20 Language and Communication

Angelo Cangelosi and Tetsuya Ogata

20.1 Introduction

Communication is a rich multimodal process combining spoken language with a variety of nonverbal behaviors such as gaze, gestures, tactile interaction, and emotional cues (Mavridis 2015; Cangelosi and Ogata 2019; Liu and Zhang 2019). For cognitive robotics and human-robot interaction, linguistic and nonverbal communication skills are fundamental cognitive capabilities necessary to interact with people. To ask a humanoid robot to perform a specific task, or to engage in a dialogue with a social robot companion, both people and robots must possess a language-like communication system. In cognitive robotics, the design of speech and nonverbal communication skills is directly inspired by communication in people.

The organization of human language and communication has been the focus of attention in linguistics and psychology. Specific levels of representation and analyses of linguistics skills, ranging from the processing of low-level phonetic features to higher-level communicative and pragmatic processes, have been identified to study language. In addition, developmental psychology has significantly contributed to the identification of the developmental stages and language-learning principles. This has been contextualized within the debate of nativist versus constructivist theories—that is, language acquisition theories giving emphasis to a genetic predisposition to language-related competence versus developmental theories stressing the role of environmental factors. These linguistics and psychology analyses have significantly contributed to the design of cognitively inspired language and communication skills in cognitive robots. Below, we first look at the developmental theories of language learning and the linguistics approach of natural language processing (NLP) and the five levels of analysis. This will inform the discussion of the different models of language acquisition in developmental robotics, of NLP models used in robots, and of the more recent machine-learning models.

20.1.1 Language Development and Learning in Humans

An important issue in language development research is the "nature" versus "nurture" debate. This is the debate between the "nativists," who hypothesize that babies are born with language-specific knowledge and skills, and the "empiricists," who propose that

babies construct linguistic knowledge through interaction with their social, language-speaking community. Within the nativist position, influential theories have proposed that there are universal syntactic rules and generative grammar principles (e.g., Chomsky's "brain organ" and "language acquisition device" hypothesis) and that these are innate in the human brain (Chomsky 1965). On the contrary, according to the nurture stance, the essence of linguistic knowledge emerges from language use during development, without any need to assume the existence of innate language-specific knowledge. This empiricist view of language development is also known as the constructivist, usage-based theory of language development (Tomasello 2003; MacWhinney 1998). The child is seen as an active constructor of their own language system through the implicit observation and learning of statistical regularities and logical relationships between the meaning of words and the words used (e.g., cognitive linguistic theories of Goldberg [2006]).

In developmental psychology research, the most significant phenomena of language acquisition occur during the first four years. The early milestones of language development follow the parallel and intertwined development of incremental phonetics-processing capabilities, increasing lexical and grammatical repertoires, and refined communicative and pragmatic faculties. Table 20.1 provides an overview of the main milestones of language development (Hoff 2013; Cangelosi and Schlesinger 2015).

In the first year, the most evident sign of linguistic development concerns phonetic capabilities such as vocal babbling. Babbling initially consists of vocal play with sounds such as cooing, squeals, and growls ("marginal babbling") and later consists of the repetition of language-like syllabic sounds such as "dada" or "bababa" (canonical/reduplicated babbling). Toward the end of the first year, children also start to produce communicative gestures (e.g., pointing) and iconic gestures (e.g., raising the fist to the ear to mean telephone). This is hypothesized to demonstrate the child's prelinguistic intentional communication and cooperation skills (Tomasello, Carpenter, and Liszkowski 2007).

Toward the beginning of the third year, the child starts to develop more complex grammatical constructs and skills. This is the case, for example, of the "verb islands" phenomenon (Tomasello 1992). Initially, children can use a variety of verbs and treat them as independent syntactic elements called "verb islands" (e.g., the child only uses very simple syntactic combinations of the same verb with different nouns of objects: "cut bread," "cut paper"). These intermediate syntactic constructions allow the child to subsequently develop more refined morphological and syntactic constructs, with more general verb islands combined with a richer set of prepositions. From the fourth year of age, the child gradually develops adultlike syntactic constructions such as simple transitives (agent-verb-patient, as in "John likes sweets") and locatives (agent-verb-patient-locative-location, as in "John puts sweets on table"; Tomasello and Brooks 1999). This gradually leads to the development of ever-more complex syntactic-morphologic constructions, more abstract and generalized grammatical categories known as word classes. These syntactic skills are accompanied by extended pragmatic and communicative skills, leading to refined narrative and discursive capabilities.

The constructivist view of language is highly consistent with the embodied and situated cognition theories (Pezzullo et al. 2013) and the relevant embodied robotics approach to the modeling of language learning (Cangelosi 2010, 2011). This embodied view stresses

Table 20.1
Typical timescale and major milestones of language development

Age (months)	Competence
0–6 months	Marginal babbling
6–9 months	Canonical (reduplicated) babbling
10–12 months	Intentional communication First gestures
12 months	Single words, holophrases Word-gesture combinations
18 months	Reorganization of phonological representations 50+ word lexicon size, vocabulary spurt Two-word combinations
24 months	Increasingly longer multiple-word sentences Verb islands
36+ months	Adultlike grammatical constructions Narrative skills

Source: Adapted from Cangelosi and Schlesinger 2015.

the fact that the body of the child, and its interaction with the environmental context, determines the type of representations, internal models, and cognitive strategies learned.

In cognitive robotics models, the embodied approach is linked to that of "symbol grounding" (Harnad 1990; Cangelosi 2010) and "grounded cognition" (Pezzulo et al. 2013). This refers to the capability of natural and artificial cognitive agents to acquire an intrinsic (autonomous) link between internal symbolic representations and referents in the external world or internal states. Cognitive robotics models implement the grounded learning of associations between words and the external and internal entities they refer to (objects, actions, internal states).

20.1.2 Levels of Analysis in Language Studies

In linguistics and psychology, a hierarchy of five levels of language analyses has been proposed: phonetic, lexical, semantic, syntactic, and pragmatic (see Cangelosi 2017). These levels are useful in cognitive robotics models because they identify the different aspects that need to be modeled and implemented to successfully achieve humanlike linguistic capabilities. For example, a robot, like a person, must be able to recognize language-specific sounds (phonetic level) to segment and identify the words (lexical level) and the grammatical structure of spoken utterances (syntactic level). This supports the understanding of the meaning of words and sentences (semantic level) and their contextualization within the interactive communication task (pragmatic level). These different levels of analysis should not, however, be considered separate modular components of language-processing models. In fact, all levels of language are strictly intertwined. For example, knowledge of the lexicon helps the lower-level recognition of phonemes and words. The pragmatic level of communication can also prime the recognition of the words and sentences that the hearer expects the speaker to choose to communicate the intended meaning.

