

Robot Command, Interrogation and Teaching via Social Interaction

Peter Ford Dominey
Institut des Sciences Cognitives, CNRS
67 Blvd. Pinel, 69675 Bron Cedex, France¹
<http://www.isc.cnrs.fr/dom/dommenu-en.htm>
(dominey@isc.cnrs.fr)

Alfredo Weitzenfeld
ITAM,
San Angel Tizapán, México DF, CP 0100
<http://robotica.itam.mx/ingles/index.phtml>
(alfredo@itam.mx)

ABSTRACT

The objective of the current research is to develop a generalized approach for human-machine interaction via spoken language that exploits recent developments in cognitive science - particularly notions of grammatical constructions as form-meaning mappings in language, and notions of shared intentions as distributed plans for interaction and collaboration. We will demonstrate this approach on two distinct robot platforms with human-robot interaction at three levels. The first level is that of commanding or directing the behavior of the system. The second level is that of interrogating or requesting an explanation from the system. The third and most advanced level is that of teaching the machine a new form of behavior. Within this context, we exploit social interaction in two manners. First, the robot will identify different human collaborators, and maintain a permanent record of their interactions in order to treat novices and experts in distinct manners. Second, the interactions are structured around shared intentions that guide the interactions in an ergonomic manner. We explore these aspects of communication on two distinct robotic platforms, the "Event Perceiver" and the Sony Aibo ERS7, and provide in the current paper the state of advancement of this work, and the initial lessons learned.

Categories and Subject Descriptors

J.4 [Social and Behavioral Sciences]: Psychology, sociology

H.5.2 [User Interfaces], Voice I/O; Natural language

General Terms

Algorithms, Performance, Design, Experimentation, Human Factors, Standardization.

Keywords

Grammatical construction, language, social cognition, shared intention, dialog.

Introduction

Ideally, research in Human-Robot Interaction will allow natural, ergonomic, and optimal communication and cooperation between humans and robotic systems. In order to make progress in this direction, we have identified two major requirements: First, we must work in real robotics environments in which technologists and researchers have already developed an extensive experience and set of needs with respect to HRI. Second, we must develop a domain independent language processing system that can be applied to arbitrary domains and that has psychological validity based on knowledge from social cognitive science. In response to the first requirement regarding the robotic context, we will study two distinct robotic platforms. The first, the "Event Perceiver" is a system that can perceive human events acted out with objects, and can thus generate descriptions of these actions. The second is the Sony AIBO ERS7 autonomous walking robot running the Tekkotsu (CMU) operating system, which provides access to a rich ensemble of sensory and motor capabilities. From the psychologically valid language context, we will base the interactions on a model of language and meaning correspondence developed by Dominey (et al. 2003) that has described both neurological and behavioral aspects of human language, and has been deployed in robotic contexts, and second, on the notion of shared intentions or plans (Tomasello 2003, et al. 2006) that will be used to guide the collaborative interaction between human and robot. The following sections introduce the two platforms, and the spoken language interface for command, control and teaching the two systems.

Comentario [AW1]: the phrase was truncated

Comentario [AW2]: The reference you have is Tomasello 2003...

The Event Perceiver

In a previous study, we reported on a system that could adaptively acquire a limited grammar based on training with human narrated video events (Dominey & Boucher 2005, 2006). An overview of the system is presented in Figure 1. Figure 1A illustrates the physical setup in which the human operator performs physical events with toy blocks in the field of view of a color CCD camera. Figure 1B illustrates a snapshot of the visual scene as observed by the image processing system. Figure 2 provides a schematic characterization of how the physical events are recognized by the image processing system. As illustrated in Figure 1, the human experimenter enacts and

¹ We thank Ann Cheylus for technical help in porting software between Unix and Windows.

simultaneously narrates visual scenes made up of events that occur between a red cylinder, a green block and a blue semicircle or “moon” on a black matte table surface. A video camera above the surface provides a video image that is processed by a color-based recognition and tracking system (Smart – Panlab, Barcelona Spain) that generates a time ordered sequence of the contacts that occur between objects that is subsequently processed for event analysis.

