Linguistic Primitives: A New Model for Language Development in Robotics

Alessio Mauro Franchi(✉), Lorenzo Sernicola, and Giuseppina Gini

DEIB Department, Politecnico di Milano, Milano, Italy
{alessiomauro.franchi,giuseppina.gini}@polimi.it

Abstract. Often in robotics natural language processing is used simply to improve the human-machine interaction. However, language is not only a powerful communication tool: it is deeply linked to the inner organization of the mind, and it guides its development. The aim of this paper is to take a first step towards a model of language which can be integrated with the diverse abilities of the robot, thus leading to its cognitive development, and eventually speeding up its learning capacity. To this end we propose and implement the Language Primitives Model (LPM) to imitate babbling, a phase in the learning process that characterizes a few months old babies. LPM is based on the same principles dictated by the Motor Primitives model. The obtained results positively compare with experimental data and observations about children, so confirming this interest of the new model.

Keywords: Emergence of vocalization · Babbling · Motor primitives

1 Introduction

Recently Natural Language Processing (NLP) has developed many voice recognition technologies, such as Apple Siri [17], mostly used for simple tasks like sending a message. Limitations emerge also in other applications; video games and robots are often able to recognize a few words, mainly related to a specific task. Instead language plays an important role in intelligent behaviours, as initially indicated by Alan Turing in his "Imitation Game" [19]. His final considerations was that language manipulation is a necessary condition for a machine to be intelligent.

In biology the ability of communicating through sounds is present in several animal species. However language has evolved differently in humans, mainly due to the fact that it is more than an external instrument for communicating; it is intrinsic to the mind itself [3]. Spokeng language is thus the epiphenomenon of the deep link existing between brain and language. It is known that cerebral areas dedicated to language are highly connected with motor ones; when one elaborates a sentence or produces a word both areas are activated [14]. This is a clue of the presence of common mental mechanisms both for motor or linguistic skills.

E. Tuci et al. (Eds.): SAB 2016, LNAI 9825, pp. 207–218, 2016.
DOI: 10.1007/978-3-319-43488-9_19

To recreate in artificial agents such an ability, natural language and perceptions should be linked together; the comprehension of natural language by robots should be based on sensory-motor experiences and not on a sort of hard coded semantics [2]. In a longer time perspective, a language based on the experience would help robots to autonomously extract knowledge about the environment, integrating also this information with those from its own actions and related sensorial feedbacks [12].

Our research mainly focuses on this relationship between motor learning and mental abilities development in humans as a way to improve the robot learning system.

We draw inspiration both from humans and from decades of studies in NLP, that despite impressive results [17] has still many problems to solve. We start from the hypothesis that the linguistic apparatus in robotics should be part of several other biologically inspired mechanisms, cooperating together towards the cognitive development of the artificial agent. We take inspirations from newborns, focusing in particular on the evolutive steps of language. Starting from data collected during the Speechome Project at M.I.T. [13], we have designed and implemented a model called the Linguistic Primitive Model (LPM). It aims at imitating babies in a specific moment of language exploration, the babbling phase, that takes place from the sixth to the tenth month and is the way they imitate sounds and words on purpose.

This new model re-uses several concepts typically associated in robots with movements, creating a parallel between motor and linguistic mechanisms that is known to exist in humans brain. Learning starts from a simple hard coded dataset of linguistic primitives; the agent tries to imitate an heard word continuolsy composing the primitives, and producing new sounds until it succeeds. Newly learned words are added to the set of primitives ready to be used or composed again to form more complex sounds. Other primitives become useless and are discarded.

In the rest of the paper we shortly review the related works and introduce our model of Linguistic Primitives. We make experiments using some data from the mentioned Speechome data. Results of our experiments are not easily comparable with state of the art, but they demonstrate that our hypotheses about language development are correct and that LPM is a basis for further researches. They highlight also that the use of a typical model for movements is a new promising point of view for the development of linguistic skills in robotics.

