Using Cognitive Semantics to Integrate Perception and Motion in a Behavior-based Robot

D. Paul Benjamin
Pace University
1 Pace Plaza
New York, NY 10038
benjamin@pace.edu

Deryle Lonsdale
Brigham Young University
Dept. Linguistics & English Language
Provo, Utah 84602
lonz@byu.edu

Damian Lyons
Fordham University
340 JMH, 441 E. Fordham Rd.
Bronx, NY 10458
dlyons@fordham.edu

Siddtharth Patel
Pace University
1 Pace Plaza
New York, NY 10038
siddtharth861@yahoo.com

Abstract

The ADAPT project is a collaboration of three universities building a unified architecture for mobile robots. The goal of this project is to endow robots with the full range of cognitive abilities, including perception, use of natural language, learning and the ability to solve complex problems. The perspective of this work is that such an architecture should be based on language and visualization.

ADAPT is based on an the structure and semantics of language, and more specifically on algebraic linguistics and visualization of semantics. ADAPT organizes its knowledge using linguistic robot schemas, which implement linguistic units within a concurrent, distributed programming language. Each schema is associated with one or more 3D visualizations that provide its semantics. These visualizations are dynamic, and are composed within a virtual world to create ADAPT's representation of itself and its environment.

1: Introduction

The ADAPT project (**A**daptive **D**ynamics and **A**ctive **P**erception for **T**hought) is a collaboration of three university research groups at Pace University, Brigham Young University, and Fordham University that is building a robot cognitive architecture that integrates the structures designed by cognitive scientists and linguists with those developed by robotics researchers for real-time perception and control. ADAPT is under development on Pioneer robots in the Pace University Robotics Lab and the Fordham University Robotics Lab. Publications describing ADAPT are [1, 2, 3, 4].

We are exploring how linguistic structures interact with perception and problem solving, and in particular how symbolic reasoning can respond to a continuous, dynamic environment. ADAPT is an architecture intended to explore the integration of perception, problem solving and natural language at a deeper structural level. We believe that the integration of these capabilities must stem from a central organizing principle, and in ADAPT that principle is the mathematical structure of language. Language provides not only the means of interaction between people and ADAPT, but also provides the basis for the robot's representation of the world, and for the integration of perception and problem solving.

2: Background

A truly cognitive architecture has not yet been implemented in robotics. Robots have been programmed to perform specific tasks such as mowing the lawn or navigating in the desert, and these accomplishments can be impressive, but robots still cannot act autonomously to choose tasks and devise ways to perform them. Even when performing their allotted tasks, they lack flexibility in reacting to unforeseen situations. Currently, the design of all important perceptual and decision-making structures is done by the programmers before the robot begins its task. The semantics for the symbols and structures the robot uses is determined and fixed by these programmers. This leads to fragmented abilities and brittle performance. The robots cannot adapt their knowledge to the task, cannot solve tasks that are even slightly different from those they have been programmed to solve, cannot communicate effectively with humans about their goals and performance, and just don't seem to understand their environment. This is a principal stumbling block that prevents robots from achieving high levels of performance on complex tasks, especially tasks involving interaction with people.

Symbolic approaches to meaning (i.e. semantics) can be loosely characterized into three differing types: (1) referential or denotational, where an attempt is made to relate symbols to external objects in the real world via logical and mathematical methods including set theory and model-theoretic representations; (2) psychological or mentalist, where an attempt is made to relate symbols to the cognitive structures in the mind that represent one's mental characterization of the real world; and (3) pragmatic or social, where an attempt is made to view communication as a social activity and where meaning is a multi-party phenomenon, a construct that emerges via such devices as interaction and the notions of self and of agency, social conventions, argumentation, negotiation, and conversation [11].

All three strands of research are actively being pursued from theoretical and application perspectives. This is even true in the field of robotics and human-computer interaction. For example, the CN architecture [18] adopts

This work was supported in part by DOE grant ER 45903.

the denotational approach, as does the Bielefeld robot [17] and the CoSy Explorer [16]. Green [7] manipulates an internal model to represent relationships in a cognitive semantics framework. van Dartel and Postma [15] use an interesting blend of approaches 1 and 2, without relating it to human-robot interaction.

