was presented by Iwahashi [83]. Iwahashi introduced a belief function that represents a mutual belief of a robot. The belief function contains several belief modules, i.e. speech, object images, motions, motion–object relationships, and behavioral contexts. The various external and internal contexts are taken into consideration to infer the speaker's intention. The robotic system is truthfully an embodied natural language processing system that can take various contexts that an ordinal amodal natural processing system cannot access. Zuo et al. applied this model for detecting robot-directed speech.[135] Sugiura et al. proposed a method for estimating the ambiguity in commands by introducing an active learning scheme to the conversation system, based on mutual belief.[136]

A mutual belief is a part of an emergent symbol system, and applies 'constraints' not only to the interpretation of utterances, but also to the generation of our speech. Context is a crucial element of our natural dialog, but is rarely taken considered in natural language processing. The SER approach is promising in this topic as well.

7.2. Active perception and learning

For a robot to have the potential to be an element of a symbol emergence system, i.e. a member of our semiotic society, it must be able to explore its environment, acquire knowledge, and communicate with people autonomously. Active perception and active learning are two of the most important capabilities of humans for achieving life-long development and communication. [43,134,137–148]

Denzler et al. proposed an information theoretic action selection method to gather information that conveys the true state of a system through an active camera.[137] They used mutual information as a criterion for action selection. Krainin et al. developed an active perception

method that allowed a mobile robot manipulate objects to build three-dimensional surface models of the objects.[138] Their method determines when and how a robot should grasp an object on the basis of the information gain (IG) criterion.

Modeling and recognizing a target object, as well as modeling a scene and segmenting objects from that scene, are important abilities for a robot in a realistic environment. Eidenberger and Scharinger [139] proposed an active perception planning method for scene modeling in a realistic environment. A partially observable Markov decision process formulation was used to model the planning problem, and the differential entropy was introduced as part of the reward function. Hoof et al. proposed an active scene exploration method. Using this method, an autonomous robot is able to segment a scene into its constituent objects by actively interacting with the objects.[140] The authors used IG as a criterion for action selection. InfoMax control for acoustic exploration was proposed by Rebguns et al. [141]. In general, IG is most often applied in active perception and learning.[137,138] Taniguchi et al. developed an optimal active perception scheme for multimodal categorization, using MLDA on the basis of the IG criterion. Localization, mapping, and navigation are also important targets of active perception.[134,142,143] In addition, various other studies on active perception have been conducted.[43, 144–148]

Intrinsic motivation has been studied in the context of RL.[149] In RL, an agent autonomously learns its policy, i.e. controller, to maximize the expected cumulative rewards. In related studies, internal reward systems have been considered, as well as external reward systems. Schmidhuber presented a survey on RL studies considering intrinsic motivation, and proposed a formal theory of fun, intrinsic motivation, and creativity.[150]

When we develop an autonomous robot, the designs of the intrinsic motivation and explanatory behavior, active perception, and in particular the IG criterion provide cues for the problem. The effectiveness of the IG criterion tells us that 'curiosity' is properly treated computationally, in contrast with other emotions.

Sugiura et al. used active perception to determine the utterances of a robot.[136] When we talk to each other, we anticipate a reply, i.e. some information, from the other person. Therefore, we might generate a sentence so as to maximize the IG or expected reward. Communication always involves some kind of decision-making problems. Active perception and learning will become ever more important in the wide range of decision-making problems concerning robots.

7.3. Compositionality and semantics

The hierarchical structure placed on segmented words that are extracted on the basis of double articulation analysis enables us to generate various meaningful sentences. A syntax is a rule that produces a meaningful sequence of words. From the viewpoint of a symbol emergence system, an important problem can be phrased as 'how can combined words have adequate meanings for an embodied and situated agent?' An important research topic concerns computational models for grounding semantic composition.[151] Classically, the principle of compositionality, which is also called Frege's principle, has been widely recognized. This is a principle stating that the meaning of a complex expression is determined by the meanings of its constituent expressions and the syntactic rules used for combining them.[152] A bottom-up approach to the principle of compositionality also represents an important topic in SER.

The notion of combinatoriality has been studied in a constructive manner. Sugita et al. and Ogata et al. developed a robotic system and neural network architecture that can simultaneously learn sentences and behaviors.[131,153,154] For this, they used an RNNPB. It was shown that compositionality emerges on the network in a distributed manner. Tuci et al. also introduced neurodynamic models that deal with compositionality problems in language and behavior association learning and in the learning of goal-directed actions.[155,156] Hinaut et al. applied reservoir computing, which is a kind of RNN, to allow a robot to acquire and produce grammatical constructions.[157,158] Tani and Cangelosi et al. have presented a comprehensive survey of related studies.[1,2]

Recently, distributional models of semantics, including word2vec, have attracted attention.[159–161] For example, Mikolov showed that the relationship between a country and a capital city can be automatically extracted from an unlabeled text data-set using only a training predictor, on the basis of a skip-gram and recurrent neural net language model. Le et al. proposed an unsupervised machine learning method, called paragraph vector, that can estimate fixed-length feature representations from variable-length sections of texts, e.g. sentences, paragraphs, and documents.[162]

Compositionality of language is deeply related to the planning of an action sequence. An important open question is how a robot becomes able to manipulate internal representations formed in a bottom-up manner, and form symbolic planning operators as well. As one of the preriminarly approaches towards the problem, Uger et al. reported that they developed a system that formed symbols and symbolic planning operators in the

continuous sensorimotor experience of the robot through self-exploration.[163]

Many preliminary studies exist concerning compositionality and semantics. Incorporating both embodied cognition and formal language structure must be important in constructing a robot that can understand uttered sentences in the real word. The connection of such learning methods to a series of studies in SER also represents an important topic.

