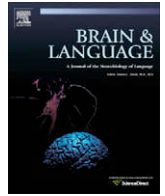




Contents lists available at ScienceDirect

Brain & Language

journal homepage: www.elsevier.com/locate/b&l

A cognitive neuroscience perspective on embodied language for human–robot cooperation

Carol Madden^{a,b}, Michel Hoen^a, Peter Ford Dominey^{a,*}

^aStem Cell and Brain Research Institute, INSERM U846, 69676 Bron Cedex, France

^bPsychology Department, T12-37, Erasmus University Rotterdam, Post bus 1738, 3000 DR Rotterdam, The Netherlands

ARTICLE INFO

Article history:

Accepted 3 July 2009

Available online xxx

Keywords:

Basal ganglia

Cortico-striatal system

Human–robot cooperation

Grammatical construction

ABSTRACT

This article addresses issues in embodied sentence processing from a “cognitive neural systems” approach that combines analysis of the behavior in question, analysis of the known neurophysiological bases of this behavior, and the synthesis of a neuro-computational model of embodied sentence processing that can be applied to and tested in the context of human–robot cooperative interaction. We propose a Hybrid Comprehension Model that links compact propositional representations of sentences and discourse with their temporal unfolding in situated simulations, under the control of grammar. The starting point is a model of grammatical construction processing which specifies the neural mechanisms by which language is a structured inventory of mappings from sentence to meaning. This model is then “embodied” in a perceptual-motor system (robot) which allows it access to sentence-perceptual representation pairs, and interaction with the world providing the basis for language acquisition.

We then introduce a “simulation” capability, such that the robot has an internal representation of its interaction with the world. The control of this simulator and the associated representations present a number of interesting “neuro-technical” issues. First, the “simulator” has been liberated from real-time. It can run without being connected to current sensory motor experience. Second, “simulations” appear to be represented at different levels of detail. Our paper provides a framework for beginning to address the questions: how does language and its grammar control these aspects of simulation, what are the neurophysiological bases, and how can this be demonstrated in an artificial yet embodied cognitive system.

© 2009 Elsevier Inc. All rights reserved.

1. Introduction

In recent decades, the field of cognitive psychology has seen a shift away from symbolic, amodal cognitive systems and towards embodied theories of cognition. While traditional views (Fodor, 1983; Pylyshyn, 1986) assumed a conceptual system that processed symbolic units that were arbitrarily related to their referents (e.g., propositional representational systems, see Kintsch, 1998), embodied theories have brought into the fore the existence of an analogical correspondence between symbols and referents and have pushed to explicitly integrate the body and its perceptual-motor systems into conceptual models (Barsalou, 1999; Damasio, 1989; Glenberg, 1997). According to this new embodied framework, perception and cognition are linked, as conceptual representations are *situated simulations* that are instantiated in the same systems that are used for perception and action. Indeed, recent studies of the human neurophysiology of embodied sentence processing have demonstrated that mem-

ory, planning, and language comprehension partially reactivate traces of experience distributed across multiple perceptual and motor modalities in the brain (Glenberg & Kaschak, 2002; Hauk, Johnsrude, & Pulvermüller, 2004; Zwaan & Madden, 2005). Related embodied approaches have had a significant influence on artificial intelligence and robotics (e.g. Anderson, 2003; Dominey, 2005; Gorniak & Roy, 2007; Lungarella, Metta, Pfeifer, & Sandini, 2003; Mavridis & Roy, 2006; Pfeifer & Gómez, 2005; Roy, Hsiao, & Mavridis, 2004; Vernon et al., 2007).

In this context, the objective of this position paper is to demonstrate how principles from embodied cognition and neuroscience have been incorporated in the building of an artificial cognitive system, with the idea that a “cognitive neural systems” approach can contribute to the understanding of embodied language processing, both in artificial systems research, and in language research on humans. The cognitive neural systems framework involves a methodical approach, with three interrelated activities by which one should: (1) characterize the behavior in question, (2) identify the neural mechanisms that implement that behavior, (3) build a system whose architecture is derived from the neurophysiology, and that can produce the behavior. The underlying idea is that the interrelations between

* Corresponding author. Fax: +33 43 791 1210.

