# Learning through Multimodal Interaction

Chen Yu, Hui Zhang and Linda B. Smith
*Indiana University*
*Bloomington, IN, 47405, USA*

***Abstract*— This paper proposes and implements a new experimental paradigm to study the role of multimodal interaction in automatic language learning. We observe that child language acquisition relies significantly on everyday social interactions with adult partners. In light of this, we argue that an important step to build machines that can also learn from social interactions with human users is to understand the nature of learning-oriented interactions. To do so, a central problem is to find a way to decouple the social interaction between two agents (e.g. a human supervisor and a machine learner), so that we can systematically manipulate and control the dynamic flow of the interaction to create and examine various interactive learning conditions. In this paper, we build a set of virtual humans as language learners possessing different social-cognitive skills, and ask real people to teach them object names. Using multisensory recording devices, we measured how well real people interact with virtual humans and how they shape their behaviors to adapt to different social-cognitive skills that virtual humans possess. Multimodal data were analyzed to shed light on both perceptual and behavioral aspects of human users in interaction. These results can be used both to guide building artificially intelligent system and to provide useful insights on human-human communication and child language learning.**

***Index Terms*— Word Learning, Social Interaction, Embodiment, Virtual Reality**

## I. INTRODUCTION

Language is a central aspect of human intelligence and essential for human-human everyday communication. A basic function of language is to provide linguistic labels that refer to objects, actions, and sensorimotor activities in the real world so that people can talk about them and share experiences through verbal communication. Nonetheless, most AI systems based on current speech recognition and generation technologies treat language as a symbolic system without considering the mappings between language and the physical world. Without any semantic knowledge grounded at the sensorimotor level, those systems may not be able to really understand and use language in a human-like way. For instance, a speech recognition and synthesis system (e.g. IBM ViaVoice) can map the sound "dog" to its text form and translate the text back to speech, but it has no knowledge about what this word refers to – the semantic meaning of the word. Therefore, when an AI system based on this kind of speech technology "sees" a dog, even if it can recognize the object kind from its visual system, it can not name it without the knowledge of the mapping between the visual object and its spoken name. More generally, this problem is termed symbol grounding by Harnad (1990), which applies not only to speech and language systems but also to many other AI systems based on mathematical and logic inferences on symbolic representations. To build human-like artificially intelligent systems, symbolic representations need to be grounded in the real world. Moreover, instead of pre-programming semantic knowledge which is practically almost impossible, intelligent systems (e.g. robots) are expected to be encoded with general learning mechanisms and can learn to connect language to the world through everyday interactions with a human supervisor.

Humans are not born with a lexicon of grounded semantics. Nonetheless, young children learn how to name things in their native language smoothly and effortlessly. How could they achieve this goal easily? Could we build a computational system that accomplishes the same learning task? If so, what attributes of a young child are crucial for the machine to emulate? We believe that studies in human language acquisition provide useful hints in various aspects to answer those questions and to guide us to build language-grounded machines (Brooks, Breazeal, Irie, Kemp, & Marjanovi, 1998; Asada, MacDorman, Ishiguro, & Kuniyoshi, 2001; Breazeal & Scassellati, 2002; Yu, Ballard, & Aslin, 2005). First, human studies suggest what kinds of technical problems needed to tackle. For the same task – mapping language to the real world, the machine needs to deal with the similar problems that the existing intelligent systems (humans, etc.) face with. More specifically, one key problem in human language learning is termed reference uncertainty (Quine, 1960): given a natural learning situation consisting of a sequence of spoken words uttered by a teacher and meanwhile multiple co-occurring objects and events in the extralinguilistic context, a word learner should discover the relevant correspondence from those co-occurring word-object and word-event pairs. The present work focuses on solving this word-to-world mapping problem in language learning.

Moreover, studies in human language acquisition document and analyze what kinds of learning mechanisms young children apply, which turn out to be quite different with current machine learning (ML) approaches. Many ML approaches first collect data with (or without) teaching labels from users and the environment, and then rely on implementing efficient mathematical algorithms and applying them onto the pre-collected data to induce language knowledge.