8. Conclusion

In this paper, we have provided an introduction to symbol emergence systems and surveyed the research field of SER. Semioticians sometimes call humans 'Homosignificans,' which means meaning-makers.[23] Comprehending signs from natural or artificial environments and applying semiosis in the mind are human characteristics. In order to develop an autonomous robot that can engage in long-term communication and collaboration with people, the robot must be able to adapt to the human symbol system. To provide a philosophical framework for the diversity and dynamics of a symbol system, we introduce a concept – the symbol emergence system – that constitutes a basic assumption in SER. People can acquire language through physical interactions with their environment and semiotic communication with other people (see Figure 1). This phenomenon is comprehended as a type of assimilation process, in which a personal symbol system that is supported by an internal representation system becomes coupled with the emergent symbol system. To achieve such assimilation, the person must have the capability to learn the language in an unsupervised manner. The same requirement arises for robots. To achieve long-term interaction with people, a robot must have the capability to learn the language in an unsupervised manner, so that the robot's symbol system becomes coupled with the emergent symbol system of the target society. Therefore, it is important to gain a computational understanding of how humans can learn a symbol system and obtain semiotic skills through their autonomous mental development.

Many challenges have been confronted in relation to the construction of robotic systems and machine learning methods that can obtain some parts of language through the embodied multimodal interaction with the environment. In order to understand human–social interactions and develop a robot that can smoothly communicate with human users, it is fundamentally important to understand symbol systems that change dynamically on the basis of the embodied cognition of participants in a constructive manner.

In this paper, we have introduced the research field called SER. This represents a constructive approach towards a symbol emergence system and an emergent symbol system. The emergent symbol system is socially self-organized through both semiotic and physical interactions with autonomous cognitive developmental agents, i.e. humans and developmental robots. Among the numerous fields connected with SER, we have described some specific topics in this paper, such as multimodal categorization, word discovery, and double articulation analysis. SER presents various future challenges involving acquiring lexicons, learning syntax, obtaining skills using metaphors, learning pragmatics, and being able to generate appropriate sentences given a particular context.

The majority of previous studies relating natural language processing with linguistics have only considered documents. However, we have to communicate and collaborate with other agents, including people and robots, in a real-world environment. The appropriateness of emergent symbol systems and robotic systems must be evaluated in relation to embodied cognition, context, and collaboration. Real-world collaborative tasks should be considered for this purpose. In this context, competitions including real-world human–robot interaction would be effective in allowing researchers to study embodied cognitive systems and symbol emergence systems, and to evaluate the appropriateness of such systems when developed. RoboCup@Home is an obvious candidate.[164–167]

However, acquiring the necessary knowledge for communication and collaboration through situated and embodied interactions requires huge costs and time. For further research, some pseudo-real-world environment will be required to increase the speed of our research. For example, Inamura et al. have developed the SIGVerse, a SocioIntelliGenesis simulator that enables human–robot interactions in a virtual world.[168] Cloud-based semiotic and physical interactions will also be important components of further studies in SER.

Establishing standard evaluation methods and organizing open data-sets for promoting the research field of SER also constitute important issues. In the context of cognitive robotics, developmental robotics, and embodied cognitive science, researchers have been emphasizing embodied interaction in the real-world environment. Of course, different robots have different sensory-motor systems, and usually act in different environments. Preparing standard data-sets has been a difficult challenge. In addition, a robot that learns and behaves in an on-line manner in a real-world environment is able to gradually change its methods of interaction. This results in the obtained data itself being altered, depending on its

history of interaction and learning. A standard data-set for learning and evaluation is popular in the field of pattern recognition, which usually does not consider sensory-motor interactions, but only deals with sensory information. In this sense, data for learning tends to be considered to represent a one-off experience, and is difficult to share in the related field of robotics. This line of thinking seems to have prevented researchers from sharing data-sets in the community. However, as described in this paper, SER involves various tasks. Some of these, e.g. multimodal categorization, unsupervised language acquisition, and motion segmentation, have a property similar to pattern recognition problems, and so it will be possible to prepare open data-sets. To organize tasks related to SER, organize standard evaluation methods, and provide data-sets is important for allowing many researchers who cannot afford robotic systems to participate in the research field, and to compare different methods objectively.

From a scientific point of view, SER is regarded as a constructive approach towards symbol emergence systems. Meanwhile, SER is a research field that facilitates the use of sensory-motor data that a robot can obtain in the real world, using statistical models and machine learning methods. This means that it shares common interests regarding the application of machine learning methods to robotics with many related approaches in robotics. For example, Ohno et al. emphasize data engineering in robotics (DER).[169] They insist that making full use of sensor data can result in new robot intelligence in the future. Clearly, SER has a strong relationship with DER. By clarifying the relationships with related research areas, approaches, and activities and actively facilitating collaboration with them, SER will contribute not only to the robotics community, but also to a wider range of research communities including artificial intelligence, cognitive science, neuroscience, and developmental psychology.

SER remains an emerging research field, but one that shows promise. Further research on SER will push long-term human–robot interaction forward, and provide a new understanding of human intelligence.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was partially supported by a Grant-in-Aid for Young Scientists (B) 2012–2014 [24700233] and a Grant-in-Aid for Young Scientists (A) 2015–2019 [15H05319] funded by the Ministry of Education, Culture, Sports, Science, and Technology, Japan. This research was supported by CREST, JST.