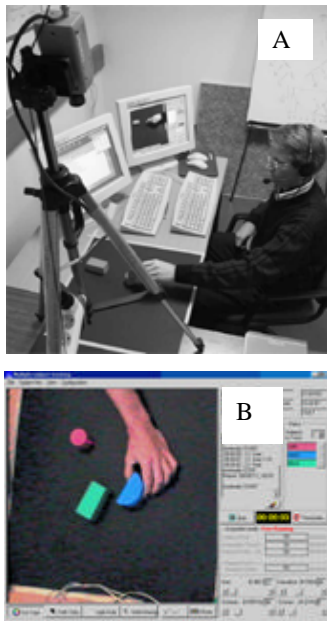


Figure 1. Overview of human-robot interaction platform. A. Human user interacting with the blocks, narrating events, and listening to system generated narrations. B. Snapshot of visual scene viewed by the CCD camera of the visual event processing system.

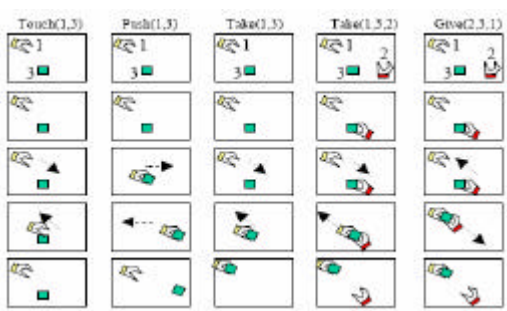


Figure 2. Temporal profile of contacts defining different event types: Touch, push, take, take-from, and give.

Using this platform, the human operator performs physical events and narrates his/her events. An image processing algorithm extracts the meaning of the events in terms of action(agent, object, recipient) descriptors. The event extraction algorithm detects physical contacts between objects (see Kotovsky & Baillargeon 1998), and then uses the temporal profile of contact sequences in order to categorize the events, based on the temporal schematic template illustrated in Figure 2. While details can be found in Dominey & Boucher (2005), the visual scene processing system is similar to related event extraction systems that rely on the characterization of complex physical events (e.g. give, take, stack) in terms of composition of physical primitives such as contact (e.g. Siskind 2001, Steels and Bailly 2003). Together with the event extraction system, a commercial speech to text system (IBM ViaVoice™) was used, such that each narrated event generated a well formed <sentence, meaning> pair.

Processing Sentences with Grammatical Constructions

These <sentence, meaning> pairs are used as input to a model that learns the sentence-to-meaning mappings as a form of template in which nouns and verbs can be replaced by new arguments in order to generate the corresponding new meanings. These templates or grammatical constructions (see Goldberg 1995) are identified by the configuration of grammatical markers or function words within the sentences (Bates et al. 1987).

Communicative Performance: We have demonstrated that this model can learn a variety of grammatical constructions in different languages (English and Japanese) (Dominey & Inui 2004). Each grammatical construction in the construction inventory corresponds to a mapping from sentence to meaning. This information can thus be used to perform the inverse transformation from meaning to sentence. For the initial sentence generation studies we concentrated on the 5 grammatical constructions below. These correspond to constructions with one verb and two or three arguments in which each of the different arguments can take the focus position at the head of the sentence. On the left are presented example sentences, and on the right, the corresponding generic construction. In the representation of the construction, the element that will be at the pragmatic focus is underlined. This information will be of use in selecting the correct construction to use under different discourse requirements.

This construction set provides sufficient linguistic flexibility, so that for example when the system is interrogated about the block, the moon or the triangle after describing the event *give(block, moon, triangle)*, the system can respond appropriately with sentences of type 3, 4 or 5, respectively. The important point is that each of these different constructions places the pragmatic focus on a different argument by placing it at the head of the sentence. Note that sentences 1-5 are specific sentences that exemplify the 5 constructions in question, and that these constructions each generalize to an open set of corresponding sentences.

The algorithm for selection of the construction type for sentence production takes as input a meaning coded in the form event(arg1, arg2, arg3), and an optional focus item (one of the three arguments). Based on this input, the system will

deterministically choose the appropriate two or three argument construction, with the appropriate focus structure, in a pragmatically relevant manner. Thus, in the dialog example below, the human user generates an event corresponding to gave(block, cylinder, moon) and then asks what happened to the moon. Based on these inputs, the system selects the three argument construction in which the recipient is the focus element (Construction 5). The predicate and arguments from the meaning are inserted into their appropriate positions, and the system thus responds: *The moon was gave the cylinder by the block.*

Sentence
1. The triangle pushed the moon.
2. The moon was pushed by the triangle.
3. The block gave the moon to the triangle.
4. The moon was given to the triangle by the block.
5. The triangle was given the moon by the block.
Construction <sentence, meaning>
1. <Agent event object, event(<i>agent</i> , object)>.
2. <Object was event by agent, event(<i>agent</i> , <i>object</i>)>
3. <Agent event object to recipient, event(<i>agent</i> , object, recipient)>
4. <Object was event to recipient by agent, event(<i>agent</i> , <i>object</i> , recipient)>
5. <Recipient was event object by agent, event(<i>agent</i> , object, recipient)>

Table 1. Sentences and corresponding constructions.