Cognitive robotics models of language benefit from the field of natural language processing (NLP), which uses a set of computation linguistics methods for the different levels

of analysis and the representation of language. Numerous NLP methods and software tools have been proposed for phonetic analysis and automatic speech recognition (e.g., Markov models), for lexical and semantic analysis (e.g., WordNet), for parsing and syntactic analysis, and for pragmatics and communication (e.g., dialogue systems). This field has very recently gone through a significant revolution with the use of deep-learning models (cf. chapter 5). For example, deep neural networks are used for state-of-the-art speech recognition systems and parsing and word tagging (LeCun et al. 2015). These changes include the increasing use of end-to-end (a.k.a seq-to-seq—i.e., sequence-to-sequence) machine-learning models. These use deep neural networks that receive the raw input (e.g., sound wave or a word list) and, without specifically decomposing the linguistic processing into different levels of analysis or mechanisms, produce the desired output (e.g., translation of the input sentence into another language). In section 20.2 we will look at both NLP and the deep-learning models used in language systems for cognitive robots.

20.2 Robot Language Models

In robot language research, we can distinguish three main approaches to the design of language communication capabilities in robots (Cangelosi and Ogata 2019). The first directly models incremental, developmental phenomena on language acquisition. This is primarily based on developmental robotics approaches (chapter 3). Another approach is based on various NLP techniques, while the third focuses on the latest machine-learning approaches (chapter 9). The NLP approach typically combines off-the-shelf techniques and language-processing tools (e.g., ready-made lexicons and knowledge bases, parsers, automatic speech recognition, and speech synthesis software) to implement in the robot the ability to respond to linguistic instructions and to utter sentences to express a request. The language-learning approach, on the other hand, uses machine-learning methods (e.g., neural networks, Bayesian methods) to train the robot to acquire language skills. In practice, however, some NLP robotic approaches do use machine-learning methods (e.g., most of the current speech recognition systems are based on statistical learning and deep neural network methods), and some robot language-learning approaches partially rely on off-the-shelf NLP tools.

20.2.1 Developmental Robot Language Models

Developmental language-learning models are typically based on the developmental robotics approach (Cangelosi and Schlesinger 2015; see also chapter 3). As such, this approach puts a strong emphasis on constraining the robot's cognitive and linguistic architecture and behavioral and learning performance to known child psychology theories, data, and developmental principles. This permits the modeling of the developmental sequence of the qualitative and quantitative stages leading to the acquisition of adultlike sensorimotor, cognitive, and linguistic skills. Developmental robotics is also naturally suited to model embodied and situated cognition for the grounding of cognition (Pezzulo et al. 2013). Specifically, for the embodied bases of language learning, the use of robots that have to learn to name objects they see and name actions they perform constitutes an ideal way to model the grounding of symbols in sensorimotor knowledge and experience (Harnad 1990; Cangelosi 2011).

Some developmental robotics models focus on the acquisition and grounding of the first words. These models directly rely on child psychology studies on language acquisition in infants in the second year of age—that is, when the first words are acquired. One seminal developmental model is that of Morse et al. (2010, 2015), as it directly replicates child psychology data on embodied language acquisition via body posture interaction (Samuelson et al. 2011). In Samuelson et al.'s (2011) child psychology study, the infant repeatedly experiences two new objects (the target and the foil) in different locations (left/right), requiring a postural change to attend to the object. Subsequently, the child hears the object name "modi" while attending to a foil object that has been placed in the location normally associated with the target object. When the infant is asked, "Where is the modi?," they select the target object—that is, the object normally associated with the posture and spatial location they were attending to, rather than the actual object they were looking at when they heard the name. This means that infants rely on memory for their own posture and the related object location to associate objects and their names.

Morse et al. (2015) have proposed an embodied model of this phenomenon with the iCub humanoid robot, replicating the original experiments by Samuelson et al. and further exploring how this spatial component can be achieved via the robot's physical interaction with objects and locations. The model is an implementation of the epigenetic robotics architecture (Morse et al. 2010), a developmental robotics cognitive architecture specifically designed for studying embodied language learning. The core of such an architecture consists of three self-organizing maps with Hebbian connections between their units (figure 20.1). The first (visual) map is used to represent in a topological way the similarity of preprocessed visual information (e.g., color and/or shape) implemented as input of a spectrogram of the color of each object in view. The second (body) map is driven by postural information (the current motor encoder values of the eyes, head, and torso of the robot). The final (word) map responds to each word encountered (preprocessed by a standard NLP speech recognition system). The visual color map and the word map are both fully connected to the body posture map, with connection weights adjusted by a normalized positive and negative Hebbian learning rule.

In one version of the experiment, the target object (a red ball) is placed to the left of the iCub. The robot looks at the target for approximately ten seconds before the target object is removed, and the foil object is placed to the right of the iCub, which again orients for approximately ten seconds. This procedure is repeated four times. In the fifth presentation cycle, the foil object is placed in the position normally associated with the target object, and the word "modi" is spoken. The original placements of each object are repeated one final time, and then both objects are positioned in new locations to test the robot by stating, "Find the modi." The robot then orients and reaches for one of the objects. Various versions of the experiment were carried out, each repeated twenty times (with all prelearning weights randomly initialized). Morse et al. (2015) conducted an additional experiment following the same procedure outlined above but with the addition of another spatial dimension of the robot's posture (from sitting to standing) for the naming event only at the fifth presentation cycle. As a result of this change, the naming event occurs in a posture that has not been previously associated with either the target or the foil object. Thus, testing the interference between previously experienced objects and that posture causes the iCub to select the foil object (the object it was observing when it first heard the name). This