Notes on contributors

Tadahiro Taniguchi received the ME and PhD degrees from Kyoto University, in 2003 and 2006, respectively. From April 2005 to March 2006, he was a member of Japan Society for the Promotion of Science (JSPS) and research fellow (DC2) at the Department of Mechanical Engineering and Science, Graduate School of Engineering, Kyoto University. From April 2006 to March 2007, he was a JSPS research fellow (PD) at the same department. From April 2007 to March 2008, he was a JSPS research fellow at the Department of Systems Science, Graduate School of Informatics, Kyoto University. From April 2008 to March 2010, he was an assistant professor at the Department of Human and Computer Intelligence, Ritsumeikan University. Since April 2010, he has been an associate professor at the same department. From September 2015 to September 2016, he is a visiting associate professor at Department of Electrical and Electronic Engineering, Imperial College London. He has been engaging in research on machine learning, symbol emergence systems and intelligent vehicles.

Takayuki Nagai received his BE, ME, and DE degrees from the Department of Electrical Engineering, Keio University, in 1993, 1995, and 1997, respectively. Since 1998, he has been with the University of Electro-Communications where he is currently a professor of the Graduate School of Informatics and Engineering. From 2002 to 2003, he was a visiting scholar at the Department of Electrical Computer Engineering, University of California, San Diego. Since 2011, he has also been a visiting researcher at Tamagawa University Brain Science Institute. He has received the 2013 Advanced Robotics Best Paper Award. He is a member of the IEEE, RSJ, JSAI, IEICE, and IPSJ.

Tomoaki Nakamura received his BE, ME, and Dr. of Eng. degrees from the University of Electro-Communications in 2007, 2009, and 2011. From 2011 to 2012, He was a research fellow of the Japan Society for the Promotion of Science. In 2013, he worked for Honda Research Institute Japan Co., Ltd. He is currently an assistant professor at the University of Electro-Communications. His research interests are intelligent robotics and machine learning.

Naoto Iwahashi received the BE degree in Engineering from Keio University, Yokohama, Japan, in 1985. He received the PhD degree in Engineering from Tokyo Institute of Technology, in 2001. In April 1985, he joined Sony Corp., Tokyo, Japan. From October 1990 to September 1993, he was at Advanced Telecommunications Research Institute International (ATR), Kyoto, Japan. From October 1998 to June 2004, he was with Sony Computer Science Laboratories Inc., Tokyo, Japan. From July 2004 to March 2010, he was with ATR. From November 2005 to March 2011, he was a visiting professor at Kobe University. In April 2008, he joined the National Institute of Information and Communications Technology, Kyoto, Japan. Since April 2014, he has been a professor at Okayama Prefectural University.

Since April 2011, he has also been a visiting researcher at Tamagawa University Brain Science Institute. His research areas include machine learning, artificial intelligence, spoken language processing, human–robot interaction, developmental multimodal dialog systems, and language acquisition robots.

Tetsuya Ogata received the BS, MS, and DE degrees in Mechanical Engineering, in 1993, 1995 and, respectively, from Waseda University. From 1999 to 2001, he was a research associate in Waseda University. From 2001 to 2003, he was a Research Scientist in the Brain Science Institute, RIKEN. From 2003 to 2012, he was an associate professor in the Graduate School of Informatics, Kyoto University. Since 2012, he has been a professor of the Faculty of Science and Engineering, Waseda University. From 2009 to 2015, he was a JST (Japan Science and Technology Agency) PRESTO Researcher (five years). From 2015, he has been a Visiting Researcher of Artificial Intelligence Research Center, AIST. His research interests include human–robot interaction, dynamics of human–robot mutual adaptation and inter-sensory translation in robot systems with neuro-dynamical models.

Hideki Asoh received his B Eng in mathematical engineering and M Eng in information engineering from the University of Tokyo, in 1981 and 1983 respectively. In April 1983, he joined in Electrotechnical Laboratory as a researcher. From 1993 to 1994 he stayed at German National Research Center for Information Technology as a visiting research scientist. He is currently a deputy director of Artificial Intelligence Research Center at National Institute of Advanced Industrial Science and Technology (AIST). His research interests are in constructing intelligent systems which can learn through interactions with the real world.

References

[1] Cangelosi A, Metta G, Sagerer G, et al. Integration of action and language knowledge: a roadmap for developmental robotics. IEEE Trans. Auton. Mental Dev. 2010;2:167–195.

[2] Tani J. Self-organization and compositionality in cognitive brains: a neurorobotics study. Proc. IEEE. 2014;102:586–605.

[3] Nakamura T, Nagai T, Iwahashi N. Grounding of word meanings in multimodal concepts using LDA. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); St. Louis, MO; 2009. p. 3943–3948.

[4] Newell A, Simon HA. Completer science as empirical inquiry: symbols and search. Commun. ACM. 1976;19:113–126.

[5] Newell A. Physical symbol systems. Cognitive Sci. 1980;4:135–183.

[6] Russell S, Norvig P. Artificial intelligence: a modern approach. 3rd ed. London: Pearson; 2009.

[7] Inamura T, Toshima I, Tanie H, et al. Embodied symbol emergence based on mimesis theory. Int. J. Rob. Res. 2004;23:363–377.

[8] Brooks R. Elephants don't play chess. Rob. Auton. Syst. 1990;6:3–15.

[9] Brooks R. Intelligence without representation. Artif. Intell. 1991;47:139–159.

[10] Breazeal C. Emotion and sociable humanoid robots. Int. J. Human Comput. Stud. 2003;59:119–155.

[11] Breazeal CL. Designing sociable robots. Cambridge (MA): MIT press; 2004.

[12] Pfeifer R, Scheier C. Understanding intelligence. Cambridge (MA): Bradford Books; 2001.

[13] Harnad S. The symbol grounding problem. Phys. D: Nonlinear Phenom. 1990;42:335–346.

[14] Cangelosi A, Greco A, Harnad S. From robotic toil to symbolic theft: grounding transfer from entry-level to higher-level categories. Connection Sci. 2000;12:143–162.