2 Related Works

As we have briefly seen language is strictly connected with mind; its learning helps the cognitive development, and viceversa [3]. From studies about babies it is clear that cognitive development in humans is a process parallel to language learning. Words are used by infants as powerful instrument for building an internal representation of the external world; they act as labels for objects in the environment [11]. New grammatical constructs interact with sensory-motor

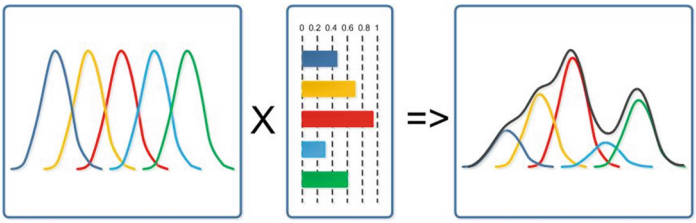

Fig. 1. A schema of the motor primitive composition mechanism

apparatus at a neuronal level [21]: when somebody listens to verbs like "walk" or "see" its brain activates also neurons of the cortical motor areas. Motor and sensors areas are thus linked togheter: language learning depends strictly on physical and sensorial experiences, and viceversa. This concept is known as embodiment: intelligence needs a body and an environment to develop [16].

A biologically inspired approach for language development should help robots in autonomously extracting simple semantic information from the context; it is the case of [10] where a moving robot is able to correctly interpret navigation commands expressed in natural language. Another challenge is the symbol grounding problem, that concerns the relation between words or symbols with their meaning [18]. An interesting study showed that sensory-motor integration can improve symbol grounding processes, just like humans do [9].

The main idea is thus that robots, just like living beings, must play an active role in their cognitive development and learning, interacting through their body with the environment. Also language emergence should follow this paradigm and should be grounded on sensorial experiences. Several evidences show that motor and linguistic learning share the same mechanisms. Nowadays a validated theory for movement learning is the motor primitive mechanism [7]; motor primitives are the "smallest" entity of voluntary movement, that activate a single muscle. Composition and coordination of several motor primitives, one for each muscle involved, result in a final complex movement Fig. 1. This theory seems to clearly explain how infants go from instinctive to voluntary movements, and may also hold for language development: babbling is for babies the mechanism to start from simple innate sounds and get to complex and intentional words by their composition [20].

3 Our Approach: Language Primitives

Understanding and producing language is a multisensory process; it is grounded on the visual, musculoskeletal and proprioceptive systems; we use our ears to listen to spoken words but several studies demonstrated that we also exploit sight for facial expression analysis or body movements recognition [1]. In the same way the production of language involves the muscular and proprioceptive systems; these should be seen as two significant hints of the relation between linguistic and motor skills.

In their first months of life, babies are not able to pronounce and to distinguish words, but can determine their phonetic differences. This mechanism leads to the so called "phonetic attunement", that is a greater sensitivity to the contrasts and to the specific tones of a particular language, that eventually leads babies to distinguish the first words. In the same time they start also to distinguish the repeated language patterns, mainly by a statistical approach. It is not necessary for a child to segment a sentence, but he is able to create early phonetic categories simply listening to sounds. This statistical approach is part of the distributional learning mechanism and it represents the origin of language learning: the information concerning the distribution of the frequencies of tones is merged with the visual information, contributing to the creation of a speech context.

This first stage paths the way for the following ones in language learning: the recognition of vowels (6–8 month), of consonants (8–12 month) and finally of phonemes duration. These phases in children follow very rapidly one another, more quickly than the only auditory inputs would allow; this gap may be explained by cross-modal association [4]. Language development is a very complex phenomenon, but also a universal process and it is the same across different environmental condition and experiences [6].

We propose a model focused mainly on the basic mechanism through which words are formed in the first months of life; the term "babbling" refers to the sounds uttered by newborns when they still aren't able to pronounce complete words. Researchers agree that this phenomenon plays a key role in the correct cognitive development of the baby. Actually, the first movements of the limbs and of the mouth of newborns are the product of involuntary reflexes. During the first two months of life, the baby utters sounds that are called protophones, which already have some features of vowels; these develop until, around the sixth month, babbling starts.

With time the protophones become no longer involuntary sounds and are intentionally produced. This voluntary act is part of a more global cognitive development of the baby, which maps the movements of the vocal tract and the resultant sounds, allowing babies to replicate a sound. This mapping leads them to voluntarily utter a word [8]. Other researches have highlighted how these basic mechanisms for language learning are in common with those for movement learning and both modules communicate to strengthen each other [13].