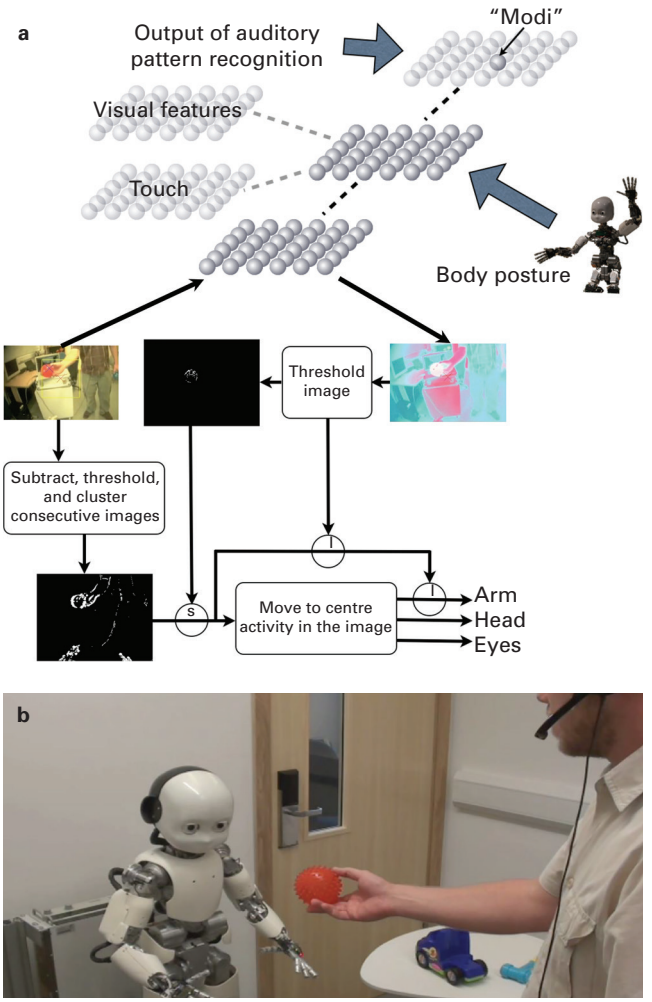

Figure 20.1
Setup for word-learning experiments (a) and cognitive architecture (b) in Morse et al. (2015).

result was also replicated in new child experiments (Morse et al. 2015). Overall, this model shows that infants, like robots, use the memory of postures as a way to organize their learning task. If two different postures are used at this early stage of development, they are used by the robot to separate different cognitive tasks.

An extended version of this model has already been used to replicate a range of other language acquisition phenomena (Morse and Cangelosi 2017; Cangelosi and Schlesinger 2018). For example, Twomey et al. (2016) used the ERA architecture to model mutual exclusivity—that is, the developmental phenomenon in which a child can learn the name of a new object if they hear a new label and are presented with an unseen (unlabeled) object among other objects with a known label. Other developmental language models have looked at the learning of both object and action labels, moving toward the first examples of syntax learning. For example, Tikhanoff et al. (2011) proposed a simulation

model of the iCub robot in the development of a lexicon based on both names of objects and of actions and their basic combinations to understand simple commands such as "pick_up blue_ball."

A few developmental robotics models have focused on grammar development—for example, modeling the emergence of semantic compositionality for syntactic compositionality for multiple word combinations and generalizations (Sugita and Tani 2005; Tuci et al. 2011; Zhong et al. 2019). For example, the robot model by Sugita and Tani (2005) investigated the emergence of compositional meanings and lexicons with no a priori knowledge of any lexical or formal syntactic representations. The environment consisted of three colored objects (red, blue, and green) in three different locations (a red object on the left-hand side of the robot's field of view, a blue object in the middle, and a green object on the right). The robot could respond with nine possible behaviors based on the combination of three actions (POINT, PUSH, HIT) with the three objects (RED, BLUE, GREEN) always in the same locations (LEFT, CENTER, RIGHT). The robot learning architecture was a parametric bias recurrent neural network (PBRNN), which is capable of learning a set of parametric bias units able to represent action sequences via language-like symbols. The robot experiments were divided into two stages: training and generalization. In the training phase, the robot acquired associations between given sample training sentences and corresponding behavioral sequences. In the testing phase, the robot's ability to generate the correct behavior by recognizing the sentences used during training and, above all, novel combinations of words was tested. A subset of fourteen object/action/location combinations was used during training, with four left for the generalization test. After the successful training stage, in the generalization test phase the four remaining novel sentences were given to the robot: "Point green," "point right," "push red," and "push left." Behavioral results showed whether the linguistic module had acquired the underlying compositional syntax correctly. The robot could generate grammatically correct sentences and understand them by giving a behavioral demonstration of the generalized actions. Detailed analyses of the robot's neural representations supporting the verb-noun compositional knowledge showed a separated substructure for the verbs and nouns. In particular, the congruence in the substructures for verbs and nouns indicated that the combinatorial semantic/syntactic structure was successfully extracted by the robot's neural network.

Yamashita and Tani (2008) proposed an extension of this work using the multiple-timescale recurrent neural network (MTRNN) for compositional action and language learning experiments. Zhong et al. (2019) further extended the MTRNN architecture to control the compositional learning and generalization of nine actions on nine objects for verb-noun learning in the iCub robot.

Developmental learning models have also been proposed to investigate the acquisition of abstract concepts and words in robots, including words referring to general-purpose motor actions such as "use" and "make" and number and counting words (Cangelosi and Stramandinoli 2018). To model the grounding and embodied bases of abstract word learning in cognitive robots, one study looked at abstract action verbs such as "to use," which can be applied to different motor contexts (e.g., "use a hammer" or "use a pen") with no common motor program. The developmental robotics model of Stramandinoli et al. (2017) exploits the hierarchical recursive structures of both the linguistic and the motor system to integrate simple motor primitives and concrete words to create the semantic referents

of abstract action words that do not have a direct mapping to the sensorimotor world. An iCub robot is first trained to recognize a set of tools of different colors, sizes, and shapes (e.g., knife, hammer, brush) and to perform object-related actions (e.g., cut, hit, paint). Subsequently, the robot is taught to name these objects and actions (e.g., "cut with knife"). Finally, the robot is taught the abstract motor words of "use" and "make" by combining these new action words with the appropriate tool name (e.g., "use knife"). The experiments investigated the effects of using different combinations of the three input modalities (i.e., vision, language, and proprioception). For example, incompatible condition tests between the perceptual and linguistic input showed that the robot ignored the linguistic command by executing the actions elicited by the seen objects. Hence, the knowledge associated with objects relies not only on the objects' perceptual features but also on the actions that can be performed on them (i.e., affordances). Further simulation experiments showed that the acquisition of concepts related to abstract action words (e.g., "use knife") requires the reactivation of similar internal representations of the network activated during the acquisition of the concrete concepts (e.g., "cut with knife") contained in the linguistic sequences used for the grounding of abstract action words (e.g., "use knife" is "cut with knife"). This finding suggests that the semantic representation of abstract action words requires the recall and reuse of sensorimotor representational capabilities (i.e., embodied understanding of abstract language). Indeed, neurophysiological evidence of the modulation of the motor system during the comprehension of both concrete and abstract language exists to support this finding.