We have already pursued the traditional Tarskian referential/denotational approach, using interpretive semantics: an input utterance is tokenized, morphologically processed, syntactically parsed, and then discrete pieces of syntactic constituency are mapped by operators to form a lexical conceptual structure, a semantic knowledge representation. This representation is then leveraged in further processing: to perform logic inferences, drive discourse understanding and generation, and feed a derived representation involving first-order predicate logic.

In this research we have reached a point where this type of derivation of meaning must be informed by further knowledge about the participants' mental models, cognitive states, and pragmatic situations.

Our current work is enhancing the current semantic processing with a further level of analysis, one based on cognitive semantics. This approach to semantic description is particularly appropriate for processing interactions that involve perception-based situations. It also has knowledge representations that allow for the encoding of perspective, figure/ground, landmarks, embodiment, spatial relationships, scalar properties, and physical traits of the environment (e.g. those necessary for navigation). All of these are not as easily encoded in an exclusively denotational semantics [6, 8].

Our research goal is to enhance the current system's semantic capabilities by adding functionality to take into consideration cognitive and pragmatic information. This will allow for novel robotics capabilities in the areas of interaction and autonomy, important linguistic insights into the integration of formalist and functionalist approaches to semantics, and timely cognitive investigation into theoretical and practical questions about how natural language and other non-linguistic tasks interrelate.

3: Comprehension by Visualization

The design of our robot architecture is based on the belief that language is central to human intelligence [5, 14] and thus should be used as a central organizing principle of an artificial intelligence. This means that language is not only used for communication, but also to represent and organize the robot’s knowledge about itself and the world, and to structure the robot’s reasoning and planning processes. Knowledge is organized according to units arising from the semantics of natural language: words, phrases, sentences, and discourse contexts. Each such unit of knowledge is called a *linguistic schema*, and is connected to other schemas that are related functionally (whether the function is physical or linguistic).

The central goal of our work is to develop effective methods for robots to comprehend their environment. In our language-based architecture, this means developing effective methods for comprehending language. Our approach models language comprehension as a process of trying to recreate the observed speech by hypothesizing various sets of goals and beliefs for the communicating agents, generating their speech based on these assumptions and comparing it with the observed speech. This knowledge-intensive approach to comprehension has a history within AI and in particular in machine learning.

We have extended this approach to apply to comprehension of all observed behaviors, whether or not they include speech, because we view language comprehension as a special case of behavior comprehension. To say it the other way around, we believe that comprehension of non-speech behaviors is necessary for language comprehension. This necessity stems from two causes. The first is that the semantics of many words (especially verbs) requires comprehension of the activity they denote. The second is that speech is typically enhanced with many non-verbal actions, such as hand movements, facial expressions and postures.

Furthermore, we believe that the comprehension requires *visualization*, and that the semantics of language requires visual representations. We view visualization as consisting of both a perceptual component and a reasoning component. The perceptual component is performed using the same perceptual mechanism that the robot uses to perceive its environment; the difference is that visualization perceives a simulation of the environment. Visual reasoning manipulates and superimposes representations that consist of a combination of symbolic knowledge and 3D animations.

Comprehension by generation requires the robot to be able to create different situations in which it can generate behaviors of robots, people and physical systems, and perceive the results of these behaviors. This requires implementing a virtual world that the robot can control.

ADAPT’s virtual world is a multimedia simulation platform capable of realistic simulations of physical phenomena. It combines the various forms of map information found in most robots: topological, metric and conceptual information. ADAPT completely controls this virtual world, and can create arbitrary objects and behaviors in it, including nonexistent objects and behaviors that were not actually observed. Central to ADAPT’s use of its virtual world is its ability to view these constructions from any point. This enables ADAPT to create visual representations with desired properties.

