Levels of Interaction Allowing Humans to Command, Interrogate and Teach a Communicating Object: Lessons Learned From Two Robotic Platform

Peter Ford Dominey [(1)] *& Alfredo Weitzenfeld* [(2)]

[(1)] Institut des Sciences Cognitives, CNRS
67 Blvd. Pinel, 69675 Bron Cedex, France
http://www.isc.cnrs.fr/dom/dommenu-en.htm
dominey@ isc.cnrs.fr

[(2)] ITAM
San Angel Tizapán, México DF, CP 0100
http://robotica.itam.mx/ingles/index.phtml
alfredo@itam.mx

Abstract

As robotic systems become increasingly capable of complex sensory, motor and information processing functions, the ability to interact with them in an ergonomic, real-time and adaptive manner becomes an increasingly pressing concern. In this context, the physical characteristics of the robotic device should become less of a direct concern, with the device being treated as a system that receives information, acts on that information, and produces information. Once the input and output protocols for a given system are well established, humans should be able to interact with these systems via a standardized spoken language interface that can be tailored if necessary to the specific system.

The objective of this research is to develop a generalized approach for human-machine interaction via spoken language that allows 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. The mapping between sentences and meanings in these interactions is guided by a neuropsychologically inspired model of grammatical construction processing. We explore these three levels of communication on two distinct robotic platforms. The novelty of this work lies in the use of the construction grammar formalism for binding language to meaning extracted from video in a generative and productive manner, and in thus allowing the human to use language to command, interrogate and modify the behavior of the robotic systems.

1. 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 study a real robotics environment in which technologists and researchers have already developed an extensive experience and set of needs with respect to HRI. Second, we must study a domain independent language processing system that has psychological validity, and that can be mapped onto arbitrary domains. In response to the first requirement regarding the robotic context, we will study two distinct robotic platforms. The first is a system that can perceive human events acted out with objects, and can thus generate descriptions of these actions. The second platform involves Robot Command and Control in the international context of robot soccer playing, in which the Weitzenfeld group competes at the international level. From the psychologically valid language context, we will study 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.

2. Platform 1

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). 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 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.

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.

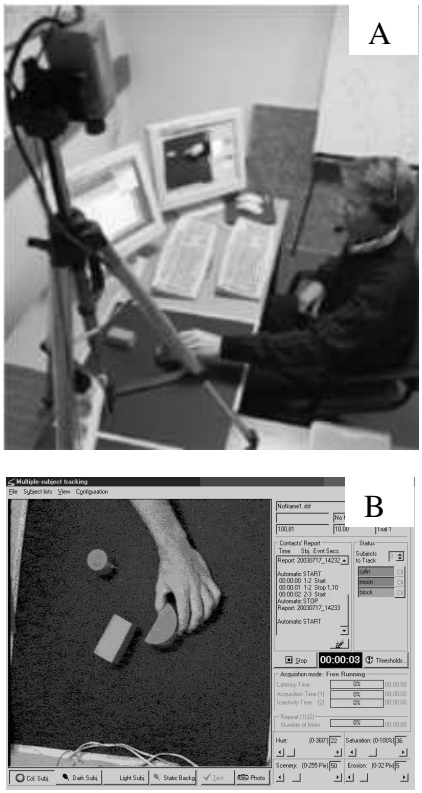

Figure 1. Overview of human-robot interaction platform. A. Human user interacting with the blocks, narrating events, and listening to system generated narrations. B. Snapshot of visual scene viewed by the CCD camera of the visual event processing system.