Finally, developmental models with humanoid robots have also been used to model abstract concepts and the representation of the underlying knowledge of numbers. Number cognition is another key example of the contribution of embodied cognition in the acquisition of abstract, symbol-like manipulation capabilities. Various embodied strategies, such as pointing and counting gestures, object touching, and finger counting, have been shown to facilitate the development of number cognition skills (e.g., Alibali and DiRusso 1999; Moeller et al. 2011). Given the implicit embodied nature of humanoid robots, some recent models have specifically looked at the modeling of the acquisition of number concepts and words via embodied strategies such as gestures (Ruciński et al. 2012) and finger counting (De La Cruz et al. 2014; Pecyna et al. 2020). For example, a developmental robotics model was used specifically to explore whether finger counting and the association of number words to each finger could bootstrap the representation of numbers in a cognitive robot. This study used a recurrent artificial neural network to model the learning of associations between (motor) finger counting, (visual) object counting, and (auditory) number word and sequence learning. In particular, this study manipulated the coupling between different modalities, such as with the comparison of the Auditory-Only condition, when the robot solely learns to hear and repeat the sequence of number words ("one," "two," . . . up to "ten"), with the Finger + Auditory condition, when the robot simultaneously learns the sequence of acoustic number words and moving fingers.

The results showed that learning the number word sequences together with finger sequencing (Finger + Auditory condition) helps to quickly build the initial representation of numbers in the robot. Robots who only learn the auditor sequences (Auditory-Only condition) achieve the worst performances. Moreover, the neural network's internal representations of these two conditions resulted in qualitatively different patterns of similarity

in the representation between numbers. Only after the Finger + Auditory sequence learning did the network represent the relative distance between numbers, which corresponded to the quantitative difference between numbers. In Finger + Auditory-trained robots, the cluster analysis diagram of the hidden layer's activation showed that the representation for the word "one" was adjacent to that of "two" and increased differently (distant) from the higher numbers. However, in the auditory-only condition, there was no correspondence between the cluster diagram similarity distance and the numerical distance.

This finger-counting model has recently been extended by Pecyna et al. (2020) to model numerosity estimation and by Di Nuovo and McClelland (2019), who combined developmental robotics and deep-learning methods to show that proprioceptive information from robot hands improves accuracy in the recognition of spoken digits. See chapter 22 for an extended discussion of abstract and number word learning.

20.2.2 NLP-Based Robot Language Models

NLP methods have been used for two different types of robot language models. In the conversational approach, the robot uses NLP tools primarily to engage in a linguistic conversation with a human user for social companionship, entertainment, or information-gathering tasks, with no actual motor tasks to perform (no language grounding required). In human-robot interaction models, robots use language primarily to respond to instructions to perform a physical action.

Conversational robots have their origins in conversational agents and chatterbots, such as the very first conversational agent developed called ELIZA (Weizembaun 1966). More recent conversational agents are often based on animated virtual 3D characters, such as A.L.I.C.E. (Wallace 2009). Conversational agents embodied in physical robots include work with the android robot ERICA (ERato Intelligent Conversational Android; Ishiguro 2016), the Robot-ERA system for supporting older people in independent living (Di Nuovo et al. 2018), and museum/station guides and robot tutors for children (Shiomi et al. 2008; Belpaeme et al. 2018). These conversational robots use a variety of NLP tools for speech recognition, parsing, and dialogue systems.

Many NLP-based robot language systems are designed with the primary function of following a user's instructions and selecting the appropriate motor behavior. These applications typically cover object manipulation tasks (e.g., "pick up blue ball," "clean the table") and navigation scenarios (e.g., "go to the exit," "take me to the restroom"; Mavridis 2015). The use of speech for language instruction understanding requires a tight coupling (grounding) of the robot's visual and motor repertoire with its language processing and knowledge representation methods. In NLP-based approaches, this link is typically predefined by the designer. There is no autonomous grounding of the robot's words via situated learning, as the robot can only use a set of "meanings" defined by the programmer. For example, Aloimonos and Pastra developed a language and action representation formalism, called PRAXICON, for action and language knowledge representation of object manipulation tasks (Pastra and Aloimonos 2012; Pastra 2008). It uses a goal-based representation of actions employing a multimodal semantic network-type representation that is directly inspired by linguistic methods, such as the mapping of a minimalist grammar of language into a minimalist grammar of action representation. PRAXICON was tested on the Baxter robot capable of learning to cook from "watching" videos available on the Web (Yang et al. 2015).

Nonverbal communication capabilities have also been proposed to complement and enhance a robot's linguistic production and communicative expressivity. For example, Csapo et al. (2012) complemented speech production with nonverbal strategies such as face tracking, nodding, gesturing, proximity detection, and interruptions. Mutlu et al. (2012) modeled humanlike gaze mechanisms to help robots signal different interaction roles to the human interlocutor to manage turn exchanges and the dynamics of the conversation.

20.2.3 Machine-Learning Robot Language Models

Multimodal integration, which directly concerns the field of language learning for connecting speech, vision, and action, has long been a difficult problem in robotics. For example, the crossmodal complementation of information loss or the application of crossmodal memory search for behavior generation problems have not been thoroughly studied. Second, literature discussions on how to fuse multimodal information to achieve stable environmental awareness have not reached a comprehensive consensus. In robotics, the sensory input acquired from different sources is still typically processed using a dedicated feature extraction mechanism (Murphy 2019). Third, multimodal synchronization modeling as a means for implementing the sensorimotor prediction of robot applications has not been adequately studied. Several studies so far have proposed a computational model that develops synchronization of behavioral effects in a developmental way toward an understanding of interaction (Kuriyama et al. 2010; Ogino et al. 2006). However, most casual models are expressed using a limited number of modalities and in many cases focus only on vision and behavior.

In recent years, different types of graphic models of multimodal classification have been reported. Lallee and Dominey (2013) proposed a multimodal convergence map based on a self-organizing map (SOM) that integrates visual-motor and language modality. Sinapov and Stoytchev (2011) developed a graph-based model that enables robots to recognize untrained objects based on their similarity to trained objects. They also let the robot take ten different actions to collect visual, auditory, and tactile data; explore one hundred objects; and categorize twenty objects with supervised learning (Sinapov et al. 2014). Ivaldi et al. (2013) developed a robot that can learn object categories by active sensing. Nakamura et al. (2009, 2015) proposed studies on multimodal classification using multilateral latent Dirichlet allocation (MLDA) and its extension. They developed robotic systems that can obtain visual, sound, and tactile information by handling objects. The robot grasps an object several times and shakes the object to acquire sound information. By applying the MLDA, they showed that robots can classify many objects into categories, which is similar to human classification results (Nakamura et al. 2009). Araki et al. (2011) developed an MLDA online and conducted experiments on completely autonomous multimodal category acquisition in the home environment.