[15] Barsalou LW. Perceptual symbol systems. Behav. Brain Sci. 1999;22:1–16.

[16] Cangelosi A, Riga T. An embodied model for sensorimotor grounding and grounding transfer: experiments with epigenetic robots. Cognitive Sci. 2006;30:673–689.

[17] Araki T, Nakamura T, Nagai T, et al. Online learning of concepts and words using multimodal LDA and hierarchical Pitman–Yor Language Model. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); Algarve, Portugal; 2012. p. 1623–1630.

[18] Sinapov J, Schenck C, Staley K, et al. Grounding semantic categories in behavioral interactions: Experiments with 100 objects. Rob. Auton. Syst. 2014;62:632–645.

[19] Tellex S, Kollar T, Dickerson S. Approaching the symbol grounding problem with probabilistic graphical models. AI Mag. 2011;32:64–76.

[20] Taddeo M, Floridi L. Solving the symbol grounding problem: a critical review of fifteen years of research. J. Exp. Theor. Artif. Intell. 2005;17:419–445.

[21] Steels L. The symbol grounding problem has been solved, so what's next ? Vol. 2005, Symbols, Embodiment and Meaning. Oxford Oxford University Press; 2008. p. 223–244.

[22] Weng J. Symbolic models and emergent models: a review. IEEE Trans. Auton. Mental Dev. 2012;4:29–53.

[23] Chandler D. Semiotics the basics. London: Routledge; 2002.

[24] Cangelosi A, Schlesinger M. Developmental robotics. Cambridge (MA): MIT press; 2015.

[25] Asada M, Hosoda K, Kuniyoshi Y, et al. Cognitive developmental robotics: a survey. IEEE Trans. Auton. Mental Dev. 2009;1:12–34.

[26] Tenenbaum JB, Kemp C, Griffiths TL, et al. How to grow a mind: statistics, structure, and abstraction. Science. 2011;331:1279–1285.

[27] Tagniguchi T. Can we create a robot that communicate with human? -constructive approach towards symbol emergence system. Tokyo: NTT publishing Co. Ltd.; 2010. Japanese.

[28] Tagniguchi T. Symbol emergence in robotics - introduction to mechanism of intelligence. Tokyo: Kidansha; 2014. Japanese.

[29] Eco U. A theory of semiotics. London: Indiana University Press; 1976.

[30] de Saussure F. Course in general linguistics (trans. Roy Harris). London: Columbia University Press: 1983.
[31] Peirce CS. Collected writings. Cambridge: Harvard University Press; 1931–1958.
[32] Von Uexküll J. A stroll through the worlds of animals and men: a picture book of invisible worlds. Semiotica. 1992;89:319–391.
[33] Sturrock J. Structuralism. London: Paladin; 1986.
[34] Piaget J. Genetic epistemology. New York (NY): W W Norton & Co Inc.; 1971.
[35] Flavell JH. The developmental psychology of Jean Piaget. Whitefish (MT): Literary Licensing, LLC; 2011.
[36] Polanyi M. The tacit dimension. Chicago: The University of Chicago Press; 1966.
[37] Maturana HR, Varela FJ. The tree of knowledge: the biological roots of human understanding. Rev. ed. Shambhala: Shambhala Publications; 1992.
[38] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems (NIPS); Stateline, Nevada; 2012. p. 1–9.
[39] Dahl GE, Yu D, Deng L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Process. 2012;20:30–42.
[40] Bishop C. Pattern recognition and machine learning (information science and statistics). Berlin: Springer; 2010.
[41] Celikkanat H, Orhan G, Pugeault N, et al. Learning and using context on a humanoid robot using latent dirichlet allocation. In: Joint IEEE International Conferences on Development and Learning and Epigenetic Robotics (ICDL-Epirob); Genoa, Italy; 2014. p. 201–207.
[42] Sinapov J, Stoytchev A. Object category recognition by a humanoid robot using behavior-grounded relational learning. In: IEEE International Conference on Robotics and Automation (ICRA); Shanghai; 2011. p. 184–190.
[43] Natale L, Metta G, Sandini G. Learning haptic representation of objects. International Conference of Intelligent Manipulation and Grasping (IMG); Genoa, Italy; 2004.
[44] Ando Y, Nakamura T, Araki T, et al. Formation of hierarchical object concept using hierarchical latent Dirichlet allocation. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); Tokyo; 2013. p. 2272–2279.
[45] Nakamura T, Nagai T, Iwahashi N. Multimodal object categorization by a robot. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); San Diego; 2007. p. 2415–2420.
[46] Nakamura T, Nagai T, Iwahashi N. Bag of multimodal LDA models for concept formation. In: IEEE International Conference on Robotics and Automation (ICRA); Shanghai; 2011. p. 6233–6238.
[47] Nakamura T, Nagai T, Iwahashi N. Multimodal categorization by hierarchical dirichlet process. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); San Francisco; 2011. p. 1520–1525.
[48] Nakamura T, Nagai T, Funakoshi K, et al. Mutual learning of an object concept and language model based on MLDA and NPYLM. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); Chicago; 2014. p. 600–607.
[49] Griffith S, Sinapov J, Sukhoy V, et al. A behavior-grounded approach to forming object categories: separating containers from noncontainers. IEEE Trans. Auton. Mental Dev. 2012;4:54–69.
[50] Iwahashi N, Sugiura K, Taguchi R, et al. Robots that learn to communicate: a developmental approach to personally and physically situated human-robot conversations. In: Dialog with Robots Papers from the AAAI Fall Symposium; Arlington, Virginia; 2010. p. 38–43.
[51] Roy DK, Pentland AP. Learning words from sights and sounds: a computational model. Cognitive Sci. 2002;26:113–146.
[52] Wenchi YEH, Barsalou LW. The situated nature of concepts. Am. J. Psychol. 2006;119:349–384.
[53] Mangin O, Oudeyer PY. Learning semantic components from subsymbolic multimodal perception. In: IEEE 3rd Joint International Conference on Development and Learning and Epigenetic Robotics (ICDL); New Delhi; 2013. p. 1–7.
[54] Mangin O, Filliat D, ten Bosch L, et al. MCA-NMF: multimodal concept acquisition with non-negative matrix factorization. Plos One. 2015;10:e0140732. Available from: http://dx.plos.org/10.1371/journal.pone.0140732.
[55] Lallee S, Ford Dominey P. Multi-modal convergence maps: from body schema and self-representation to mental imagery. Thousand Oaks (CA): Sage Publications; 2013.
[56] Ivaldi S, Nguyen SM, Lyubova N, et al. Object learning through active exploration. IEEE Trans. Auton. Mental Dev. 2014;6:56–72.
[57] Nakamura T, Ando Y, Nagai T, et al. Concept formation by robots using an infinite mixture of models. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); Hamburg, Germany; 2015.
[58] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J. Mach. Learn. Res. 2003;3:993–1022.
[59] Griffiths TL, Steyvers M. Finding scientific topics. Proc. Nat. Acad. Sci. USA (PNAS). 2004;101:5228–5235.
[60] Araki T, Nakamura T, Nagai T, et al. Autonomous acquisition of multimodal information for online object concept formation by robots. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); San Francisco; 2011. p. 1540–1547.
[61] Teh Y, Jordan M, Beal M, et al. Hierarchical dirichlet processes. J. Am. Stat. Assoc. 2006;101:1566–1581.
[62] Sudderth EB, Torralba A, Freeman W, et al. Describing visual scenes using transformed dirichlet processes. Vol. 18, In: Advances in Neural Information Processing Systems (NIPS); Vancouver; 2005. p. 1297–1304.
[63] Teh Y, Jordan M. Hierarchical Bayesian nonparametric models with applications. Bayesian nonparametrics. Cambridge: Cambridge University Press; 2009. p. 158.
[64] Nakamura T, Nagai T, Iwahashi N. Bag of multimodal hierarchical dirichlet processes: Model of complex conceptual structure for intelligent robots. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); Algarve, Portugal; 2012. p. 3818–3823.
[65] Blei DM, Griffiths TL, Jordan MI. The nested Chinese restaurant process and Bayesian nonparametric