The methodology largely assumes that a learner (e.g. a machine) passively receives information from a language teacher (e.g. a human supervisor) in a one-way flow. In contrast, a young child is situated in social contexts and learns language through his everyday social interactions with caregivers. Language teachers dynamically adjust their behaviors based on their understanding of the learner's mental state. Thus, teachers provides "on-demand" information in real time learning. Meanwhile, the learner also plays an important role in learning-oriented interactions by actively generating actions to interact with the physical environment and to shape the teachers' responses and acquire just-in-need data for his learning. Thus, current machine learning studies focus on one aspect of learning – what kind of **learning device** can perform effective computations on the pre-collected data, but ignore an equally important aspect of the learning — the **learning environment** that a learner is situated in. An important way in which child language learning does not resemble passive machine learning approaches is that there is an active social partner in the learning environment. Parents always follow their child's bodily actions as cues to the child's attention and then name those things to which the child is attending. In this way, the reference uncertainty problem could be significantly simplified.

In light of the role of social interaction in child language learning, the present study attempts to systematically investigate its role in automatic language learning in machines. To do so, we propose and implement a new paradigm based on virtual reality techniques. The central idea is to use virtual humans as well-controlled agents to interact with real users. In this way, we can pre-program virtual humans to generate different kinds of behaviors and demonstrate different kinds of social capabilities. Then we can use them as a tool to systematically manipulate the learning environment in social interactions and measure the adaptive behaviors of real humans. Following this general idea, the present work builds a set of virtual learners who demonstrate different kinds of social understanding (e.g. following the eye gaze of real teachers) when real teachers are asked to interact with them and teach them object names. The questions we seek to answer are (1) how well real humans perceive behaviors and social skills of virtual humans; (2) whether and if so, in what ways real humans shape their behaviors based on their observation of virtual learners' states; and (3) what kind of learning environment real teachers provide through social interactions when they interact with different virtual humans.

## II. Related Work

Grounding language in the real world has recently attracted much attention in the AI community (Siskind, 1996; Steels & Vogt, 1997; Cohen, Oates, Adams, & Beal, 2001; Roy & Pentland, 2002; Weng, Zhang, & Chen, 2003; Yu et al., 2005). Among others, Siskind (1996) developed a mathematical model based on cross-situational learning and the principle of contrast, which learns word-meaning associations when presented with paired sequences of pre-segmented tokens and semantic representations. Roy and Pentland (2002) used the correlation of speech and vision to associate spoken utterances with a corresponding object's visual appearance. The learning algorithm is based on cross-modal mutual information to discover words and their visual associations. However, the role of social cues and social interaction are not considered in those systems.

Different from the above studies, Steels and Vogt (1997) proposed that language and the meanings of words could emerge as a result of social interactions between a group of distributed agents. They reported experiments in which autonomous visually grounded agents bootstrap meanings and language through adaptive language games without central control. Thus, automatic language learning is not a focus of their work. Yu et al. (2005) developed a multimodal learning system that uses egocentric multisensory data to first spot words from continuous speech and then associate action verbs and object names with their perceptually grounded meanings. The central idea is to utilize body movements as deictic references to associate temporally co-occurring data from different modalities. This work demonstrates the role of social cues in word learning but it doesn't study those embodied social cues in the context of dynamic social interactions between a language teacher (a human user) and a language learner (a computational system).

## III. Decoupling Social Interactions between Two Agents

A typical scenario of human language learning is like this: a language teacher provides spoken names of things in the physical world. Meanwhile, a language learner perceives and processes information collected from the learning environment and the social partner to build his vocabulary. In this kind of interaction, both the learner and the teacher dynamically adjust their behaviors based on the responsive actions from the other agent. For instance, imagine the learning situation wherein a mother plays with her child with a set of toys. At one moment, the language teacher may lead the interaction by attracting the learner's attention first and then naming those toys for the learner. At the other moment, the learner may lead the interaction by grasping a toy and playing with it while the teacher would follow the learner's attention and provide the linguistic label of the toy that the learner is playing with. From moment to moment, when two agents engage in a social interaction, the complex dynamic flow of the interaction depends on the learner's reaction to the teacher's behaviors which in turn will influence the teacher's response actions to the learner's reaction. Without interfering with the interaction, we can control neither the learner's responses nor the teacher's

actions. Without systematically manipulating some variables in the interaction, we cannot perform quantitative analyses of the role of social cues in human learning. Without a better understanding of human-human interaction in language learning, we cannot build a machine learner with necessary social capacities and behaviors encoded so that it can learn the meanings of words through its social interactions with a human teacher.