Notwithstanding the above multimodal machine-learning examples, there has been little research on scalable learning frameworks for handling a large amount of sensorimotor data of a high dimension. The latest robots are equipped with state-of-the-art sensor devices such as high-resolution image sensors, distance sensors, and multichannel microphones as the demand for perception accuracy with respect to the surrounding environment increases (Kaneko et al. 2008; Sakagami et al. 2002). Thus, a remarkable improvement in the amount of sensorimotor information available has been achieved. However, due to the

scalability limitations of conventional machine-learning algorithms, few computational models achieve robust behavior control and environmental recognition by fusing multimodal perceptual inputs into a single representation. To overcome the problem of the scalability limitation, deep-learning approaches such as deep neural networks (DNNs), used as perceptual feature extraction and multimodal integration learning mechanisms, have attracted the attention of the robotics and machine-learning community in recent years. One of the main advantages of applying a DNN is the ability to self-organize highly generalized sensory functions from large-scale raw data. For example, DNNs have been successfully applied to unsupervised feature learning for a single modality such as text, images, and voice. The same approach has also been applied to the learning of integrated representation among multiple modalities, resulting in a significant improvement in speech recognition performance. In another context using unsupervised learning, Le (2013) showed that DNNs with large-scale data can automatically construct high-level features from image data. Connecting acquired representation by neural networks and multimodal classification is an important research field (Bengio et al. 2013). However, the application of DNNs for more dynamic information such as robot motion and language has just begun to be considered.

Multimodal integration based on DNNs is generally accomplished by two approaches. First, in the feature extraction method, feature vectors from some plural modalities are transformed to acquire an integrated feature vector. For example, Ngiam et al. (2011) utilized a DNN that extracts directly integrated expressions from multimodal signal input by compressing the input dimension. Huang and Kingsbury (2013) used deep belief networks (DBNs) for audiovisual speech recognition tasks by combining intermediate-level features learned by a DBN of a single modality. However, these methods have difficulty explicitly and adaptively selecting their respective information gains in response to dynamic changes in the reliability of multimodal information sources. Alternatively, in the fusion method the outputs of the unimodal classifiers are merged to determine the final classification. Unlike the feature extraction approaches, the fusion methods can improve robustness by incorporating the stream reliability associated with multiple information sources as a measure of the information gain of the recognition model.

Specifically for robot language models, Noda et al. (2015) proposed a speech recognition model that uses a DNN both for noise reduction of speech features and for using visual information in a complementary style. The perception features acquired from the audio signal and the corresponding mouth region image are then integrated. Two kinds of DNNs, a deep denoising autoencoder (DDA) and a convolution neural network (CNN), are used for the feature extraction of audio information and visual information, respectively. In addition, the multistream hidden Markov model (MSHMM) is applied to integrate the two perceptual features acquired from the speech signal and the mouth region image. They show that the CNN outputs higher recognition rates than the visual features extracted by PCA (principal component analysis), and the effect of the different image resolutions is not prominent. The word recognition rate, visual features acquired by the CNN, is approximately 22.5 percent.

The DNN language and multimodal integration models provide intuitive and direct ways to accomplish temporal sequence recognition tasks. The focus of the task is to "recognize" by symbolizing the raw sensory signal. However, since recognition methods using

probabilistic models specialize in obtaining symbolic representation from the raw signal, they are not suitable for sensorimotor coordination tasks, such as robot behavior generation. Therefore, this approach needs to design external mechanisms to generate behaviors corresponding to the recognized state. To address this, Heinrich et al. (2015) utilized multiple timescale recurrent neural networks (MTRNN) to integrate visual, auditory, and motor information.

Noda et al. (2014) also proposed a multimodal temporal sequence integration learning framework using a DNN for multimodal time series integrated learning, as well as feature extraction by dimensional compression. They showed the framework with multiple DNNs as a crossmodal memory retriever and as a temporal sequence predictor. Specifically, they integrated image, sound signal, and motor modalities with multiple deep autoencoders (DAs). The learning experiments were conducted on six types of object manipulations by the humanoid robot NAO, generated by direct teaching. The data of high dimension, such as images and sound signals, are compressed to thirty dimensions by the DA. The image and sound data obtained from this process and the motor command obtained from the robot are integrated using a DA instead of an HMM. The data were extracted within a sliding time window of thirty steps. Results showed that this model self-organizes not only the sensory features but also the motion patterns from the time series of sensorimotor data corresponding to the plural robot motions. The principal component analysis of the acquired internal representation showed that each motion does not correspond to the motion cluster designed by the human teacher. Some motions have multiple clusters reflected by the characteristics of the learning condition. Some motions overlap the other motions, thereby associating with each other. Thus, the real world, the body structure, and the learning model self-organize the expressions of behaviors coupled with recognition. They realized a crossmodal memory association by using this internal representation. For example, robot motion is generated from images and sound data; the visual image (movie) is produced from body motion or sound. This demonstrates a significant advantage of using DNN multimodal learning to generate expressions of a very large dimension.

DNNs have also been used to extend developmental models of language learning, integrating recurrent neural networks, such as long short-term memory (LSTM), with simultaneous action and language processing. For example, Antunes et al. (2019) used a bidirectional multiple timescales LSTM for the grounding of actions and verbs without explicitly learning an intermediate representation. The model self-organizes such representations at the level of a slowly varying latent layer connecting the language and the action route (figure 20.2). The model is also trained in a bidirectional way, learning how to produce a sentence from a certain action sequence input and, simultaneously, how to generate an action sequence given a sentence as input. This network was evaluated on motor actions performed by an iCub robot and their corresponding letter-based description. Yamada et al. (2017) also used recurrent DNNs to train a robot to translate sentences that included logic words, such as "not," "and," and "or," into robot actions. The model analysis showed that referential words are merged with visual information and the robot's own current state, while logical words are represented by the model in accordance with their functions as logical operators.