inference of topic hierarchies. J. ACM (JACM). 2007;57:1–30. 0710.0845.
[66] Tagniguchi T, Takano T, Yoshino R. Multimodal hierarchical dirichlet process-based active perception. 2016. arXiv:1510.00331.
[67] Ngiam J, Khosla A, Kim M, et al. Multimodal deep learning. In: Proceedings of The 28th International Conference on Machine Learning (ICML); Bellevue; 2011. p. 689–696.
[68] Noda K, Arie H, Suga Y, et al. Intersensory causality modeling using deep neural networks. In: IEEE International Conference on Systems, Man, and Cybernetics (SMC); Manchester; 2013. p. 1995–2000.
[69] Le QV, Ranzato M, Monga R, et al. Building high-level features using large scale unsupervised learning. International Conference in Machine Learning (ICML); Bellevue; 2011.
[70] Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. 2013;35:1798–1828.
[71] Saffran JR, Newport EL, Aslin RN. Word segmentation: the role of distributional cues. J. Memory Lang. 1996;35:606–621.
[72] Saffran JR, Aslin RN, Newport EL. Statistical learning by 8-month-old infants. Science. 1996;274:1926–1928.
[73] Thiessen ED, Saffran JR. When cues collide: use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. Dev. Psychol. 2003;39:706–716.
[74] Brent MR. An efficient, probabilistically sound algorithm for segmentation and word discovery. Mach. Learn. 1999;34:71–105.
[75] Venkataraman A. A statistical model for word discovery in transcribed speech. Comput. Linguistics. 2001;27:351–372.
[76] Goldwater S, Griffiths TL, Johnson M, et al. Contextual dependencies in unsupervised word segmentation. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (ACL); Sydney; 2006. p. 673–680.
[77] Goldwater S, Griffiths TL, Johnson M. A Bayesian framework for word segmentation: exploring the effects of context. Cognition. 2009;112:21–54.
[78] Mochihashi D, Yamada T, Ueda N. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP); Singapore; 2009. p. 100–108.
[79] Johnson M, Goldwater S. Improving nonparameteric Bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In: Annual Conference of the North American Chapter of the Association for Computational Linguistics (ACL); Colorado; 2009. p. 317–325.
[80] Chen M, Chang B, Pei W. A joint model for unsupervised Chinese word segmentation. In: Conference on Empirical Methods in Natural Language Processing (EMNLP); Doha; 2014. p. 854–863.
[81] Magistry P. Unsupervized word segmentation : the case for Mandarin Chinese. Vol. 2, In: Annual Meeting of the Association for Computational Linguistics (ACL); Jeju, Korea; 2012. p. 383–387.
[82] Sakti S, Finch A, Isotani R, et al. Unsupervised determination of efficient Korean LVCSR units using a Bayesian Dirichlet process model. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); Prague, Czech Republic; 2011. p. 4664–4667.
[83] Iwahashi N. Language acquisition through a human-robot interface by combining speech, visual, and behavioral information. Inform. Sci. 2003;156:109–121.
[84] Iwahashi N. Interactive learning of spoken words and their meanings through an audio-visual interface. IEICE Trans. Inform. Syst. 2008;312–321.
[85] Teh YW. A hierarchical Bayesian language model based on Pitman-Yor processes. In: International Conference on Computational Linguistics and the Annual Meeting of the Association for Computational Linguistics (ACL); Sydney; 2006. p. 985–992.
[86] Neubig G, Mimura M, Mori S, et al. Bayesian learning of a language model from continuous speech. IEICE Trans. Inform. Syst. 2012;E95-D:614–625.
[87] Heymann J, Walter O. Unsupervised word segmentation from noisy input. In: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU); Olomouc, Czech Republic; 2013. p. 458–463.
[88] Heymann J, Walter O, Haeb-umbach R, et al. Iterative bayesian word segmentation for unsupervised vocabulary discovery from phoneme lattices. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); Florence, Italy; 2014. p. 4085–4089.
[89] Elsner M, Goldwater S, Feldman N, et al. A joint learning model of word segmentation, lexical acquisition, and phonetic variability. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington: USA; 2013. p. 42–54.
[90] Hinoshita W, Arie H, Tani J, et al. Emergence of hierarchical structure mirroring linguistic composition in a recurrent neural network. Neural Networks. 2011;24:311–320.
[91] Kamper H, Jansen A, Goldwater S. Fully unsupervised small-vocabulary speech recognition using a segmental bayesian model. Interspeech; Dresden, Germany; 2015; p. 678–682.
[92] Brandl H, Wrede B, Joublin F, et al. A self-referential childlike model to acquire phones, syllables and words from acoustic speech. In: IEEE International Conference on Development and Learning (ICDL); Futuroscope-Chasseneuil, France; 2008. p. 31–36.
[93] Cy Lee. Donnell TJO, Glass J. Unsupervised lexicon discovery from acoustic input. Trans. Assoc. Comput. Linguistics. 2015;3:389–403.
[94] Taniguchi T, Nakashima R, Nagasaka S. Nonparametic Bayesian double articulation analyzer for direct language acquisition from continuous speech signals. 2015. arXiv:1506.06646.
[95] Taguchi R, Yamada Y, Hattori K, et al. Learning place-names from spoken utterances and localization results by mobile robot. In: Interspeech; 2011. p. 1325–1328.
[96] Taniguchi T, Nagasaka S. Double articulation analyzer for unsegmented human motion using Pitman–Yor language model and infinite hidden Markov model.