We argue that the key to solve the above puzzles is to decouple the social interactions between two agents without interfering with the interaction itself. More specifically, if the behaviors of one agent can be fully controlled and manipulated, then we can measure and analyze the behaviors of the other agent under various well-controlled learning situations. Here the controlled agent which could be either the language teacher and the language learner is expected to meet two basic requirements in the proposed paradigm: (1) the agent being study should behave naturally without noticing that the controlled agent is in an unnatural condition; (2) the controlled agent is expected to perform the same actions consistently across multiple users so that we can acquire sufficient amount of data under the exactly same condition for further analyses. These two requirements are not trivial to satisfy considering the fact that we cannot simply ask an experimenter to perform a set of pre-defined actions repeatedly and consistently across multiple users and ensure that they perceive the same actions at the same moments.

The present study proposes a novel approach to decouple social interaction using virtual reality techniques. The idea is to ask real people to interact with human-like virtual agents. The virtual humans we developed possess a set of social capabilities embodied by sensorimotor primitive actions, such as facial expressions, pointing at objects by hand, gazing and following a real user's attention. We design different virtual humans with different social capabilities and use them to measure and analyze the behaviors and responses of real humans when they interact with virtual humans.

## IV. A Virtual Reality Platform

The new experimental paradigm we developed uses virtual reality (VR) technologies to decouple complex social interactions between two agents. We build virtual humans equipped (pre-programmed) with different kinds of social-cognitive skills and ask real people to interact with virtual humans in a virtual environment.

Our VR interaction system consists of four components as shown in Figure 1:

- **A virtual environment** includes a virtual laboratory with furniture and a set of virtual objects that real people can manipulate in real time via a touch screen.
- **Virtual humans** can demonstrate different kinds of social skills and perform actions in the virtual environment.

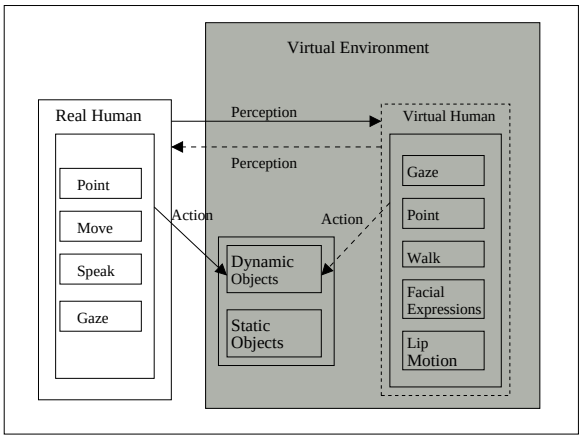


Fig. 1. Overview of system architecture. A real person and a virtual human interact in a virtual environment. We control the actions of the virtual person and measure the behavioral responses of a real person.

- **Multimodal interaction** between virtual humans and real people includes speaking, eye contact, pointing at, gazing at and moving virtual objects.
- **Data recording** monitors and records a participant's body movements including pointing and moving actions on virtual objects, eye gaze, and speech acts in real time.

One of the most important issues in our design is the "behavioral realism" of the virtual agents, which means that virtual humans should act and respond like real humans, or in other words, they should be believable in terms of both the physical actions of virtual humans themselves, and their social interactions with real humans (Turk, Bailenson, Beall, Blascovich, & Guadagno, 2005; Jasso & Triesch, 2005). In our design, we use Boston Dynamics's DI-Guy libraries to animate lifelike human characters that can be created and readily programmed to generate realistic human-like behaviors in the virtual world, including **gazing** and **pointing** at an object or a person in a specific 3D location, **walking** to a 3D location, and **moving lips** to synchronize with speech while **speaking**. In addition, the virtual human can generate 7 different kinds of **facial expression**, such as smile, trust, sad, mad and distrust. All these combine to result in smooth and lifelike behaviors being generated automatically.

## V. Experiment: Real Humans Teach Virtual Learners

### A. Design and Procedure

The present experiment studies how learners' reactions shape the behaviors of language teachers. As shown in Figure 2, real people were asked to teach virtual foreigners the names of several everyday objects. They were allowed to point to, gaze at and move those objects through a touch screen. There was no constraint about what they have to say or what they have to do. There were three conditions in this experiment wherein three virtual agents demonstrated different levels of engagement in interaction - engaged in

10%, 50% or 90% of total interaction time. When a virtual human is fully engaged in interaction, she would share visual attention with a real teacher by gazing at the object attended by a real teacher and generating positive facial expressions (e.g. smile, trust, etc.). While she is not engaged, she would look at somewhere else with negative facial expressions (e.g. sad, conniving, etc.). The objects attended by a real person are detected based on where he is looking as well as his actions on those objects through the touch screen. The attentional information is then sent to the virtual human so that she can switch her attention to the right objects in real time when she is in the engaged state.