The next task at hand is to integrate these pieces, including (a) scene processing for event recognition, (b) sentence generation from scene description and response to questions, (c) speech recognition for posing questions, and (d) speech synthesis for responding - into an interactive environment. The CSLU Speech Tools Rapid application Development (RAD) (<http://cslu.cse.ogi.edu/toolkit/index.html>) provides useful capability in this context. The system provides dialog management with a flexible and powerful graphical user interface, coupled with the global ability to link speech recognition and synthesis to the conditional execution of code on the same machine or on remote machines via ftp and socket protocols. This results in a hub architecture with RAD at the hub and the vision processing, language model, speech-to-text and voice synthesis at the periphery.

Example Dialog for Interrogation with the Event Perceiver

Here we provide an example of typical dialog. Speech is indicated in *italics*.

Robot: *Welcome to the talking robot demonstration. Show me an object.*
Human: (Picks up and moves the block)
Robot: *That is the block. Another object?*
Human: *Yes* (moves the moon)
Robot: *That is the moon. Another object?*
Human: *No.*

Robot: *Shall we stop or go on?*
Human: *Go on.*
Robot: *When you are ready, make a movement with the objects so that I can see what you are doing. I will then describe what you did.*
Human: (picks up the block and uses it to move the cylinder to the moon).
Robot: *The block gave the cylinder to the moon. Was I correct?*
Human: *Yes.*
Robot: *Good for me. (stores results in database) Do you have any questions?*
Human: *Yes.*
Robot: *Would you like to know about the block, the moon or the cylinder?*
Human: *The moon.*
Robot: *The moon was gave the cylinder by the block. Do you have any questions?*
Human: *Yes.*
Robot: *Would you like to know about the block, the moon or the cylinder?*
Human: *The cylinder.*
Robot: *The cylinder was gave to the moon by the block.*

The sample dialog illustrates how vision and speech processing are combined in an interactive manner. Two points are of particular interest. In the response to questions, the system uses the focus element in order to determine which construction to use in the response. This illustrates the utility of the different grammatical constructions. However, we note that the two passivized sentences have a grammatical error, as "gave" is used, rather than "given". This type of error can be observed in inexperienced speakers either in first or second language acquisition. Correcting such errors requires that the different tenses are correctly associated with the different construction types, and will be addressed in future research.

These results demonstrate the capability to command the robot (with respect to whether objects or events will be processed), and to interrogate the robot, with respect to who did what to whom.

Shared Intentions for Learning

Perhaps the most interesting aspect of the three part "tell, ask, teach" scenario involves learning. Our goal is to provide a generalized platform independent learning capability that acquires new <percept, response> constructions. That is, we will use existing perceptual capabilities, and existing behavioral capabilities of the given system in order to bind these together into new, learned <percept, response> behaviors.

The common idea is to create new <percept, response> pairs that can be permanently archived and used in future interactions. Ad-hoc analysis of human-human interaction during teaching-learning reveals the existence of a general intentional plan that is shared between teachers and learners, which consists of three components. The first component involves specifying the percept that will be involved in the <percept, response> construction. This percept can be either a verbal command, or an internal state of the system that can originate from vision or from another sensor such as the battery charge state. The second component involves specifying what should be done in response to this percept. Again, the response can be either a verbal response or a motor response from the existing behavioral repertoire. The third component involves

the binding together of the <percept, response> construction, and validation that it was learned correctly. This requires the storage of this new construction in a construction data-base so that it can be accessed in the future. This will permit an open-ended capability for a variety of new types of communicative behavior.

For the Event Perceiver, this capability will be used for teaching the system to name and describe new geometrical configurations of the blocks. The human user will present a configuration of objects and name the configuration (e.g. four object placed in a square, and say «this is a square »). The system will learn this configuration, and the human will test with different positive and negative examples.

For the ERS7 this capability will be used to teach the system to respond with physical action or other behavioral (or internal state) responses to perceived objects, or perceived internal states. The user enters into a dialog context, and tells the robot that we are going to learn a new behavior. The robot asks what is the perceptual trigger of the behavior and the human responds. The robot then asks what is the response behavior, and the human responds. The robot links the <percept, response> pair together so that it can be used in the future. The human then enters into a dialog context from which he tests whether the new behavior has been learned.