The Speechome Project is our main inspiration. Among the huge amount of collected data, several audio files recorded a baby repeating the same word in different instants and house places, starting from the very first trials of imitation to its voluntary pronunciation. As an example we report a brief transcription of the word "water":

"gaga" - "gata" - "wata" - "wate" - "water"

From the analysis of these registrations has emerged that each single consonant-vocal couple may be considered as the most similar particle to a linguistic primitive we can extract. In nature a baby tries to imitate the sound he is listening

to composing all the "linguistic primitives" he has, in a similar way to motor learning mechanism. In the above example the baby starts repeating the innate "ga" particle; as his vocal tract and facial muscles modify in time he learns more complex sounds such as "ta" or "wa", replacing simpler particles and resulting in a more accurate reproduction of the target word. The baby finally learns the "r" and succeeds in pronouncing "water".

Our linguistic primitives model aims at reproducing this process of imitation of a spoken word starting from a hard coded dataset of linguistic primitives; this set has a direct equivalent in humans as the internal mapping between a sound and a specific movement of the mouth and of the diaphragm. We consider linguistic primitives as innate, as they are a direct consequence of non-voluntary changes, and are independent from the language family and context.

To generate the primitives dataset we have analyzed various registrations reproducing sounds made by babies during their babbling stages. We have first discarded videos not tagged with the age of the baby and then classified those selected into the five different stages of language learning [8]:

1. Cooing (1st–4th month), repetition of single sound, e.g.: oooooo, aaaaaaah;
2. Consonant-Vowel (CV) or Vowel-Consonant (VC) sounds combinations (4th–6th month), e.g.: maaaa, uuuum, baaaa ;
3. Reduplicated babbling (6th–10th month), e.g.: babababa, gagagaga, dadadada;
4. Non-reduplicated babbling (6th–10th month), e.g.: bama, gagamee
5. Quasiwords (10th–12th month), e.g.: watee.

Stages 4 and 5 see the first attempts to compose these primitives intentionally. Since during these stages the dataset of primitives is quite limited, babies are not able to compose real words but only simple terms such as "mama" or "dad", made of two or three primitives concatenated. The continuous enrichment of this internal dataset eventually leads to the production of complex sounds.

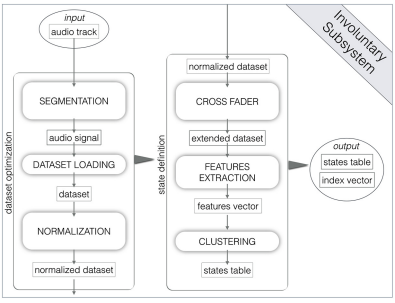

Fig. 2. The involuntary subsystem.

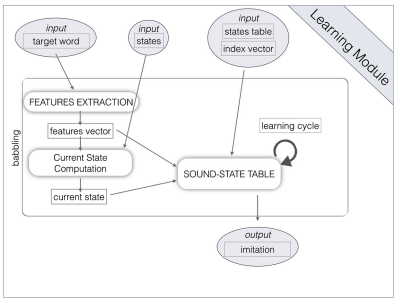

Fig. 3. The voluntary subsystem.

4 The Implemented System

An artificial architecture able to simulate the language development skills as seen above should comprehend both the innate mental abilities and all the mechanisms of voluntary learning that rely on multi-modal sensorial inputs. Such an architecture is a long term goal; in our model we will consider only those aspects related to canonical babbling and to the auditory stimulus processing.

Two subsystems compose our model: the involuntary (ISS) Fig. 2 and voluntary (VSS) Fig. 3. The first reproduces all the aspects related to the physiologic growth of the body, starting from the earlier months of life until first voluntary mechanisms of imitation starts. It receives as input several examples of babbling and extracts the linguistic primitives; all the possible combinations of two primitives are then generated, a mechanism that corresponds to the innate development of the proto-word. A Sound-State table is generated, which recreates the natural mapping between "words" and "states" each baby learns during time.