We recruited 26 subjects who received course credits for participation. They were asked to interact with three virtual humans in total and one per condition. We randomly assigned the virtual humans to three levels of engagement, counterbalancing across participants. There were six trials in each engagement condition and three virtual objects were introduced in each trial. Thus, participants needed to teach $3\times$ 6=18 objects in each condition and 54 objects in all of the three conditions. Whenever they thought that the virtual learner already acquired three object names in the current trial, they could move to the next trial. We recorded real people's behaviors in interaction including their pointing and moving actions, speech acts and eye gaze. Moreover, they were asked to complete questionnaires at the end of the experiment. The questionnaires measured social intelligence of three virtual learners. They were also asked to provide their estimates of the percentage of time the virtual humans followed the human teacher's attention.

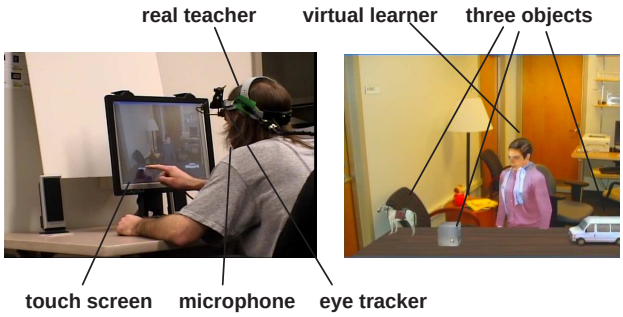


Fig. 2. Left: a participant wearing an eye tracker and a microphone interacts with the virtual human in a virtual environment through a touch screen. Right: the VR scene consists of a virtual human and three objects on a table in each trial.

### *B. Measures and Results*

A 5-point Likert scale was used for a set of 10 questions in our questionnaire. Those questions focus on different aspects of participants' perception of the social-cognitive skills of three virtual humans:

- **Joint attention and eye contact** We measured how much the participants felt that eye movements of virtual humans were natural, social and friendly. A representative question contributed to this measure is "I felt that the agent did not look enough at me".
- **Social intelligence/engagement** We calculated a score to measure how much the participants felt that virtual learners were engaged during interaction (0-not engaged at all, 5-fully engaged). A representative question in this measure is "the agent and I interacted very smoothly".
- **Overall intelligence** We calculated a score to measure participants' estimates of virtual learners' intelligence. An example question used here is "the agent is smart".
- **Gaze time estimation**: Participants were also asked to estimate the amount of time (on a scale of 0 to 100 percent) that virtual humans paid attention to their behaviors.

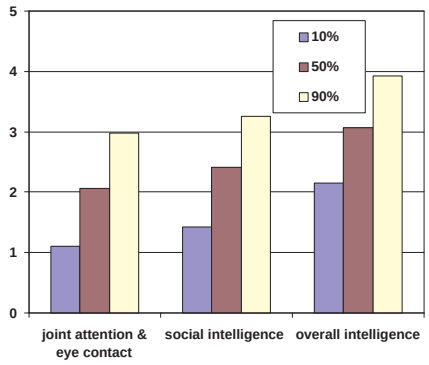


Fig. 3. A comparison of participants' evaluation of three virtual humans.

TABLE I

THE ESTIMATED ENGAGEMENT TIMES OF VIRTUAL HUMANS

| | 10% | 50% | 90% |
|---|---|---|---|
| gaze time | M= 22.50% SD= 22.10% | M=54.37% SD= 23.89% | M= 86% SD=16.1% |

Figure 3 shows a comparison of the results of three virtual humans with different engagement levels. Clearly, participants were aware of social behaviors of virtual humans and provided quite consistent estimates of their social sensitivities. Thus, the significant differences between three conditions are not surprising. We note that even when the virtual human almost fully engaged in interaction by following the real person's actions in 90% of the total time, most people were still not satisfied with the virtual human's social behaviors. Another observation is that they gave more credits to the high-level questions such as the overall intelligence of the virtual humans, but were less satisfied with more concrete issues, such as eye contact. This is true in all of the three conditions.