This approach to visualization is very different from previous work on reasoning about spatial relationships. ADAPT does not just turn spatial relationships into symbolic terms to be used in reasoning, but instead can reason visually about spatial relationships by constructing instances of those relationships, viewing them from various angles, and superimposing them.

In the current implementation, ADAPT's world model is the Ogre3D open source gaming platform (http://www.ogre3d.org). Ogre gives the robot the ability to create a detailed and dynamic virtual model of its environment, by providing sophisticated graphics and rendering capabilities together with a physics engine based on the PhysX physics engine. Ogre models a wide variety of dynamic environments, including modeling other agents moving and acting in those environments.

ADAPT uses this virtual world in a novel way. Typical robotics architectures connect their sensory mechanisms to their world models, so that sensory data is processed and modeled in the world model. The reasoning engine then operates on the world model to plan the robot's behaviors. This type of architecture treats perception as a separate process from the central reasoning, and typically the implementation reflects this, e.g. a computer vision module processes the vision data and puts symbolic representations of the recognized objects and their relationships in the world model, and the reasoning engine then manipulates these symbols to plan and learn. The reasoning engine does not process the sensory data.

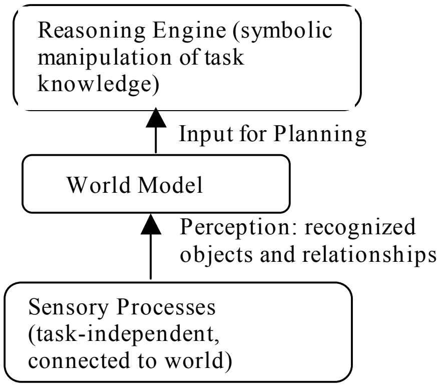
Reasoning Engine (symbolic manipulation of task knowledge)
Input for Planning
World Model
Perception: recognized objects and relationships
Sensory Processes (task-independent, connected to world)

Fig. 1. A typical robot architecture organization.

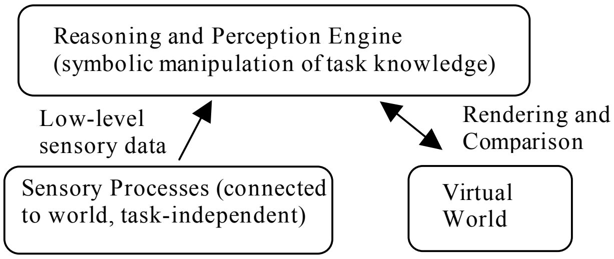
Reasoning and Perception Engine (symbolic manipulation of task knowledge)
Low-level sensory data
Rendering and Comparison
Sensory Processes (connected to world, task-independent)
Virtual World

Fig. 2. ADAPT's organization

In contrast, ADAPT's virtual world is not connected to its sensory processes. ADAPT's sensory data is placed directly in the reasoning engine (after some low-level processing); the reasoning engine's principal task in ADAPT is to reason about how to model the data. It does this in the following way:

It creates virtual entities and behaviors in Ogre.

It senses in the virtual world, using the same position and orientation as in the real world, and using the same sensors. For example, if ADAPT is modeling visual data, it grabs graphics input from Ogre, and if it is modeling sonar data, it grabs distance data from Ogre in the directions of the actual sonars.

It compares the virtual sensory data with the real sensory data, using a least-squares measure to find the degree of disagreement.

The reasoning engine searches alternative combinations of virtual entities and behaviors to attempt to minimize the measured disagreement. In this way, *perception becomes a problem-solving process*. This enables all the knowledge of the system to be brought to bear on perception, and unifies the reasoning and learning processes of problem solving with those of perception.

This search can be long and expensive; for this approach to comprehension to be practical, an effective speedup learning mechanism is required to store the results of this search. ADAPT contains a knowledge compilation method that stores generalized results of each successful search. One of the main research goals of our project is to quantify the effectiveness of this approach.

Visualization is also used in ADAPT for *predictive vision*: the robot predicts what it expects to see based on its virtual world and pays attention only to significant differences. This part of the project is detailed in [2].