In: IEEESICE International Symposium on System Integration (SII); Kyoto, Japan; 2011. p. 250–255.
[97] Takano W, Imagawa H, Kulić D, et al. What do you expect from a robot that tells your future? The crystal ball. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); Taipei, Taiwan; 2010. p. 1780–1785.
[98] Takenaka K, Bando T, Nagasaka S, et al. Contextual scene segmentation of driving behavior based on double articulation analyzer. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); Algarve, Portugal; 2012. p. 4847–4852.
[99] Tani J, Nol S. Learning to perceive the world as articulated: an approach for hierarchical learning in sensory-motor systems. Neural Networks. 1999;12:1131–1141.
[100] Rubin J. Boundaries of visual motion. Vol. 835, AI Memo. Cambridge (MA): MIT Artificial Intelligence Lab Publications; 1985.
[101] Fod A, Matarić M, Jenkins O. Automated derivation of primitives for movement classification. Auton. Rob. 2002;12:39–54.
[102] Kawashima H, Matsuyama T. Multiphase learning for an interval-based hybrid dynamical system. IEICE Trans. Fundam. Electron. Commun. Comput. Sci. 2005;E88-A:3022–3035.
[103] Barbič J, Safonova A, Pan J, et al. Segmenting motion capture data into distinct behaviors. In: Proceedings of Graphics Interface GI; London; 2004. p. 185–194.
[104] Li Y, Wang T, Shum H. Motion texture: a two-level statistical model for character motion synthesis. In: Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques; San Antonio,TX; 2002. p. 465–472.
[105] Okada M, Nakamura D, Nakamura Y. Selforganizing symbol acquisition and motion generation based on dynamics-based information processing system. In: Proceedings of the Second International Workshop on Man-machine Symbiotic Systems; Kyoto,Japan; 2004. p. 219–229.
[106] Kadone H, Nakamura Y. Segmentation, memorization, recognition and abstraction of humanoid motions based on correlations and associative memory. In: IEEE-RAS40 International Conference on Humanoid Robotics; Genoa, Italy; 2006. p. 1–6.
[107] Chiappa S, Peters J. Movement extraction by detecting dynamics switches and repetitions. In: Advances in Neural Information Processing Systems (NIPS); Vancouver; 2010. p. 388–396.
[108] Takano W, Imagawa H, Kulić D, et al. Organization of behavioral knowledge from extraction of temporal-spatial features of human whole body motions. In: 3rd IEEE RAS and EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob); Tokyo; 2010. p. 52–57.
[109] Kulić D, Ott C, Lee D, et al. Incremental learning of full body motion primitives and their sequencing through human motion observation. Int. J. Rob. Res. 2012;31:330–345.
[110] Takano W, Nakamura Y. Integrating whole body motion primitives and natural language for humanoid robots. In: IEEE-RAS International Conference on Humanoid Robots (Humanoids); Daejon, Korea; 2008. p. 708–713.
[111] Takano W, Nakamura Y. Incremental learning of integrated semiotics based on linguistic and behavioral symbols. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); St. Louis; 2009. p. 2545–2550.
[112] Fox EB, Sudderth EB, Jordan MI, et al. A sticky HDP-HMM with application to speaker diarization. Ann. Appl. Stat. 2009;5:1020–1056.
[113] Taniguchi T, Nagasaka S, Hitomi K, et al. Semiotic prediction of driving behavior using unsupervised double articulation analyzer. In: IEEE Intelligent Vehicles Symposium (IV); Madrid; 2012. p. 849–854.
[114] Taniguchia T, Nagasaka S, Hitomi K, et al. Unsupervised hierarchical modeling of driving behavior and prediction of contextual changing points. IEEE Trans. Intell. Trans. Syst. 2014;16:1746–1760.
[115] Nagasaka S, Taniguchi T, Yamashita G, et al. Finding meaningful robust chunks from driving behavior based on double articulation analyzer. In: IEEE/SICE International Symposium on System Integration (SII); Fukuoka, Japan; 2012. p. 535–540.
[116] Bando T, Takenaka K, Nagasaka S, et al. Unsupervised drive topic finding from driving behavioral data. In: IEEE Intelligent Vehicles Symposium (IV); London; 2013. p. 177–182.
[117] Bando T, Takenaka K, Nagasaka S, et al. Automatic drive annotation via multimodal latent topic model. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); Tokyo; 2013. p. 2744–2749.
[118] Takenaka K, Bando T, Nagasaka S, et al. Drive video summarization based on double articulation structure of driving behavior. In: ACM Multimedia (ACMM); Nara, Japan; 2012. p. 1169–1172.
[119] Taniguchi T, Nagasaka S, Hitomi K, et al. Sequence prediction of driving behavior using double articulation analyzer. IEEE Trans. Syst. Man Cybern.: Syst. 2015;99: p. 1
[120] Mori M, Takenaka K, Bando T, et al. Automatic lane change extraction based on temporal patterns of symbolized driving behavioral data. In: IEEE Intelligent Vehicles Symposium (IV); Barcelona; 2015.
[121] Wolpert DM, Kawato M. Multiple paired forward and inverse models for motor control. Neural Networks. 1998;11:1317–1329.
[122] Demiris Y, Khadhouri B. Hierarchical attentive multiple models for execution and recognition of actions. Rob. Auton. Syst. 2006;54:361–369.
[123] Taniguchi T, Sawaragi T. Assimilation and accommodation for self-organizational learning of autonomous robots: proposal of dual-schemata model. Vol. 1, In: IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA); Kobe, Japan; 2003. p. 277–282.
[124] Taniguchi T, Sawaragi T. Self-organization of inner symbols for chase: symbol organization and embodiment. Vol. 2, In: IEEE International Conference on Systems, Man and Cybernetics (SMC); The Hague; 2004. p. 2073–2078.
[125] Tagniguchi T, Sawaragi T. Incremental acquisition of multiple nonlinear forward models based on differentiation process of schema model. Neural Networks. 2008;21:13–27.