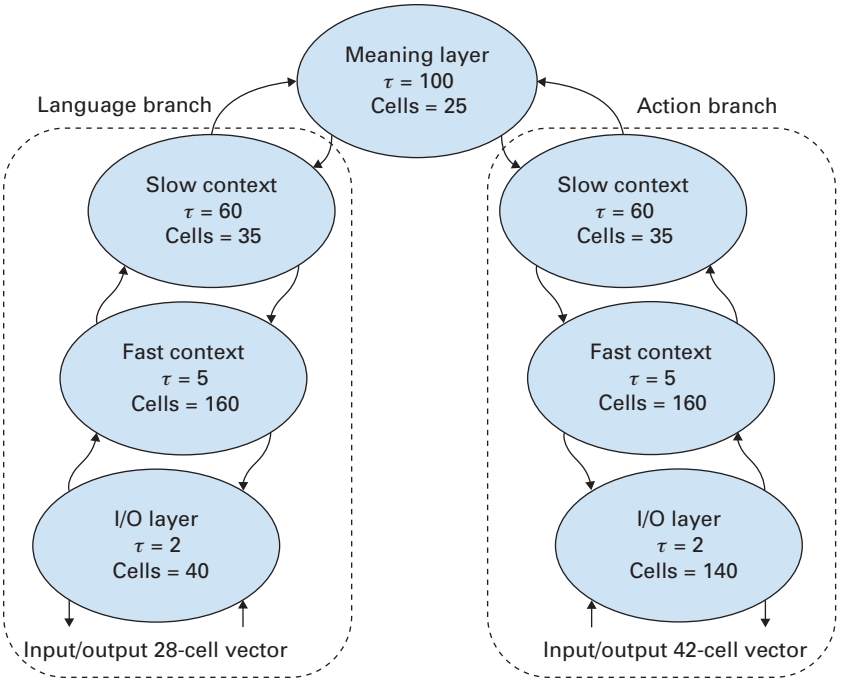

Figure 20.2
A bidirectional LSTM for action and language learning. *Source:* Adapted from Antunes et al. 2019.

20.3 Conclusion

This chapter summarizes the cognitively inspired approaches to the design of language learning and language grounding and processing capabilities in robots. Developmental language-learning models have been able to replicate humanlike developmental trajectories in the early acquisition of words and simple grammatical structures. They also exploit embodied strategies, such as posture bias and finger-counting skills, in learning and grounding concrete and abstract words. However, the level of complexity of the robot's language repertoire is limited to small lexicons. NLP-based models, on the other hand, have been widely used to handle dialogue with conversational agents and complex lexicons. However, in these models the robot is not able to autonomously ground the words it uses for sensorimotor knowledge, and it must rely on the hand coding of the word-meaning mappings defined by the system designer.

An important development in robot language research is the very recent progress on learning methods for language and multimodal information based on machine-learning models. However, on its own, DNN cannot address the whole problem of robot language grounding. For example, deep learning takes a batch-learning and a supervised-learning approach, and generally, it cannot work online. It acquires representations approximating the given input data, and it cannot easily define novel symbols (and meanings) about the world, as humans do with language generativity. It is also important to acknowledge that although DNNs can match human performance in some particular data-processing tasks,

they do have significant limitations. The most critical issue with DNNs for robot language models is that it is extremely challenging to understand a DNN's internal mechanism. Even when high performance is achieved, it is difficult to identify the cause when a mistake occurs. This is a serious problem in the behavior learning of real-world systems such as an interactive robots or automatic driving cars. In DNNs, the internal representation is embedded not only in its large structure but also in its small structure. These mechanisms enable DNNs to self-organize very large and complicated structures of data and to show high performance rates. However, simple statistical analysis and modeling are not directly effective for explaining the mechanism of deep learning. Thus, a mathematical understanding of the DNN as a multidimensional complex system—that is, a dynamic system—is an important area for future work that will have significant implications for the use of deep learning in robot language models.

Finally, an important direction for future research is to focus on a developmental approach, where symbol acquisition emerges from the incremental interaction between the robot, the human user, and their environment. This requires the long-term and open-ended development of a human-robot interaction and communications system that allows a developmental learning robot to bootstrap its multimodal, grounded language-learning skills and repertoire.

Additional Reading and Resources

- An extensive position paper proposing a developmental robotics approach to communication and language integration: Cangelosi, Angelo, Giorgio Metta, Gerhard Sagerer, Stefano Nolfi, Chrystopher Nehaniv, Kerstin Fischer, Jun Tani, et al. 2010. "Integration of Action and Language Knowledge: A Roadmap for Developmental Robotics." *IEEE Transactions on Autonomous Mental Development* 2 (3): 167–195.
- A comprehensive paper on the symbol-emergence approach to language development modeling: Taniguchi, Tadahiro, Takayuki Nagai, Tomoaki Nakamura, Naoto Iwahashi, Tetsuya Ogata, and Hideki Asoh. 2016. "Symbol Emergence in Robotics: A Survey." *Advanced Robotics* 30 (11–12): 706–728.
- A recent extensive review of language and speech models for humanoid robotics: Cangelosi, Angelo, and Tetsuya Ogata. 2019. "Speech and Language in Humanoid Robots." In *Humanoid Robotics: A Reference*, edited by P. Vadakkepat and A. Goswami. Berlin: Springer.

References

Alibali, Martha Wagner, and Alyssa A. DiRusso. 1999. "The Function of Gesture in Learning to Count: More than Keeping Track." *Cognitive Development* 14 (1): 37–56.

Antunes, Alexandre, Alban Laflaquière, Tetsuya Ogata, and Angelo Cangelosi. 2019. "A Bi-directional Multiple Timescales LSTM Model for Grounding of Actions and Verbs." In *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2614–2621. New York: IEEE.

Araki, Takaya, Tomoaki Nakamura, Takayuki Nagai, Kotaro Funakoshi, Mikio Nakano, and Naoto Iwahashi. 2011. "Autonomous Acquisition of Multimodal Information for Online Object Concept Formation by Robots." In *Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1540–1547. New York: IEEE.

Belpaeme, Tony, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. "Social Robots for Education: A Review." *Science Robotics* 3 (21).

Bengio, Yoshua, Aaron Courville, and Pascal Vincent. 2013. "Representation Learning: A Review and New Perspectives." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (8): 1798–1828.

Cangelosi, Angelo. 2010. "Grounding Language in Action and Perception: From Cognitive Agents to Humanoid Robots." *Physics of Life Reviews* 7 (2): 139–151.

Cangelosi, Angelo. 2011. "Solutions and Open Challenges for the Symbol Grounding Problem." *International Journal of Signs and Semiotic Systems* 1 (1): 49–54.

Cangelosi, Angelo. 2017. "Language Processing." In *From Neuron to Cognition via Computational Neuroscience*, edited by M. Arbib and J. Bonaiuto, 693–718. Cambridge, MA: MIT Press.

Cangelosi, Angelo, and Tetsuya Ogata. 2019. "Speech and Language in Humanoid Robots." In *Humanoid Robotics: A Reference*, edited by P. Vadakkepat and A. Goswami, 2261–2292. Berlin: Springer.