4: Natural Language in ADAPT

Communication between humans and the robot is handled in ADAPT via a natural language system implemented within a cognitive modeling framework. The system supports spoken human language input via an interface with Sphinx. Textual inputs representing best-guess transcriptions from the ASR system are pipelined as whole utterances into the natural language component.

ADAPT processes each word individually and performs the following operations in order to understand the input text:

- lexical access (retrieving morphological, syntactic, and semantic information for each word from its lexicon)
- syntactic model construction (linking together pieces of an X-bar parse tree)
- semantic model construction (fusing together pieces of a lexical-conceptual structure)
- discourse model construction (extracting global coherence from individual utterances)

As is typically implemented for human/robotic interaction, our system uses a dialogue-based discourse interface between the robot and the NL component. The system's discourse processing involves aspects of input text comprehension (including referring to the prior results of syntax and semantics where necessary) and generation (i.e. the production of linguistic utterances). Both applications of discourse processing involve planning and plan recognition, linguistic principles, real-world knowledge, the virtual model of the world, and interaction management. The robotics domain requires a limited command vocabulary size of some 1500 words initially, and utterances are comparatively straightforward. This will also improve the recognition rate of the speech engine and support more diverse interaction environments. To begin with, the robot will understand imperative utterances, but other types of comprehension capabilities, as well as language generation, will be incrementally added.

Using dialogue processing in the human/robot interface allows, but also requires, the robot to maintain a model of the world and to maintain a record of the dialogue. Without a discourse/dialogue component, utterances would be difficult to connect to the robot's environment.

ADAPT implements a discourse recipe-based model (DRM) for dialogue comprehension and generation. It learns the discourse recipes, which are generalizations of an agent's discourse plans, as a side effect of dialogue planning. This way, plans can be used for comprehension and generation. If no recipe can be matched, the system resorts to dialogue plans. This allows both a top-down and bottom-up approach to dialogue modeling. It also supports elements of BDI/DME functionality such as maintaining a common ground with information about shared background knowledge and a conversational record.

Initiative is an important aspect in dialogue. Different approaches to managing dialogue vary from system-initiative (where the robot controls interaction) to user-initiative (where the human controls interaction) to, ideally, mixed or joint-initiative (where the robot and the human take turns controlling and relinquishing control as situations unfold). A highly reactive robot requires mixed initiative. Part of the work in this project will involve investigating and demonstrating the relative advantages and disadvantages of BDI vs. DRM approaches for supporting (successively) human-, system-, and mixed-initiative robotic interactions.

4.1: Using Visual Schemas for Semantics

The previous section explains the overall organization of the language system. Let us examine in more detail how the semantics are handled.

The central use of this world model is to enable the robot to "see" what utterances might mean, and thus to help select appropriate semantics from among numerous possibilities. Langacker's Cognitive Grammar [12, 13, 14] provides a well-founded integration of grammar and semantics with imagery, using spatial primitives to give semantics for many common actions and relationships. His grammar provides a mechanism for reasoning about linguistic composition by superimposition of images. For example, Figure 3 shows image schemas for "walk" and "John" and "snake". Given the sentence, "John walks", the schema for "walk" can be completely assigned to "John". But when given the sentence, "A snake walks", the schema for "walk" cannot be completely assigned to the schema for "snake". In this way, the system can figure out that the first sentence makes sense and the second one doesn't.

Holmqvist [9, 10] partially implemented this grammar, but no complete implementation yet exists. We are currently implementing Langacker's Grammar, and extending his grammar by animating these schemas so that the "walk" schema will not be a static picture of legs, but rather a working model of virtual legs.