[126] Samejima K, Katagiri K, Doya K, et al. Symbolization and imitation learning of motion sequence using competitive modules. IEICE Trans. Inform. Syst. 2002;85:90–100.
[127] Wolpert D, Doya K, Kawato M. A unifying comuputational framework for motor control and social interaction. Philos. Trans. R. Soc. B: Biol. Sci. 2003;358:593–602.
[128] Doya K, Samejima K, Katagiri Ki, et al. Multiple model-based reinforcement learning. Neural Comput. 2002;14:1347–1369.
[129] Taniguchi T, Sawaragi T. Incremental acquisition of behaviors and signs based on a reinforcement learning schemata model and a spike timing-dependent plasticity network. Adv. Rob. 2007;21:1177–1199.
[130] Haruno M, Wolpert D, Kawato M. Hierarchical MOSAIC for movement generation. Int. Congress Ser. 2003;1250:575–590.
[131] Tani J, Ito M, Sugita Y. Self-organization of distributedly represented multiple behavior schemata in a mirror system: reviews of robot experiments using RNNPB. Neural Networks. 2004;17:1273–1289.
[132] Heinrich S, Magg S, Wermter S. Analysing the multiple timescale recurrent neural network for embodied language understanding. Vol. 4, In: Koprinkova-Hristova P, Mladenov V, Kasabov NK, editors. Artificial neural networks. Berlin: Springer International Publishing; 2015, p. 149–174.
[133] Murata S, Yamashita Y, Arie H, et al. Generation of sensory reflex behavior versus intentional proactive behavior in robot learning of cooperative interactions with others. In: Joint IEEE International Conferences on Development and Learning and Epigenetic Robotics (ICDL-EPIROB); Genoa, Italy; 2014. p. 242–248.
[134] Roy N, Thrun S. Coastal navigation with mobile robots. In: Advances in neural processing systems (NIPS); Denver; 1999.
[135] Zuo X, Iwahashi N, Taguchi R, et al. Detecting robot-directed speech by situated understanding in object manipulation tasks. In: IEEE International Workshop on Robot and Human Interactive Communication (IEEE RO-MAN); Viareggio, Italy; 2010. p. 608–613.
[136] Sugiura K, Iwahashi N, Kawai H, et al. Situated spoken dialogue with robots using. Adv. Rob. 2011;25:2207–2232.
[137] Denzler J, Brown CM. Information theoretic sensor data selection for active object recognition and state estimation. IEEE Trans. Pattern Anal. Mach. Intell. 2002;24:1–13.
[138] Krainin M, Curless B, Fox D. Autonomous generation of complete 3D object models using next best view manipulation planning. In: IEEE International Conference on Robotics and Automation (ICRA); Shanghai; 2011. p. 5031–5037.
[139] Eidenberger R, Scharinger J. Active perception and scene modeling by planning with probabilistic 6D object poses. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); Taipei, Taiwan; 2010. p. 1036–1043.
[140] van Hoof H, Kroemer O, Ben Amor H, et al. Maximally informative interaction learning for scene exploration. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); Algarve, Portugal; 2012. p. 5152–5158.
[141] Rebguns A, Ford D, Fasel I. InfoMax control for acoustic exploration of objects by a mobile robot. In: AAAI11 Workshop on Lifelong Learning; San Francisco; 2011. p. 22–28.
[142] Burgard W, Fox D, Thrun S. Active mobile robot localization. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI); Nagoya, Japan; 1997. p. 1346–1352.
[143] Stachniss C, Grisetti G, Burgard W. Information gain-based exploration using Rao-Blackwellized particle filters. Robotics Science and Systems (RSS); Cambridge, MA; 2005.
[144] Gouko M, Kobayashi Y, Kim CH. Online exploratory behavior acquisition of mobile robot based on reinforcement learning. In: Recent Trends in Applied Artificial Intelligence; Berlin: Springer; 2013. p. 272–281.
[145] Saegusa R, Natale L, Metta G, et al. Cognitive robotics - active perception of the self and others. In: International Conference on Human System Interactions (HSI); Perth, Australia; 2011. p. 419–426.
[146] Ji S, Carin L. Cost-sensitive feature acquisition and classification. Pattern Recogn. 2006;40:1474–1485.
[147] Tuci E, Massera G, Nolfi S. Active categorical perception of object shapes in a simulated anthropomorphic robotic arm. IEEE Trans. Evol. Comput. 2010;14:885–899.
[148] Schneider A, Sturm J, Stachniss C, et al. Object identification with tactile sensors using bag-of-features. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); St. Louis, MO; 2009. p. 243–248.
[149] Singh S, Chentanez N, Barto AG. Intrinsically motivated reinforcement learning. In: Advances in Neural Information Processing Systems (NIPS); Vancouver; 2005. p. 1281–1288.
[150] Schmidhuber J. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). IEEE Trans. Auton. Mental Dev. 2010;2:230–247.
[151] Daoutis M, Mavridis N. Towards a model for grounding semantic composition. In: The 50th Annual Convention of the AISB; London; 2014.
[152] Morris JF. The Principle of Semantic Compositionality. Topoi. 1994;13:11–24.
[153] Sugita Y, Tani J. Learning semantic combinatoriality from the interaction between linguistic and behavioral processes. Adapt. Behav. 2005;13:33–52.
[154] Ogata T, Murase M, Tani J, et al. Two-way translation of compound sentences and arm motions by recurrent neural networks. In: IEEE International Conference on Intelligent Robots and Systems (IROS); San Diego; 2007. p. 1858–1863.
[155] Tuci E, Ferrauto T, Zeschel A, et al. An experiment on behavior generalization and the emergence of linguistic compositionality in evolving robots. IEEE Trans. Auton. Mental Dev. 2011;3:176–189.
[156] Sandamirskaya Y, Zibner SKU, Schneegans S, et al. Using dynamic field theory to extend the embodiment stance toward higher cognition. New Ideas Psychology. 2013;31:322–339.