Table I shows the estimated times that virtual humans pay attention to participants' behaviors. Although the means of two out of three estimated times are close to 50% and 90% separately. Surprisingly, participants provided quite different

TABLE II

TWO EXAMPLES OF PARTICIPANTS' BEHAVIORAL DATA

| participant | # of actions | total time (seconds) | # of pointing | moving speed (pixels/second) | moving duration (seconds) | moving distance (pixels) | pause (seconds) |
|---|---|---|---|---|---|---|---|
| s1 10% | 51 | 285 | 6 | 307 | 5.60 | 270 | 0.77 |
| s1 50% | 70 | 307 | 8 | 148 | 4.39 | 687 | 1.25 |
| s1 90% | 37 | 383 | 3 | 188 | 10.36 | 176 | 0.83 |
| s2 10% | 86 | 395 | 3 | 192 | 4.71 | 667 | 1.33 |
| s2 50% | 52 | 293 | 0 | 86 | 5.63 | 473 | 0.99 |
| s2 90% | 93 | 295 | 14 | 114 | 3.50 | 385 | 1.27 |

estimates in all of three conditions. For instance, the low limit for the estimates in the 10% condition is 0%, indicating that some participants significantly overestimated the virtual human's engagement time. Meanwhile, some of them underestimate the times in the 90% condition as well. Further investigation is needed to explain this observation.

The purpose of these measures is to evaluate how participants perceive social-cognitive skills of virtual humans. The next question we ask is that based on their perception and estimates of a virtual human's social sensitivities, how they adapt their behaviors and actions in the task of teaching object names.

### *C. Behavioral Analysis at the Sensorimotor Level*

In addition to questionnaires, we also recorded multimodal behavioral data including participants' speech and actions on the touch screen mounted a computer monitor. Speech signals were sampled at 8000Hz and the sampling rate of actions on the touch screen is 60Hz.

The following global behavioral statistics were computed based on participants' actions on the touch screen – pointing at or moving one of the three virtual objects from moment to moment: (1) **number of actions** generated in each condition; (2) **total interaction time** in each condition; (3) **number of pointing action** performed; (4) **moving speed** measures how fast participants move an object on the table; (5) **moving duration** measures the amount of time on each action; (6) **moving distance** measures the average moving distance of each action, and (7) **pause duration** measures the average of time between two actions. Table II shows the data collected from two participants.

From those behavioral data, we first attempted to discover any consistent tendencies in three engagement conditions across all the participants. Among these seven measures, we found only one measure was changed consistently across all the participants – they tended to move objects at a faster speed when a virtual agent was not attending to their actions. Our working hypothesis is that moving object at a fast speed is a way to attract the virtual person's attention. Thus, if the virtual agent didn't pay attention to the moving object, then slowing down the movement would not help. For other six measures, we found that participants demonstrated different adaptive behaviors to attempt to attract the attention of virtual humans. For instance, some of them generated more actions with short durations in the 10 % condition while other generated a fewer actions but each of them took a longer duration. We also noticed that there might be potential correlations between different measures, suggesting we might be able to discover several behavioral patterns.

We next aligned the above measures to form a behavioral feature vector for each individual participant and used a hierarchical clustering algorithm to gradually group participants into several categories based on the similarities of their feature vectors. As a result, participants that share similar adaptive behavioral patterns in dealing with different conditions are clustered into one group. The following is a list of three representative groups:

- Group 1 (size = 6) is characterized by spending more time in interaction, moving objects more frequently and dramatically, when the virtual human is not engaged (10%). Participants in this group tried their best to change their actions to attract the virtual human's attention when she is not engaged.
- Group 2 (size = 5) is characterized by pausing (between actions) for a longer time, moving more dramatically in the 50% condition, and spending more time on each action in the 90% condition. This group seems to not bother to change their behaviors in the 10% condition.
- Group 3 (size = 5) is characterized by pausing for a longer time in the 10% condition, moving faster and in a longer duration in the 50% condition. This group seems to be more willing to interact with the virtual human that is partially engaged, but neither fully engaged nor rarely engaged.

The analysis above is just our first step to extract behavioral patterns of participants in different engagement conditions. We also recorded their speech and are working on calculating the synchrony between pointing/moving actions and speech. Meanwhile, recorded eye movement data can also be used to find when and for how long a real person is attending to the virtual human and when he is looking at the objects being manipulated. By integrating those multimodal data and computing potential correlations between them, we expect to obtain a better understanding of real people's actions in this multimodal learning interaction.