Cangelosi, Angelo, and Matthew Schlesinger. 2015. *Developmental Robotics: From Babies to Robots*. Cambridge, MA: MIT Press.

Cangelosi, Angelo, and Matthew Schlesinger. 2018. "From Babies to Robots: The Contribution of Developmental Robotics to Developmental Psychology." *Child Development Perspectives* 12 (3): 183–188.

Cangelosi, Angelo, and Francesca Stramandinoli. 2018. "A Review of Abstract Concept Learning in Embodied Agents and Robots." *Philosophical Transactions of the Royal Society B: Biological Sciences* 373 (1752): 20170131.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Csapo, Adam, Emer Gilmartin, Jonathan Grizou, Jingguang Han, Raveesh Meena, Dimitra Anastasiou, Kristiina Jokinen, and Graham Wilcock. 2012. "Multimodal Conversational Interaction with a Humanoid Robot." In *2012 IEEE 3rd International Conference on Cognitive Infocommunications*, 667–672. New York: IEEE.

De La Cruz, Vivian Milagros, Alessandro Di Nuovo, Santo Di Nuovo, and Angelo Cangelosi. 2014. "Making Fingers and Words Count in a Cognitive Robot." *Frontiers in Behavioral Neuroscience* 8:13.

Di Nuovo, Alessandro, F. Broz, N. Wang, T. Belpaeme, A. Cangelosi, R. Jones, R. Esposito, F. Cavallo, and P. Dario. 2018. "The Multi-modal Interface of Robot-Era Multi-robot Services Tailored for the Elderly." *Intelligent Service Robotics* 11 (1): 109–126.

Di Nuovo, Alessandro, and Jay L. McClelland. 2019. "Developing the Knowledge of Number Digits in a Child-Like Robot." *Nature Machine Intelligence* 1 (12): 594–605.

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

Harnad, Stevan. 1990. "The Symbol Grounding Problem." *Physica D* 42:335–346.

Heinrich, Stefan, Sven Magg, and Stefan Wermter. 2015. "Analysing the Multiple Timescale Recurrent Neural Network for Embodied Language Understanding." In *Artificial Neural Networks*, edited by P. Koprinkova-Hristova, V. Mladenov, and N. K. Kasabov, 149–174. Cham, Switzerland: Springer.

Hoff, Erika. 2013. *Language Development*. Boston: Cengage Learning.

Huang, Jing, and Brian Kingsbury. 2013. "Audio-Visual Deep Learning for Noise Robust Speech Recognition." In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 7596–7599. New York: IEEE.

Ishiguro, Hiroshi. 2016. "Android Science." In *Cognitive Neuroscience Robotics A*, edited by M. Kasaki, H. Ishiguro, M. Asada, M. Osaka, and T. Fujikado, 193–234. Tokyo: Springer.

Ivaldi, Serena, Natalia Lyubova, Alain Droniou, Vincent Padois, David Filliat, Pierre-Yves Oudeyer, and Olivier Sigaud. 2013. "Object Learning through Active Exploration." *IEEE Transactions on Autonomous Mental Development* 6 (1): 56–72.

Kaneko, Kenji, Kensuke Harada, Fumio Kanehiro, Go Miyamori, and Kazuhiko Akachi. 2008. "Humanoid Robot HRP-3." In *Proceedings of the 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2471–2478. New York: IEEE.

Kuriyama, Takatsugu, Takashi Shibuya, Tatsuya Harada, and Yasuo Kuniyoshi. 2010. "Learning Interaction Rules through Compression of Sensori-motor Causality Space." In *Proceedings of The Tenth International Conference on Epigenetic Robotics (Epirob10)*, 57–64. Lund University Cognitive Studies, 149.

Lallee, Stephane, and Peter Ford Dominey. 2013. "Multi-modal Convergence Maps: From Body Schema and Self-Representation to Mental Imagery." *Adaptive Behavior* 21 (4): 274–285.

Le, Quoc V. 2013. "Building High-Level Features Using Large Scale Unsupervised Learning." In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 8595–8598. New York: IEEE.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521 (7553): 436–444.

Liu, Rui, and Xiaoli Zhang. 2019. "A Review of Methodologies for Natural-Language-Facilitated Human–Robot Cooperation." *International Journal of Advanced Robotic Systems* 16 (3): 1729881419851402.

MacWhinney, B. 1998. “Models of the Emergence of Language.” *Annual Review of Psychology* 49:199–227.

Mavridis, Nikolaos. 2015. “A Review of Verbal and Non-verbal Human–Robot Interactive Communication.” *Robotics and Autonomous Systems* 63:22–35.

Moeller, Korbinian, Laura Martignon, Silvia Wessolowski, Joachim Engel, and Hans-Christoph Nuerk. 2011. “Effects of Finger Counting on Numerical Development–the Opposing Views of Neurocognition and Mathematics Education.” *Frontiers in Psychology* 2:328.

Morse, Anthony F., Viridian L. Benitez, Tony Belpaeme, Angelo Cangelosi, and Linda B. Smith. 2015. “Posture Affects How Robots and Infants Map Words to Objects.” *PLoS One* 10 (3): e0116012.

Morse, Anthony F., and Angelo Cangelosi. 2017. “Why Are There Developmental Stages in Language Learning? A Developmental Robotics Model of Language Development.” *Cognitive Science* 41:32–51.

Morse, Anthony F., Joachim de Greeff, Tony Belpeame, and Angelo Cangelosi. 2010. “Epigenetic Robotics Architecture (ERA).” *IEEE Transactions on Autonomous Mental Development* 2 (4): 325–339.

Murphy, Robin R. 2019. *Introduction to AI Robotics*. Cambridge, MA: MIT Press.

Mutlu, Bilge, Takayuki Kanda, Jodi Forlizzi, Jessica Hodgins, and Hiroshi Ishiguro. 2012. “Conversational Gaze Mechanisms for Humanlike Robots.” *ACM Transactions on Interactive Intelligent Systems* 1 (2): 1–33.

Nakamura, Tomoaki, Yoshiki Ando, Takayuki Nagai, and Masahide Kaneko. 2015. “Concept Formation by Robots Using an Infinite Mixture of Models.” In *Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 4593–4599. New York: IEEE.

Nakamura, Tomoaki, Takayuki Nagai, and Naoto Iwahashi. 2009. “Grounding of Word Meanings in Multimodal Concepts Using LDA.” In *Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3943–3948. New York: IEEE.

Ngiam, Jiquan, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. “Multimodal Deep Learning.” In *ICML ’11: Proceedings of the 28th International Conference on Machine Learning*. Madison, WI: Omnipress.