[157] Hinaut X, Petit M, Pointeau G, et al. Exploring the acquisition and production of grammatical constructions through human-robot interaction with echo state networks. Front. Neurorobotics. 2014;8:1–17.
[158] Hinaut X, Lance F, Droin C, et al. Corticostriatal response selection in sentence production: Insights from neural network simulation with reservoir computing. Brain Lang. 2015;150:54–68.
[159] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. In: C Burges, L Bottou, M Welling, Z Ghahramani, K Weinberger, editors. Advances in Neural Information Processing Systems (NIPS); Lake Tahoe, Nevada; 2013. p. 3111–3119. Available from: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrasesand-their-compositionality.pdf.
[160] Mikolov T, Corrado G, Chen K, et al. Efficient estimation of word representations in vector space. In: International Conference on Learning Representations (ICLR); Scottsdale, Arizona; 2013. p. 1–12.
[161] Mikolov T, Yih Wt, Zweig G. Linguistic regularities in continuous space word representations. In: The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL); Atlanta, Georgia; 2013. p. 746–751.
[162] Le Q, Mikolov T. Distributed representations of sentences and documents. 32, In: International Conference on Machine Learning (ICML); Beijing; 2014. p. 1188–1196.
[163] Ugur E, Piater J. Bottom-up learning of object categories, action effects and logical rules: from continuous manipulative exploration to symbolic planning. In: IEEE International Conference on Robotics and Automation (ICRA); Seattle, Washington; 2015. p. 2627–2633.
[164] Stückler J, Behnke S. Integrating indoor mobility, object manipulation, and intuitive interaction for domestic service tasks. In: IEEE-RAS International Conference on Humanoid Robots (HUMANOIDS); Paris; 2009. p. 506–513.
[165] Iocchi L, Holz D, Ruiz-del Solar J, et al. RoboCup@Home: analysis and results of evolving competitions for domestic and service robots. Artif. Intell. 2015;1:1–24.
[166] Nakamura T, Attamimi M, Sugiura K, et al. An extended mobile manipulation robot learning novel objects. J. Intell. Rob. Syst. 2012;66:187–204.
[167] Stückler J, Droeschel D, Gräve K, et al. Towards robust mobility, flexible object manipulation, and intuitive multimodal interaction for domestic service robots. Vol. 7416, Lecture notes in computer science. In RoboCup 2011: Robot Soccer World Cup XV. Berlin: Springer; 2012. p. 51–62.
[168] Inamura T, Shibata T, Sena H, et al. Simulator platform that enables social interaction simulation - SIGVerse: socioIntelliGenesis simulator. In: IEEE/SICE International Symposium on System Integration (SII); Sendai, Japan; 2010. p. 212–217.
[169] Ohno K, Yamazaki K, Shimosaka M. Data engineering robotics –sensor data produce new robot intelligence. J. Rob. Soc. Jpn. 2015;33:97–99. Japanese.