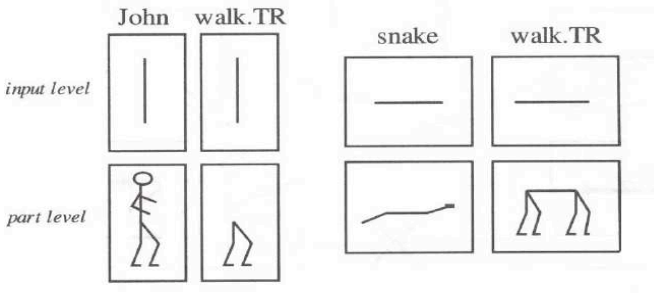

Fig. 3. Image schemas for "John walks" and "A snake walks"

In this way, perceptual patterns from the vision system are used not only to guide motion, but also to guide ADAPT's search among alternative semantics for utterances, both the system's own and those it hears. This is an illustration of the deeper integration of perception, language and action in ADAPT.

Let's examine how ADAPT uses visualization to understand a simple navigational primitive: the term "near". In Figure 4 we see a screenshot of ADAPT's virtual world.

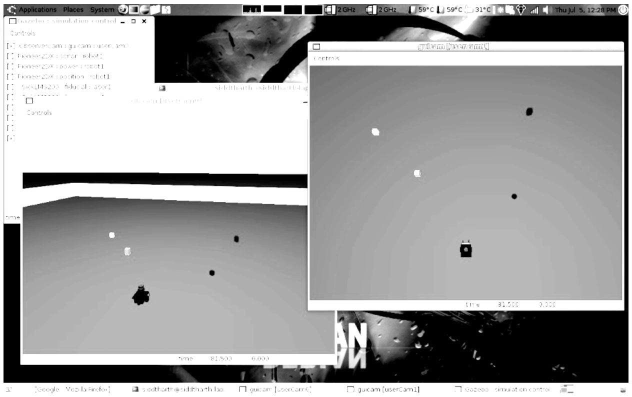

Figure 4. Screenshot from ADAPT.

There are two principal windows open on the screen. The left one shows ADAPT's virtual world. We see the virtual copy of the robot itself, and four blocks: one white, one yellow, and two red blocks, one small and one large (if you are reading a black-and-white copy of this paper then these blocks are listed left-to-right). The right window shows the same scene viewed from a virtual camera suspended directly above the robot. This virtual camera moves with the robot as it moves, and shows the robot's current visual context. In the situation shown, the robot's task is to maneuver among the blocks, and thus the proper visual context is a region of the environment that contains all the blocks.

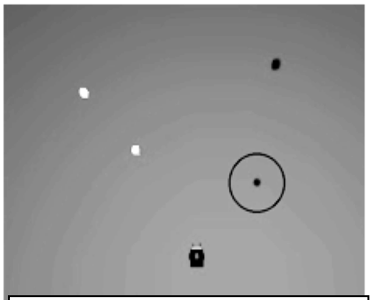

Figure 5. The neighborhood defining "near the small red block".

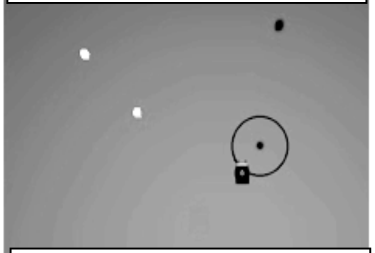

Figure 6. The robot is near the small red block.

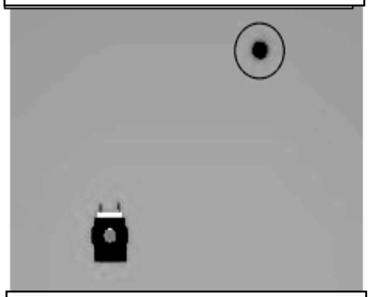

Figure 7. The visual context for picking up the small red block.

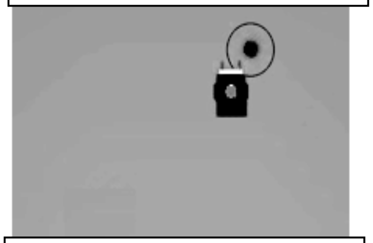

Figure 8. The robot is near the small red block in the context of picking it up.