E-mail addresses: carol.madden@inserm.fr (C. Madden), michel.hoen@inserm.fr (M. Hoen), peter.dominey@inserm.fr (P.F. Dominey).

these three activities will produce both insights and new questions that would not arise otherwise.

Part of the outcome of this methodology has been our proposal of a hybrid system in which compact representations of actions, action sequences, coordinated action sequences and shared plans, etc., are stored in a propositional format as “situation constructions”. When necessary, they are “expanded” in situated simulations, providing the listener with the desired level of detail. In this context we hold that language allows the speaker to “direct the film”, to precisely control the initiation, unfolding and termination of appropriate simulations in the mind of the listener, through precise grammatical mechanisms that have evolved for this purpose.

This work begins with an overview of our model of grammatical construction processing and its neurophysiological underpinnings in Section 2. Section 3 introduces the first level of embodiment of the model within a sensory-motor robotic system. In Section 4, the second level of embodiment – the integral use of situated simulations – and its neurophysiological bases are introduced, along with motivation for integration of this capability into the robotic system. Section 5 identifies effects wherein grammar acts as a set of cues to direct simulations in humans. These effects are discussed in terms of their implementation within the robot model. Finally, Section 6 offers conclusions and future directions based on the preceding sections.

2. Grammatical construction processing

Across human languages, the grammatical structure of sentences is specified by a combination of cues including word order, grammatical function words (and or grammatical markers attached to the word roots) and prosodic structure (Bates, McNew, MacWhinney, Devescovi, & Smith, 1982; MacWhinney, 1982). These grammatical structures are thought to code the mappings from surface structures of utterances to their underlying meanings (Goldberg, 1995). Thus, the goal of any construction-based language comprehension system (artificial or natural) is to (1) identify grammatical constructions and (2) correctly map these constructions to their corresponding meanings.

2.1. A model for processing grammatical constructions

Fig. 1 depicts the process by which our language model identifies grammatical structures and maps them onto meaning (from Dominey, Hoen, & Inui, 2006). As the sentence is processed word by word, a lexical categorization process identifies open class words (unlimited set of words that carry the semantic content) and closed class words (limited set of words that carry the grammatical structure, e.g., determiners, prepositions, auxiliary verbs). This is a basic and viable feature of the language system, as this type of lexical categorization has been demonstrated in neural net-