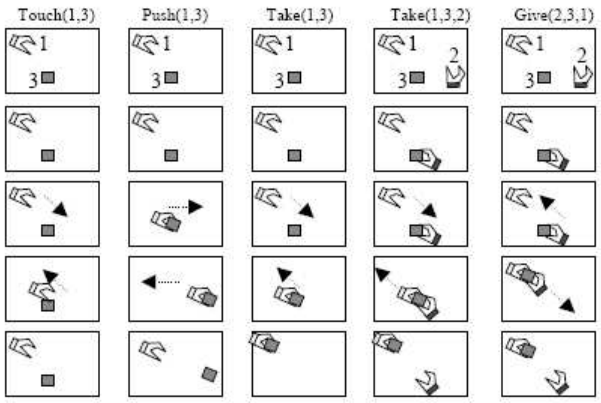

Figure 2. Temporal profile of contacts defining different event types: Touch, push, take, take-from, and give.

2.1 Processing Sentences with Grammatical Constructions

These <sentence, meaning> pairs are used as input to the model in Figure 3 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). Here we provide a brief overview of the model, and define the representations and functions of each component of the model using the example sentence "The ball was given to Jean by Marie," and the corresponding meaning "gave(Marie, Ball, John)" in Figure 2.

Sentences: Words in sentences, and elements in the scene are coded as single bits in respective 25-element vectors, and sentences can be of arbitrary length. On input, Open class words (ball, given, Jean, Marie) are stored in the Open Class Array (OCA), which is thus an array of 6 x 25 element vectors, corresponding to a capacity to encode up to 6 open class words per sentence. Open class words correspond to single word noun or verb phrases, and determiners do not count as function words.

Identifying Constructions: Closed class words (e.g. was, to, by) are encoded in the Construction Index, a 25 element vector, by an algorithm that preserves the identity and order of arrival of the input closed class elements. This thus uniquely identifies each grammatical construction type, and serves as an index into a database of <form, meaning> mappings.

Meaning: The meaning component of the <sentence, meaning> pair is encoded in a predicate-argument format in the Scene Event Array (SEA). The SEA is also a 6 x 25 array that encodes meaning in a predicate-argument representation. In this example the predicate is *gave,* and the arguments corresponding to agent, object and recipient are *Marie, Ball, John*. The SEA thus encodes one predicate and up to 5 arguments, each as a 25 element vector. During learning, complete <sentence, meaning> pairs are provided as input. In subsequent testing, given a novel sentence as input, the system can generate the corresponding meaning.

Sentence-meaning mapping: The first step in the sentence-meaning mapping process is to extract the meaning of the open class words and store them in the Predicted Referents Array (PRA). The word meanings are extracted from the real-valued WordToReferent matrix that stores learned mappings from input word vectors to output meaning vectors. The second step is to determine the appropriate mapping of the separate items in the PredictedReferentsArray onto the predicate and argument positions of the SceneEventArray. This is the "form to meaning" mapping component of the grammatical construction. PRA items are thus mapped onto their roles in the Scene Event Array (SEA) by the FormToMeaning mapping, specific to each construction type. FormToMeaning is thus a 6x6 real-valued matrix. This mapping is retrieved from ConstructionInventory, based on the ConstructionIndex that encodes the closed class words that characterize each sentence type. The ConstructionIndex is a 25 element vector, and the FormToMeaning mapping is a 6x6 real-valued matrix, corresponding to 36 real values. Thus the ConstructionInventory is a 25x36 real-valued matrix that defines the learned mappings from ConstructionIndex vectors onto 6x6 FormToMeaning matrices. Note that in 3A and 3B the ConstructionIndices are different, thus allowing the corresponding FormToMeaning mappings to be handled separately.