Noda, Kuniaki, Hiroaki Arie, Yuki Suga, and Tetsuya Ogata. 2014. “Multimodal Integration Learning of Robot Behavior Using Deep Neural Networks.” *Robotics and Autonomous Systems* 62 (6): 721–736.

Noda, Kuniaki, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G. Okuno, and Tetsuya Ogata. 2015. “Audio-Visual Speech Recognition Using Deep Learning.” *Applied Intelligence* 42 (4): 722–737.

Ogino, Masaki, Hideki Toichi, Yuichiro Yoshikawa, and Minoru Asada. 2006. “Interaction Rule Learning with a Human Partner Based on an Imitation Faculty with a Simple Visuo-motor Mapping.” *Robotics and Autonomous Systems* 54 (5): 414–418.

Pastra, Katerina. 2008. “PRAXICON: The Development of a Grounding Resource.” *In Proceedings of the International Workshop on Human-Computer Conversation*, Bellagio, Italy.

Pastra, Katerina, and Yiannis Aloimonos. 2012. “The Minimalist Grammar of Action.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 367 (1585): 103–117.

Pecyna, Leszek, Angelo Cangelosi, and Alessandro Di Nuovo. 2020. “A Robot That Counts Like a Child: A Developmental Model of Counting and Pointing.” *Psychological Research*. https://doi.org/10.1007/s00426-020-01428-8.

Pezzulo, Giovanni, Lawrence W. Barsalou, Angelo Cangelosi, Martin H. Fischer, Ken Mcrae, and Michael Spivey. 2013. “Computational Grounded Cognition: A New Alliance between Grounded Cognition and Computational Modeling.” *Frontiers in Psychology* 3:612.

Ruciński, Marek, Angelo Cangelosi, and Tony Belpaeme. 2012. “Robotic Model of the Contribution of Gesture to Learning to Count.” In *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics*, 1–6. New York: IEEE.

Sakagami, Yoshiaki, Ryujin Watanabe, Chiaki Aoyama, Shinichi Matsunaga, Nobuo Higaki, and Kikuo Fujimura. 2002. “The Intelligent ASIMO: System Overview and Integration.” In Vol. 3, *Proceedings of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2478–2483. New York: IEEE.

Samuelson, Larissa K., Linda B. Smith, Lynn K. Perry, and John P. Spencer. 2011. “Grounding Word Learning in Space.” *PLoS One* 6 (12): e28095.

Shiomi, Masahiro, Daisuke Sakamoto, Takayuki Kanda, Carlos Toshinori Ishi, Hiroshi Ishiguro, and Norihiro Hagita. 2008. “A Semi-autonomous Communication Robot—a Field Trial at a Train Station.” In *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction*, 303–310. New York: IEEE.

Sinapov, Jivko, Connor Schenck, Kerrick Staley, Vladimir Sukhoy, and Alexander Stoytchev. 2014. “Grounding Semantic Categories in Behavioral Interactions: Experiments with 100 Objects.” *Robotics and Autonomous Systems* 62 (5): 632–645.

Sinapov, Jivko, and Alexander Stoytchev. 2011. "Object Category Recognition by a Humanoid Robot Using Behavior-Grounded Relational Learning." In *2011 IEEE International Conference on Robotics and Automation*, 184–190. New York: IEEE.

Stramandinoli, Francesca, Davide Marocco, and Angelo Cangelosi. 2017. "Making Sense of Words: A Robotic Model for Language Abstraction." *Autonomous Robots* 41 (2): 367–383.

Sugita, Yuuya, and Jun Tani. 2005. "Learning Semantic Combinatoriality from the Interaction between Linguistic and Behavioral Processes." *Adaptive Behavior* 13 (1): 33–52.

Tikhanoff, V., A. Cangelosi, and G. Metta. 2011. "Language Understanding in Humanoid Robots: iCub Simulation Experiments." *IEEE Transactions on Autonomous Mental Development* 3 (1): 17–29.

Tomasello, Michael. 1992. *First Verbs: A Case Study of Early Grammatical Development*. Cambridge: Cambridge University Press.

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

Tomasello, Michael, and Patricia J. Brooks. 1999. "Early Syntactic Development: a Construction Grammar Approach." In *Development of Language*, edited by M. Barrett, 161–190. London: Psychology Press.

Tomasello, Michael, Malinda Carpenter, and Ulf Liszkowski. 2007. "A New Look at Infant Pointing." *Child Development* 78 (3): 705–722.

Tuci, Elio, Tomassino Ferrauto, Arne Zeschel, Gianluca Massera, and Stefano Nolfi. 2011. "An Experiment on Behavior Generalization and the Emergence of Linguistic Compositionality in Evolving Robots." *IEEE Transactions on Autonomous Mental Development* 3 (2): 176–189.

Twomey, Katherine E., Anthony F. Morse, Angelo Cangelosi, and Jessica S. Horst. 2016. "Children's Referent Selection and Word Learning: Insights from a Developmental Robotic System." *Interaction Studies* 17 (1): 93–119.

Wallace, Richard S. 2009. "The Anatomy of ALICE." In *Parsing the Turing Test*, edited by R. Epstein, G. Roberts, and G. Beber, 181–210. Dordrecht: Springer.

Weizenbaum, Joseph. 1966. "ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine." *Communications of the ACM* 9 (1): 36–45.

Yamada, Tatsuro, Shingo Murata, Hiroaki Arie, and Tetsuya Ogata. 2017. "Representation Learning of Logic Words by an RNN: From Word Sequences to Robot Actions." *Frontiers in Neurorobotics* 11:70.

Yamashita, Yuichi, and Jun Tani. 2008. "Emergence of Functional Hierarchy in a Multiple Timescale Neural Network Model: A Humanoid Robot Experiment." *PLoS Computational Biology* 4 (11): e1000220.

Yang, Yezhou, Yi Li, Cornelia Fermuller, and Yiannis Aloimonos. 2015. "Robot Learning Manipulation Action Plans by 'Watching' Unconstrained Videos from the World Wide Web." In *Twenty-Ninth AAAI Conference on Artificial Intelligence*. Menlo Park, CA: AAAI Press.

Zhong, Junpei, Martin Peniak, Jun Tani, Tetsuya Ogata, and Angelo Cangelosi. 2019. "Sensorimotor Input as a Language Generalisation Tool: A Neurorobotics Model for Generation and Generalisation of Noun-Verb Combinations with Sensorimotor Inputs." *Autonomous Robots* 43 (5): 1271–1290.