The visual context is the central construct used to determine the meanings of words that have physical meanings. The same word may have many different meanings in different situations, e.g. the word "near" may mean "within a few feet" for a motion task but may mean "within a foot" for a grasping task. The typical approach to semantics has been to store all meanings so that they can used in appropriate situations. This requires the robot to learn a large number of meanings for each word and to classify situations to be able to apply the correct meaning. Such an approach makes it difficult to understand the same word in new situations, e.g. "It is near lunchtime." or "The student is near graduation." Also, this approach faces the difficulty of enforcing semantic agreement among many different words; a notion of context is needed. Note that WordNet, which is a widely used lexical resource used in computational semantics, does not encode the meanings of prepositions such as "near". Cognitive grammars have become especially important in representing the meaning of such functional items in language, an area where symbolic denotational semantics has been weak.

ADAPT's approach is to use a *single meaning* in as many situations as possible, and to change the visual context according to the current task and goals of the robot. Rather than being encoded in an "a priori" arbitrary list of symbolic senses that then has to be consulted whenever a word must be disambiguated, the semantics of a word is defined by a fixed visual construct whose effect changes as the visual context changes. The visual context consists of the view from a virtual camera above the robot, seen in a fixed window. This means that the amount of the world that is visible changes as the virtual camera zooms in and out.

The following example illustrates this process. The semantics for the word "near" is defined by a visual neighborhood of a *fixed distance* from the given object. In Figure 5, we see such a neighborhood depicted by a black circle around the small red block. This defines the meaning of "near the small red block".

If the robot is told to go near the small red block, it will create this neighborhood in its virtual world and plan a motion that will take it anywhere inside the circle. In Figure 6, we see that the robot has accomplished this, so the value of "near the small red block" is true.

Then we tell the robot to pick up the small red block. This is a new task, and the context shifts: it no longer includes all the blocks, but only the small red one. This causes a shift of focus to the region immediately around the small red block. The visual context zooms in to magnify the region around the small red block, as shown in Figure 7. The task of picking up the block requires the robot to be near the block, but the meaning of "near" is now different, because the robot must be much closer to grasp the block than it must be to see it or navigate around it. In Figure 7, we see that the same black circle is around the small red block; however, it no longer denotes the same region of the world but rather a much smaller one, and the robot is no longer seen as near the small red block. Given the task of picking up the block, the robot must now plan motions to take it within the black circle. Figure 8 shows the situation after this has been done.

At this point the robot can begin the special small maneuvers required to pick up an object.

5: Current Work and Summary

We are currently working on two applications that require perception, planning and interaction with humans using natural language. The first is serving as a tour guide for people who wish to tour our lab facilities. This is an application that has been successfully used elsewhere, and serves as a good starting point. To make this task more dynamic, and to test learning, we relocate objects in the lab environment so that the robot must locate them and adjust its behavior and speech to incorporate the changes.

The second class of tasks simulates a team performing an assembly task. The robot and a human cooperate in pushing boxes of various sizes around and stacking them to create a desired configuration of boxes. For example, the goal may be to sort the boxes in piles according to their size. The robot and the human must perform this task with minimal interference between them. This requires the use of language to communicate, and also requires the robot to model the human's actions to avoid interference. This class of tasks includes the full range of problems for the robot to solve, from abstract task planning to real-time scheduling of motions, and including perception, navigation, communication with humans and grasping of objects. In addition, the robot must learn how to push one or more objects properly. This range of demands is ideal for our purposes, because it creates a situation in which complex hierarchies of features and constraints arise.

To this point, most of the work has been on basic implementation. Getting all the software components to talk nicely to each other has been very hard. We have completed this implementation. ADAPT's inference engine can create basic entities in the virtual world in real time based on its vision data and update them to reflect new percepts as the robot moves. This has been implemented for a very small hand-coded library of known objects.

Also, we have demonstrated successfully that ADAPT can listen to a person, generate an appropriate response using a discourse model, and speak the response. The discourse models are also constructed by hand. Basic image schemas have been implemented by hand and used to provide semantics for simple navigational concepts. Our robots can maneuver successfully using visualization to determine the semantics of "near", "around", and "far", and to follow spoken commands using these terms.