The second subsystem deals with babbling, a phase of learning that emerges in parallel to the acquirement of new words and to the first voluntary imitations of heard sounds. The VSS receives as input a sound representing the word to be learned and the architecture starts to "babble", i.e. it produces sounds by composing linguistic primitives. The first type of composition we implemented consists in the concatenation of two or more linguistic primitives; more advanced mechanism may be added in the future. These sounds are compared with the target input and the learning process is stopped when the similarity is greater than a fixed threshold; this event triggers the activation of the Sound-State Table, mimicking the neural activation it is known to appear when a baby accomplishes his imitation task [15].

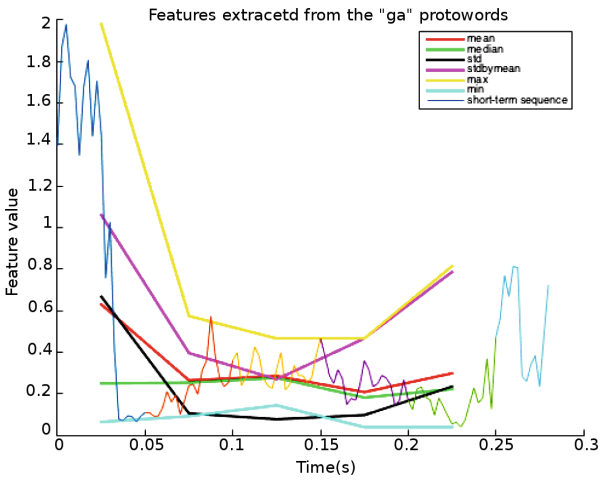

Fig. 4. The six features extracted in the short-term processing from the sound "ga"

4.1 Involuntary Subsystem

The main task accomplished by the ISS is the generation of the "hard-coded" dataset of linguistic primitives. The ISS receives as input several segmented audio signals, each containing exactly one "babble", i.e. a vocal-consonant couple. Every signal is processed with the following filters: stereo-to-mono converter, sampling frequency normalizator, pitch shifter and an RMS equalizator. This process equalizes signals coming from different sources. A second module is responsible for the concatenation of these primitives by a cross-fading technique. This process can be considered part of the involuntary subsystem as in babies the generation of first linguistic primitives appears as a direct consequence of their physical development more than of an a-posteriori learning.

The last step is the selection of the "States" for the table. A state is a compact representation of the environment; it is actually composed by the sound the robot has listened to. For state selection we perform two kinds of elaboration on the signals, mid-term windowing and short-term processing: both make use of a framing mechanism for trimming primitives into very short segments which are then analyzed independently from each other. Several features are extracted; their mean and variance form the vector of features we use in classification for the state selection and for similarity computation Fig. 4.

4.2 Voluntary Subsystem

The second module deals with the babbling and learning phase; it receives in input a target word, the Sound-State Table and the dataset of primitives. The features vector of the input word is computed and projected onto the collection of states and the most similar one is selected.

The States-Sounds Table is used as a sort of neural network to register each tentative of imitation; rows are dedicated to states, columns correspond to produced sounds. For each tentative the system will do, the similarity values between the target and the produced sound is computed and stored in the corresponding entry of the table; this process is repeated until the imitation performance is satisfying. As the number of trials grows, the mean number of tentatives needed by the agent significantly decreases, indicating that the system is learning new words.

5 Experimental Results

For validating the proposed model we defined different experiments; stated the innovative approach here presented a direct comparison with other works is very difficult. Our goal is not to improve others' models, but to propose a new point of view for language learning that is bio-inspired, grounded on the agent's experience, and that shares its mechanisms with those of the motor system.

Experiments are intended to evaluate the ability of the system both to correctly imitate an input sound and to learn them as the number of tentatives grows. Two

metrics are evaluated. The first is similarity, that describes how much the produced word is similar to the listened one and is computed as the distance between the features vectors of both sounds, normalized in the range [0..1]:

$$similarity = \tanh\left(\frac{1}{vectorsDistance}\right) \tag{1}$$

where vectorsDistance is the squared norm of the difference between the two vectors; similarity is thus the probability that two sounds represent the same word.

The second is the number of cycles the system needs to produce a "good" output, where good means above a predefined similarity threshold we empirically evaluated. For all the following experiment we split our dataset of sounds into training and testing subset, composed of 600 and 300 signals respectively.