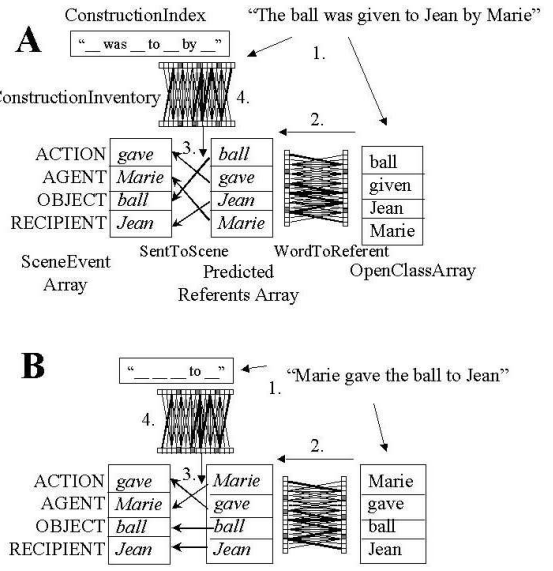

Figure 3. Model Overview: Processing of active and passive sentence types in A, B, respectively. On input, Open class words populate the Open Class Array (OCA), and closed class words populate the Construction index. Visual Scene Analysis populates the Scene Event Array (SEA) with the extracted meaning as scene elements. Words in OCA are translated to Predicted Referents via the WordToReferent mapping to populate the Predicted Referents Array (PRA). PRA elements are mapped onto their roles in the Scene Event Array (SEA) by the SentenceToScene mapping, specific to each sentence type. This mapping is retrieved from Construction Inventory, via the ConstructionIndex that encodes the closed class words that characterize each sentence type. Words in sentences, and elements in the scene are coded as single ON bits in respective 25-element vectors.

2.2 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.

Table 1. Sentences and corresponding constructions.

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(agent, object>.
2. <Object was *event* by *agent, event(agent, object>*
3. <Agent event object to *recipient, event(agent, object, recipient)>*
4. <Object was *event* to *recipient* by *agent, event(agent, object, recipient)>*
5. <Recipient was *event object* by *agent, event(agent, object, recipient)>*

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.*

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 with Platform 1

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. Gorniak and Roy (2004) have demonstrated a related capability for a system that learns to describe spatial object configurations, but with less flexibility in the use of appropriate grammatical constructions.

3 Platform 2 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.

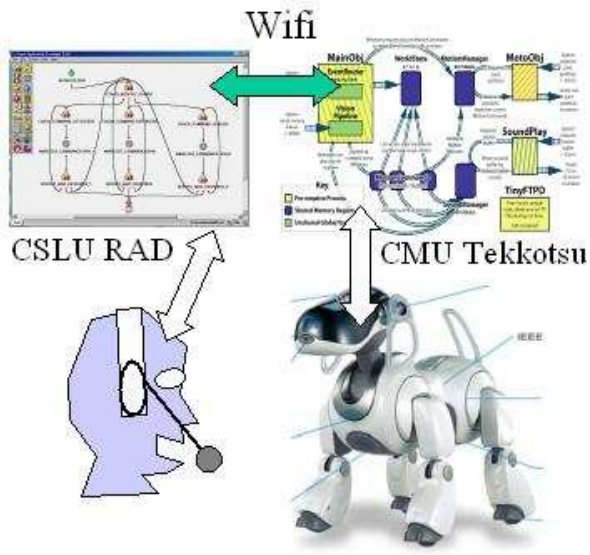

Figure 4. A schematic view of the Tekkotsu architecture for vision and behavioral event processing that runs on the Sony AIBO ERS7™ robot, and their integration with the CSLU RAD spoken language application toolkit.

The objective of the current exercise is to provide a spoken language interface to the AIBO using the CSLU-RAD framework as described in Platform 1. The 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.

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.

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 this same context it will be 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 In this sense we have demonstrated the first steps towards the development of a generic communication architecture that can be adapted to different robot platforms.

3. Learning

The final 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 (Dominey et al. 2006).