Example Dialog for Teaching and Interrogation with the Event Perceiver

Here we provide an example of typical dialog. Speech is indicated in *italics*.

Robot: *Hello Peter, how are you since we last spoke today? Show me a relation.*
 Human: (Picks up and moves the blocks, forming a triangle)
 Robot: *What would you like to call that relation?*
 Human: *Triangle*
 Robot: *Are you sure you want to call that relation triangle?*
 Human: *Yes.*
 Robot: *OK, do you want to teach me a new relation, or let me show you that I can recognize relations?*
 Human: *Recognize relations.*
 Robot: *OK, show me a relation with the blocks and I will name it.*
 Human: (Picks up and moves the blocks, forming a square)
 Robot: *Sorry, I don't recognize that relation. Am I correct?*
 Human: *Yes.*
 Robot: *Good. Since this is a new relation, would you like me to learn it?*
 Human: *Yes.*
 Robot: *What would you like to call this relation?*
 Human: *Square.*
 Robot: *Ok, Do you have another relation to show me?*
 Human: *Yes.*
 Robot: *OK, show me a relation with the blocks and I will name it.*
 Human: (Picks up and moves the blocks, forming a square)
 Robot: *That is a square.*

This sample dialog illustrates how the system invites the user to teach it new relations, and to test it with learned relations. Ad-hoc testing with users revealed the utility of allowing the user to define unfamiliar relations that occur during the testing phase, as for the case of the square above. Gorniak

and Roy (2004) have demonstrated a related capability for a system that learns to describe spatial object configurations.

The Aibo ERS7

In order to demonstrate the generalization of this approach to an entirely different robotic platform we have begun a series of studies using the AIBO ERS7 mobile robot platform illustrated in Figure 4. We have installed on this robotic system an open architecture operating system, the Tekkotsu framework developed at CMU (<http://www-2.cs.cmu.edu/~tekkotsu/>), graphically depicted in Figure 4B. The Tekkotsu system provides vision and motor control processing running on the AIBO, with a telnet interface to a control program running on a host computer connected to the AIBO via wireless internet. Via this interface, the AIBO can be commanded to perform different actions in the Tekkotsu repertoire, and it can be interrogated with respect to various internal state variables.

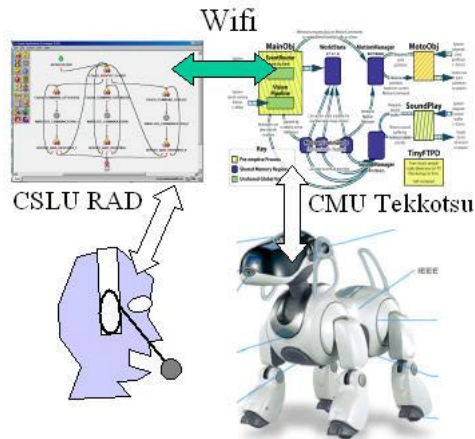


Figure 4. A schematic view of the Tekkotsu architecture for vision and behavioral event processing that runs on the Sony AIBO ERS7™ robot, and their integration with the CSLU RAD spoken language application toolkit.

Again, the spoken language interface to the AIBO is provided by the CSLU-RAD framework as described in Platform 1. In the future, this interface will be used between Aibo Robots in the RoboCup robot soccer context, and coaches that direct the activity of the robots. The RoboCup environment has been chosen because it is a well documented standardized robot environment that the ITAM team has extensive experience with, and thus provides a quantitative domain for evaluation of success (Martínez-Gómez, Medrano, Chavez, Muciño & Weitzenfeld 2005).

In order to demonstrate the feasibility of this approach we have developed an initial platform prototype that we now describe. Using the CSLU-RAD interface, the AIBO offers the human a set of possible actions that can be commanded, including shaking hands, chasing the red ball, or following the red ball with the gaze direction. The user then speaks one of

these commands and the robot performs the command. This is illustrated in the sample dialog below.