5.1 Preliminary Step

In a preliminary step we had to optimize the open parameters of the system; the most fundamental is the similarity threshold, that is the minimum value of similarity we require to consider an imitation as valid.

The setup for this preliminary experiment is:

- training dataset: 600 words;
- testing dataset: 300 words;
- number of tests: 20;

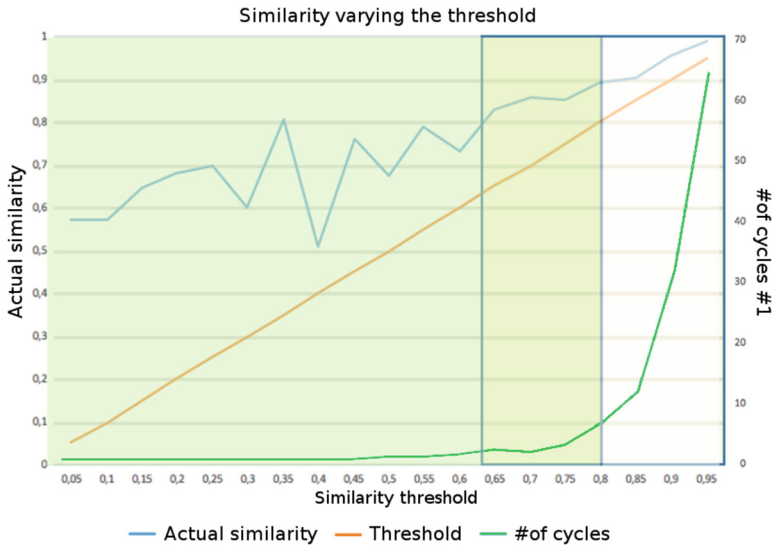

Fig. 5. The graph shows the combined result and the optimal values for the similarity threshold is highlighted

– threshold value step: 0.05 (from 0.05 to 0.95);
– input words per test: 200 words (selected randomly from the 300 words).

For each threshold value in the range, the mean number of cycles necessary to correctly imitate the input word is logged; as Fig. 5 shows this value is low (<15) for threshold lower than 0.85, a value corresponding to a very accurate imitation of the input word. Moreover the mean value of the actual similarity is always greater that the defined treshold. The combination of these two considerations defines a range of optimal similarity threshold in [0.65, 0.8], a good trade-off that guarantees a similarity above 0.8 and a mean number of cycles lower than 10.

5.2 First Experiment

We firstly evaluated the importance of the number of words in input to the system; by this parameter we can copy the natural tendency of caregivers to use a simplified lexicon, with non-conjugate verbs and a restricted vocabulary.

We thus reduced the number of words in the testing dataset down to only 20, keeping other parameters unaltered:

– training dataset: 600 words;
– testing dataset: 20 words;
– number of tests: 19;
– threshold value step: 0.05 (from 0.05 to 0.95);
– input words per test: 200 words (selected randomly from the 20 words).

By comparing experimental data with previous results it emerges that our system is able to correctly imitate and learn words in a lower number of cycles, especially in cases of high similarity threshold values Fig. 6. Moreover the quality of learning is good even if the number of input word decreases Fig. 7.

This behavior is biologically validated by results from the Speechome Project: in nature the learning of a new word happens as caregivers repeat it more frequently and homogeneusly.

5.3 Second Experiment

The second experiment is mostly focused on the ability of the system to learn new words. We have analyzed the trend of the learning rate as input words follow each other and we expect it to decrease in time.

The parameters for this second experiment are:

– training dataset: 600 words;
– testing dataset: 300 words;
– number of test: 1;
– threshold values: 0.8;
– input words per test: 200 words (selected randomly from the 300 words).