In both of these Platform contexts common idea is to create new <percept, response> pairs that can be permanently archived and used in future interactions. This requirement breaks down into three components. The first component involves specifying to the system the nature of 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 to the system 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 is the binding together of the <percept, response> construction, and 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 Platform 1 this capability has been used for teaching the system to name and describe new geometrical configurations of the blocks. The human user presents 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 Platform 2 this capability has been 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 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: *Stand up*
Robot: Dog stands up

4. Lessons Learned and Conclusions

The research described here represents work in progress towards a generic control architecture for communicating systems that allows the human to "tell, ask, and teach" the system. This is summarized in Table 2. For the principal lessons learned there is good news and bad news (or rather news about hard work ahead, which indeed can be considered good news.) The good news is that given a system that has well defined input, processing and output behavior, it is technically feasible to insert this system into a spoken language communication context that allows the user to tell, ask, and teach the system to do things. This may require some system specific adaptations concerning communication protocols and data formats, but these issues can be addressed. The tough news is that this is still not human-like communication. A large part of what is communicated between humans is not spoken, and rather relies on the collaborative construction of internal representations of shared goals and intentions (Tomasello et al in press). What this means is that more than just building verbally guided interfaces to communicative systems, we must endow these systems with representations of their interaction with the human user. These representations will be shared between the human user and the communicative system, and will allow more human-like interactions to take place (Tomasello 2003). Results from our ongoing research permit the first steps in this direction (Dominey 2005).

Table 2. Status of "tell, ask, and teach" capabilities in the two robotic platforms.

	Robot Platforms	
Capability	**Platform 1. Event Vision and Description**	**Platform 2. Behaving Autonomous Robot**
1. Tell		Command different actions (shake, chase the ball, etc.)
2. Ask	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.

In conclusion, there are two distinct novel components of this work. The first lies in the manner of linking of language to meaning that has been extracted from video. The meaning extraction is similar to that described by Siskind (2001). The resulting meaning in a predicate-argument format is mapped onto language in a novel fashion via learned grammatical constructions (Goldberg 1995). Learned constructions (e.g. Table 1) can then be used in comprehension and production of novel sentences. Part of the novelty is that the system is not language dependant, and has been demonstrated in English and Japanese (Dominey & Inui 2004). The second novel aspect is that we address how this framework can be used in order to teach the systems to recognize new perceptual events (Platform 1), and to respond to perceptual events with behaviors in new ways (Platform 2). That is, we use language to modify the behavior of robotic systems, in order to allow a more flexible and adaptive human-robot interaction (Dominey et al. 2005).

Acknowledgements

Supported by the French-Mexican LAFMI, and the ACI TTT Projects.

References

[1] Bates E, McNew S, MacWhinney B, Devescovi A, Smith S (1982) Functional constraints on sentence processing: A cross linguistic study, *Cognition* (11) 245-299.

[2] Chang NC, Maia TV (2001) Grounded learning of grammatical constructions, *AAAI Spring Symp. On Learning Grounded Representations*, Stanford CA.

[3] Dominey PF (2000) Conceptual Grounding in Simulation Studies of Language Acquisition, *Evolution of Communication*, 4(1), 57-85.

[4] 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*

[5] Dominey PF, Alvarez M, Gao B, Jeambrun M, Weitzenfeld A, Medrano A (2005) Robot Command, Interrogation and Teaching via Social Interaction, *Proc. IEEE Conf. On Humanoid Robotics* 2005.

[6] Dominey PF, Boucher (2005) Developmental stages of perception and language acquisition in a perceptually grounded robot, In press, *Cognitive Systems Research*

[7] 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

[8] 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

[9] Goldberg A (1995) *Constructions*. U Chicago Press, Chicago and London.

[10] Gorniak P, Roy D (2004). Grounded Semantic Composition for Visual Scenes, Journal of Artificial Intelligence Research, Volume 21, pages 429-470.

[11] Kotovsky L, Baillargeon R, The development of calibration-based reasoning about collision events in young infants. 1998, *Cognition*, 67, 311-351

[12] 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

[13] Steels, L. and Baillie, JC. (2003). Shared Grounding of Event Descriptions by Autonomous Robots. Robotics and Autonomous Systems, 43(2-3):163--173. 2002

[14] Tomasello, M. (2003) Constructing a language: A usage-based theory of language acquisition. Harvard University Press, Cambridge.