Example Dialog for Commanding the ERS7

Robot: *Hello I am AIBO the talking dog. I can look at the ball, chase the ball and shake hands. Ask me to do something.*
Human: *Chase the ball.*
Robot: *OK I will chase the ball for a bit and then I will stop. (Chases the ball for 10 seconds). So, you wanted me to chase the ball. Do you want to do something else?*
Human: *Yes.*
Robot: *I can look at the ball, chase the ball and shake hands. Ask me to do something.*
Human: *Shake hands.*
Robot: *(Raises and lowers its left paw). So, you wanted me to shake hands. Do you want to do something else?*
Human: *No.*
Robot: *Ok, I think I will take a rest now.*

This dialog demonstrates the feasibility of the use of a “generic” natural language interface to the AIBO via the Tekkotsu architecture, and provides a demonstration of the ability to verbally command the robot in this context. In the interrogation and learning context, it is straightforward to read status data from the AIBO in order to ask questions about the state of the battery, whether or not the AIBO can see the ball, etc. and to use the construction grammar framework for formulating the answers, and to associate different internal states with behavioral responses for learning.

Example Dialog for Teaching and Asking the ERS7

Robot: *Hello I am AIBO the talking dog. Do you want to teach me a new action, or test what I already know.*
Human: *New action.*
Robot: *Watch what I can do. (Stands up straight). Do you want to name this action, or associate it with a button press.*
Human: *Name.*
Robot: *What would you like to call this action?*
Human: *Stand up.*
Robot: *Ok I have learned to call this action Stand up. Do you want to teach me another action?*
Human: *Yes.*
Robot: *Watch what I can do. (raises its left fore-leg). Do you want to name this action, or associate it with a button press.*
Human: *Button.*
Robot: *Ok, press one of my buttons.*
Human: *(Presses the head button)*
Robot: *Ok, this action has been linked with the button you pressed.*
.....
Robot: *Do you want to test another action I know?*
Human: *Yes*
Robot: *Do you want to speak to me or to press a button*
Human: *Speak*
Robot: *Select one action by saying its name*
Human: *Bark*
Robot: *Dog Barks*
...
Robot: *Do you want to test another action I know?*

Human: *Yes*
Robot: *Do you want to speak to me or to press a button*
Human: *Press a button*
Robot: *Select one action by pressing a button*
Human: *(Presses the head button)*
Robot: *Dog raises its foreleg to “shake hands”*
...
Robot: *Ask me what you want to know about?*
Human: *Battery level*
Robot: *The battery level is normal, it is at 61 percent. Do you want to know anything else?*
Human: *Yes*
Robot: *Ask me what you want to know about?*
Human: *Battery temperature*
Robot: *The battery temperature is normal, it is 31 degrees. Do you want to know anything else?*
Human: *No.*

These dialog fragments illustrate how spoken language can be used to allow the human user to link verbal commands and button presses with different actions that the Aibo can perform. During the course of use of the robot with the learned commands, the user may discover that it would be more appropriate to associate a given action with both a verbal command and a button press, or to change existing associations. This flexibility is provided by the spoken language interface.

Incorporating Lessons Learned

The stated objective of the current research is to develop a generalized approach for human-machine interaction via spoken language that exploits recent developments in cognitive science - particularly notions of grammatical constructions as form-meaning mappings in language, and notions of shared intentions as distributed plans for human-robot interaction and collaboration. In order to do this, we tested human-robot interaction with the Event Perceiver, and the Aibo ERS7.

With respect to grammatical constructions, the Event Perceiver selected between different grammatical constructions to provide relevant descriptions of “who did what to whom” in the event descriptions. Our future research will exploit the flexibility provided by different grammatical constructions in the natural language input from the human user.

Capability	Robot Platforms	
	Event Perceiver	AIBO ERS7
1. Command		Command different actions (shake, chase the ball, etc.)
2. Interrogate	Ask who did what in a given action	Ask what is the battery state ? Can you see the ball ?
3. Teach	This is a stack This is a square, etc.	Associate perceptual events with behaviors. Head-touch -> Bark.

Table 1. “Tell, ask, and teach” capabilities in the two robotic platforms.

With respect to shared intentions, in social cognition, shared intentions are distributed plans in which two or more collaborators have a common representation of an action plan in which each plays specific roles with specific responsibilities with the aim of achieving some common goal (see Tomasello 2003, et al. 2006). In the current study, the common goals were well defined in advance (e.g. teaching the robots new relations or new behaviors), and so the shared intentions could be built into the dialog management system. Subsequent ad-hoc testing revealed cases where these rigidly structured intentional plans did not allow flexibility, and we could thus modify them to render them more ergonomic. Future research will address how shared intentions can be acquired and modified.