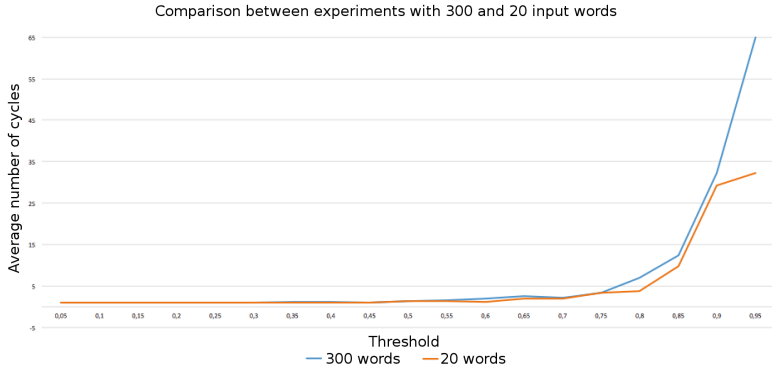

Fig. 6. The number of cycles needed to reproduce a word decrease if a reduced input set of word is used

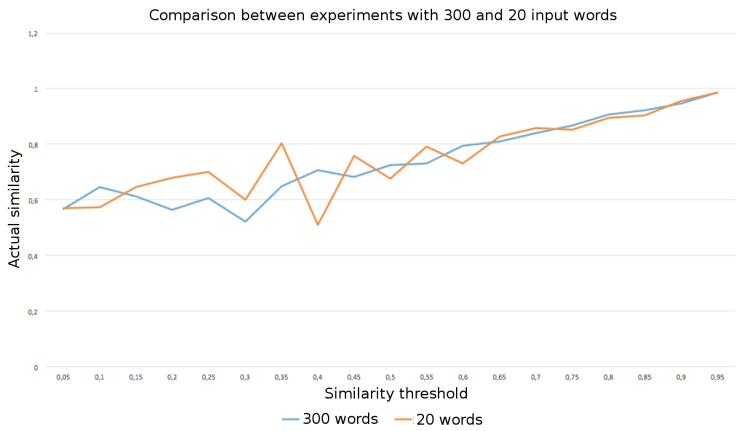

Fig. 7. The trend of the similarity values is not affected by the number of input words

From this experiment we have extracted the number of cycles necessary to imitate each single word sent as input to the system, computing then the moving average to remove noise in data. The decreasing mean number of cycles and the frequency of input words requiring a single tentative to be correctly imitated show an ongoing learning of novel words or proto-words Fig. 8. This result is supported by scientific evidence showing that babies speed up language learning by memorizing the correct imitation tentatives they make and reusing words already learned.

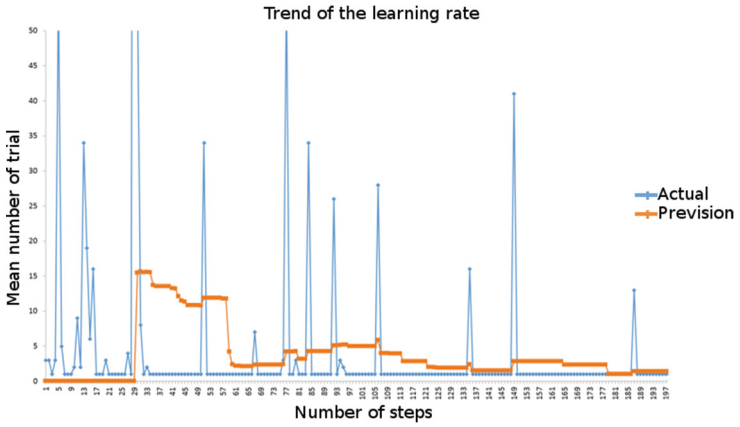

Fig. 8. The moving average of the mean number of tentatives the system need to imitate the target word

6 Conclusion

The use of natural language in robotics has always been an independent field of research that was born with the aim to obtain intelligent and immediate interaction between men and machines. However, the importance of language does not exclusively lie in the field of communication: it actually represents the very image of the mind, it is deeply linked to its inner structure, and it guides its development through innumerable phases.

The aim of this work is to take a first step towards a model of language that can be integrated with the other cognitive abilities of the robot, with the purpose of contributing and collaborating towards a faster and more reliable development of its mind and of its learning ability.

We focused our attention on the initial stages of language development, which takes place in babies during their first years of life during which they switch from an involuntary production of sounds to the voluntary use of vowels and syllables: the babbling phase. We consequently elaborated the Model of Language Primitives (LPM), which is based on the same principles lying under the motor primitives, transposed into the language learning process.

In order to test the LPM we performed some experiments with the aim to evaluate its imitation ability and to test whether the system is effectively able to learn. The obtained results not only validate this model, but also show a behavior very similar to the one observed in babies. This supports the idea of the strict parallelism between language and motor primitives, the core of the proposed model.