## VI. Discussion and Conclusion

Compared with using a real robot in a real environment, virtual humans are easy to implement and use mainly because we can neglect low-level technical problems, such as motor control of joint angles, which perfectly matches our research purposes. We are most interested in high-level social-cognitive skills in language learning. We attempt to answer how the behavioral-level actions, such as gazing and pointing, generated from both a language teacher and a language learner, dynamically coupled in real time to create the social learning environment, and how the language learner appreciates those social cues signaled by the teacher. Therefore, as far as those primitive actions generated by virtual humans look realistic, real people would treat them as social partners and are willing to interact with them. Moreover, the virtual platform has several special advantages in the study of social interaction: (1) Various virtual environments can be easily created and we can dynamically change or switch between different virtual scenes easily during an experiment; (2) the degree to fully control both virtual humans' behaviors and the virtual environment that real users and virtual humans share cannot be achieved with neither real robots nor human experimenters, which allows us to systematically study what aspects of the social environment are crucial for learning; and (3) we can easily maintain the consistency of the experimental environment and perfectly reproduce the experiments across multiple participants.

The present study proposes and implements a new experimental paradigm to study learning from multimodal interaction. We build virtual humans and control their behaviors to create different social partners that real people interacted with. We measured how well real people interact with virtual humans and how they shape their behaviors to adapt to different social-cognitive skills that virtual humans possess. We found that real people treat virtual humans as social partners when they interact with them, suggesting that we can further apply this experimental setup to create different interaction conditions by systematically manipulating the virtual human's behaviors. The preliminary results from the new method show the adaptive behaviors generated by real people are various. This observation demonstrated the difficulties in studying social interaction. Even after real people interact with the same virtual human in the same interaction environment while performing the same teaching task, they may behave quite differently.

To sum up, we report our first steps to systematically and quantitatively study social cues in word-to-world mappings, suggesting that learning words through social interaction will have the best chance to approaching human capacity. In addition, we argue that the new proposed paradigm itself is promising to study the roles of social cues in both human language learning and machine intelligence.

## References


Asada, M., MacDorman, K., Ishiguro, H., & Kuniyoshi, Y. (2001). Cognitive developmental robotics as a new paradigm for the design of humanoid robots. *Robotics and Autonomous Systems*, *37*(2), 185-193(9).

Breazeal, C., & Scassellati, B. (2002). Robots that imitate humans. *Trends in Cognitive Sciences*, *6*, 481-487.

Brooks, R. A., Breazeal, C., Irie, R., Kemp, C. C., & Marjanovi, M. (1998). Alternative essences of intelligence. In *Aaai '98/iaai '98: Proceedings of the fifteenth national/tenth conference on artificial intelligence/innovative applications of artificial intelligence* (p. 961-968). Menlo Park, CA, USA: American Association for Artificial Intelligence.

Cohen, P. R., Oates, T., Adams, N., & Beal, C. R. (2001). Robot baby 2001. In *Lecture notes in artificial intelligence* (Vol. 2225, p. 32 - 56). Washington D.C.

Harnad, S. (1990). The symbol grounding problem. *physica D*, *42*, 335-346.

Jasso, H., & Triesch, J. (2005). A virtual reality platform for modeling cognitive development. In *Biomimetic neural learning for intelligent robots: Intelligent systems, cognitive robotics, and neuroscience* (p. 211-224).

Quine, W. (1960). *Word and object.* Cambridge, MA: MIT Press.

Roy, D., & Pentland, A. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, *26*(1), 113-146.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*, 39-61.

Steels, L., & Vogt, P. (1997). Grounding adaptive language game in robotic agents. In C. Husbands & I. Harvey (Eds.), *Proc. of the 4th european conference on artificial life.* London: MIT Press.

Turk, M., Bailenson, J., Beall, A., Blascovich, J., & Guadagno, R. (2005). Multimodal transformed social interaction. In *Proceedings of the 6th international conference on multimodal interfaces* (p. 46-52). ACM Press.

Weng, J., Zhang, Y., & Chen, Y. (2003). Developing early senses about the world: 'object permanence' and visuoauditory real-time learning. In *Proc. international joint conf. on neural networks* (p. 2710 - 2715). Portland.

Yu, C., Ballard, D. H., & Aslin, R. N. (2005). The role of embodied intention in early lexical acquisition. *Cognitive Science*, *29*(6), 961-1005.