The first important lesson learned from this activity is that once the input/output interfaces for a robot have been established - allowing an external system to command the system, and to have access to sensor values and other internal state information, - a layer of social interaction can be applied on top of this I/O level, providing a rich, flexible and adaptive human-robot interface capability.

The second lesson has to do with the ergonomics of the system. An initial evaluation period revealed that while technically we had demonstrated command, interrogation and teaching, the user interface ergonomics was somewhat clumsy. In particular the dialog pathways were somewhat constrained, with several levels of hierarchical structure in which the user had to navigate the control structure with several single word commands in order to teach the system a new relation, and then again to demonstrate the knowledge, rather than being able to do these operations in more natural single sentences. In order to address this issue, we reorganized the tree structure of the dialog management into a hub-based structure in which context changes are made in a single step. Also, in order to focus the interactions, we worked around scenarios in which the human and robot (ERS7) collaborate around the shared goal of finding the ball and moving it towards a landmark so that the ERS7 can see both at the same time.

Human: Can you see the ball?
 Robot: No, I don't see it.
 Human: Look to the left.
 Robot: I still don't see it.
 Human: Look a little more to the left.
 Robot: OK I see it now.
 Human: Ok, now go over to the ball.
 Robot: (starts walking to ball).
 Human: OK stop.
 Human: Do you see the goal?
 Robot: What is the goal?
 Human: It is that blue thing.
 Robot: Ok the goal is blue.
 Robot: No, I don't see it.
 Human: Turn right
 Robot: OK I see it now.
 Human: You see what?
 Robot: I see the goal.
 Human: Ok, now take the ball to the goal.
 Robot: OK (starts walking, pushing the ball towards the goal).
 Human: Stop
 Robot: (stops)
 Human: Back up 3 steps.
 Robot: OK

Human: Now can you see the ball and the goal?
 Robot: Yes.

Acknowledgements

Supported by the French-Mexican LAFMI, the ACI TTT Projects in France and the UC MEXUS CONACYT, CONACYT grant #42440, and "Asociación Mexicana de Cultura, S. A." in Mexico.

References

- Bates E, McNew S, MacWhinney B, Devescovi A, Smith S (1982) Functional constraints on sentence processing: A cross linguistic study, *Cognition* (11) 245-299.
- Dominey PF (2005) Towards a Construction-Based Account of Shared Intentions in Social Cognition, Comment on Tomasello et al. Understanding and sharing intentions: The origins of cultural cognition, *Behavioral and Brain Sciences*
- Dominey PF, Boucher (2005) Developmental stages of perception and language acquisition in a perceptually grounded robot, In press, *Cognitive Systems Research*
- Dominey PF, Boucher (2005) Learning To Talk About Events From Narrated Video in the Construction Grammar Framework, *Artificial Intelligence*, In press
- Dominey PF, Hoen M, Lelekov T, Blanc JM (2003) Neurological basis of language in sequential cognition: Evidence from simulation, aphasia and ERP studies, *Brain and Language*, 86(2):207-25
- Dominey PF, Inui T (2004) A Developmental Model of Syntax Acquisition in the Construction Grammar Framework with Cross-Linguistic Validation in English and Japanese, *Proceedings of the CoLing Workshop on Psycho-Computational Models of Language Acquisition*, Geneva, 33-40
- Goldberg A (1995) *Constructions*. U Chicago Press, Chicago and London.
- Gorniak P, Roy D (2004). Grounded Semantic Composition for Visual Scenes, *Journal of Artificial Intelligence Research*, Volume 21, pages 429-470.
- Kotovsky L, Baillargeon R, (1998), The development of calibration-based reasoning about collision events in young infants, *Cognition*, 67, 311-351
- Martínez-Gómez, J.A., Medrano, A., Chavez, A., Muciño, B. and Weitzenfeld, A., (2005), Eagle Knights AIBO Team, Proc. VII World Robocup 2005, Osaka, Japan, July 13-17 (accepted for publication)
- Siskind JM (2001) Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal of AI Research* (15) 31-90
- Steels, L. and Baillie, JC. (2003). Shared Grounding of Event Descriptions by Autonomous Robots. *Robotics and Autonomous Systems*, 43(2-3):163-173. 2002
- Tomasello, M. (2003) Constructing a language: A usage-based theory of language acquisition. Harvard University Press, Cambridge.
- Tomasello M, Carpenter M, Call J, Behne T, Moll H (2006) Understanding and sharing intentions: The origins of cultural cognition, *Behavioral and Brain Sciences*

Comentario [AW3]: Not referenced