This preliminary results are encouraging but several open problems still exist. The next step we want to explore is the integration of this model into our intentional architecture IDRA [5], to exploit its potentiality in processing different types of sensorial input, in learning associations, and in the autonomous generation of new objectives starting from innate instincts. The integration of the LPM in IDRA should strengthen their learning abilities.

References

1. Calvert, G., Spence, C., Stein, B.E.: The Handbook of Multisensory Processes. MIT Press, Cambridge (2004)
2. Cangelosi, A.: Grounding language in action and perception: from cognitive agents to humanoid robots. Phy. Life Rev. **7**, 139–151 (2010)
3. Dominey, P.F.: How are grammatical constructions linked to embodied meaning representations? AMD Newslett. **10**, 3 (2013)
4. Erneling, C.E.: Understanding Language Acquisition: The Framework of Learning. Cambridge University Press, Cambridge (1993)
5. Franchi, A.M., Mutti, F., Gini, G.: From learning to new goal generation in a bioinspired robotic setup. In: Advanced Robotics (in press, 2016). doi:10.1080/01691864.2016.1172732
6. Gleitman, L.R., Newport, E.L.: The invention of language by children: environmental and biological influences on the acquisition of language. In: Gleitman, L.R., Liberman, M. (eds.) An Invitation to Cognitive Science, 2nd edn, pp. 90–116. MIT Press, Cambridge (2005)
7. Konczak, J.: On the notion of motor primitives in humans and robots. In: Proceedings of the Fifth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems, pp. 47–53 (2005)
8. Kuhl, P.K., Meltzoff, A.N.: Infant vocalizations in response to speech: vocal imitation and developmental change. J. Acoust. Soc. Am. **100**, 2425–2438 (1996)
9. MacDorman, K.F.: Grounding symbols through sensorimotor integration. J. Robot. Soc. Jpn. **17**, 20–24 (1999)
10. Matuszek, C., Herbst, E., Zettlemoyer, L., Fox, D.: Learning to parse natural language commands to a robot control system. In: Desai, J.P., Dudek, G., Khatib, O., Kumar, V. (eds.) Experimental Robotics. Springer Tracts in Advanced Robotics, vol. 88, pp. 403–415. Springer, New York (2013)
11. Pastra, K.: Autonomous acquisition of sensorimotor experiences: any role for language? AMD Newslett. **25**, 12–13 (2013)
12. Popescu, A.M., Etzioni, O., Henry, K.: Towards a theory of natural language interfaces to databases, vol. 1, pp. 149–157 (2013)
13. Roy, D., Patel, R., DeCamp, P., Kubat, R., Fleischman, M., Roy, B., Mavridis, N., Tellex, S., Salata, A., Guiness, J., Levit, M., Gorniak, P.: The human speechome project, vol. 1, pp. 192–196 (2006)
14. Saffran, J.R., Pollak, S.D., Seibel, R.L., Shkolnik, A.: Dog is a dog is a dog: infant rule learning is not specific to language. Cognition **105**, 669–680 (2007)
15. Scott, M., Yeung, H.H., Gick, B., Werker, J.F.: Inner speech captures the perception of external speech. J. Acoust. Soc. Am. **133**, 2425–2438 (2013)
16. Spenko, M.J., Haynes, G.C., Saunders, J.A., Cutkosky, M.R., Rizzi, A.A., Full, R.J., Koditschek, D.E.: Biologically inspired climbing with a hexapedal robot. J. Field Robot. **25**, 223–242 (2008)
17. Stern, J.: Apple's siri: loved, but underused. The ABC News, pp. 83–92 (2012)
18. Taddeo, M., Floridi, L.: Solving the symbol grounding problem: a critical review of fifteen years of research. J. Exper. Theor. Artif. Intell. **17**, 419–445 (2005)
19. Turing, A.M.: Computing machinery and intelligence. Mind **59**, 433–460 (1950)
20. Vihman, M.M.: Phonological Development: The Origins of Language in the Child. Wiley, Chichester (1996)
21. Weng, J.: These questions arose because you used symbolic representations. AMD Newslett. **25**, 11 (2013)