We are currently expanding the libraries of discourse models, visual schemas, and virtual objects that can be modeled.

6: References

[1] Benjamin, D. Paul, Damian Lyons and Deryle Lonsdale, "Embodying a Cognitive Model in a Mobile Robot", Proceedings of the SPIE Conference on Intelligent Robots and Computer Vision, Boston, October, 2006.

[2] Benjamin, D. Paul, Damian Lyons and Thomas Achtemichuk, "Obstacle Avoidance using Predictive Vision based on a Dynamic 3D World Model", Proceedings of the SPIE Conference on Intelligent Robots and Computer Vision, Boston, October, 2006.

[3] Benjamin, D. Paul, Damian Lyons and Deryle Lonsdale, "Designing a Robot Cognitive Architecture with Concurrency and Active Perception", Proceedings of the AAAI Fall Symposium on the Intersection of Cognitive Science and Robotics, Washington, D.C., October, 2004.

[4] Benjamin, D. Paul, Damian Lyons and Deryle Lonsdale, " Cognitive Robots: Integrating Perception, Action and Problem Solving in Behavior-Based Robots", AAMAS-2004 Proceedings, pp. 1308-1309, 2004.

[5] Daniel C. Dennett, The Role of Language in Intelligence, in What is Intelligence?, The Darwin College Lectures, ed. J. Khalfa, Cambridge Univ. Press. 1994.

[6] Gärdenfors, Peter. 1995. Meanings as Conceptual Structures. Lund University Cognitive Studies 40.

[7] Green, Rebecca. "Internally-Structured Conceptual Models in Cognitive Semantics," in The Semantics of Relationships - An Interdisciplinary Perspective, R. Green, C. Bean, and S. Myaeng, Eds. Dordrecht, The Netherlands: Kluwer, 2002, pp. 73-89.

[8] Holmqvist, Kenneth, Dimensions of Cognition. In: Spinning Ideas, Electronic Essays: Dedicated to Peter Gärdenfors on His Fiftieth Birthday. (not dated): http://www.lucs.lu.se/spinning/categories/cognitive/Holmqvist/kenneth.pdf.

[9] Holmqvist, K., "Implementing Cognitive Semantics", Lund: Department of Cognitive Science, 1993.

[10] Holmqvist, K., "Conceptual Engineering", in Cognitive Semantics: Meaning and Cognition, Allwood and Gardenfors (Eds.), John Benjamins, pp.153-171, 1999.

[11] Jackendoff, Ray (1996). Semantics and Cognition. In: The Handbook of Contemporary Semantic Theory (Shalom Lappin, Ed.). Blackwell Publishers. pp. 539-559.

[12] Langacker, R., "Foundations of Cognitive Grammar, Vol. I", Stanford University Press, 1987.

[13] Langacker, R., "Foundations of Cognitive Grammar, Vol. II", Stanford University Press, 1991.

[14] Langacker, R., "Concept, Image, and Symbol", Berlin, New York, Mouton de Gruyter, 1991.

[15] M. van Dartel & E. Postma, Symbol manipulation by internal simulation of perception and behaviour, in Berthouze, L., et al. (Eds.), Proceedings of the Fifth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems, Lund University Cognitive Studies, 123, 2005.

[16] Hendrik Zender, P. Jensfelt, O. M. Mozos, Geert-Jan M. Kruijff and W. Burgard, An Integrated Robotic System for Spatial Understanding and Situated Interaction in Indoor Environments, Proc. AAAI07, 2007.

[17] J. Zhang, Colloni and Knoll, Interactive Assembly by a Two-Arm Robot Agent, Robotics and Autonomous Systems, Elsevier, 1999.

[18] Y. Zhang and A. K. Mackworth, "Modeling behavioral dynamics in discrete robotic systems with logical concurrent objects," in *Robotics and Flexible Manufacturing Systems* (S. G. Tzafestas and J. C. Gentina, eds.), pp. 187–196, Elsevier Science, 1992.