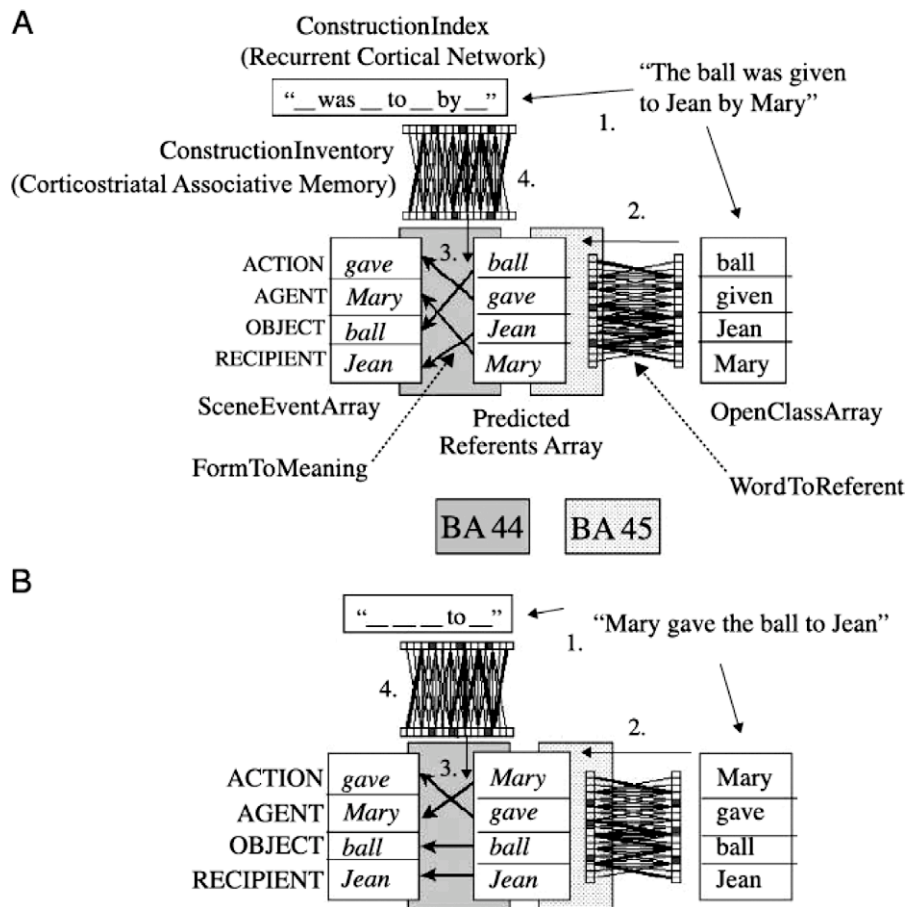


Fig. 1. (From Dominey et al., 2006) Schematic Structure-Mapping Architecture. (A) Passive Sentence Processing: Step 1. Lexical categorization – Open and closed class words Open Class Array and ConstructionIndex. Step 2. Open class words in Open Class Array are translated to their referent meanings via the WordToReferent mapping. Insertion of this referent semantic content into the Predicted Referents Array (PRA) is realized in pars triangularis BA45. Step 3. PRA elements are mapped onto their roles in the SceneEventArray by the FormToMeaning mapping, specific to each sentence type. Step 4. This mapping is retrieved from ConstructionInventory (a cortico-striatal associative memory), via the ConstructionIndex (a cortico-cortical recurrent network) that encodes the closed and open class word patterns that characterize each grammatical construction type. The structure mapping process is associated with activation of pars opercularis BA44. In the current implementation, neural network associative memory for the ConstructionInventory is replaced by a procedural look-up table. (B) Active sentence. Note difference in ConstructionIndex, and in FormToMeaning.

work studies (Blanc, Dodane, & Dominey, 2003; Shi, Werker, & Morgan, 1999), and even newborns have the ability to perform this categorization (Shi et al., 1999). The current implementation recognizes only nouns and verbs as open class words, with adjectives and adverbs to be addressed in future implementations. The open class word meanings are retrieved from the lexicon and stored in a working memory called the PredictedReferentsArray. Next, during thematic role assignment, these referent meanings are mapped onto the appropriate components of the meaning structure. In Fig. 1, this corresponds to the mapping from the PredictedReferentsArray onto the meaning coded in the SceneEventArray.

During the lexical categorization process, the closed class words are explicitly represented in the ConstructionIndex, which is a global representation of the sentence structure, re-coded from the local structure of the open and closed class words in the utterance. The requirement is that every different grammatical construction type should yield a unique ConstructionIndex, corresponding to the cue ensemble of Bates et al. (1982). This ConstructionIndex can then be used as an index into an associative memory to store and retrieve the correct FormToMeaning mapping. As seen in parts A and B of Fig. 1, this mapping varies depending on the grammatical construction type, emphasizing the importance of the model's ability to store and retrieve different FormToMeaning mappings for different sentence types. In this way, thematic roles for the content words can be correctly determined by their relative position in the sentences with respect to the other content words and the function words.

2.2. Proposed neurological underpinnings

We argue that the cortico-striato-thalamo-cortical system (CSTC) provides the building blocks for constructing such a machine (Dominey, Inui, & Hoen, 2009), based on its inherent capabilities for sequential structure coding (Dominey, 1995) and abstract structure mapping (Dominey, Lelekov, Ventre-Dominey, & Jeannerod, 1998). The formation of the ConstructionIndex as a neural pattern of activity relies on sequence processing of closed class elements in recurrent cortical networks in BA47. The retrieval of the corresponding FormToMeaning component relies on a corticostriatal associative memory (Calabresi, Picconi, Tozzi, & Di Filippo, 2007) that links cortical representations of the ConstructionIndex with the appropriate mapping in striatum (Dominey, Hoen, Blanc, & Lelekov-Boissard, 2003; Dominey et al., 2006), allowing the assignment of lexical open class elements into the global meaning representation. In particular we have proposed that BA45 encodes the current open class element, while a working memory in BA45/47 maintains the sentence level ensemble of open class words available for possible reanalysis and repair (Dominey et al., 2009). Implementation of the mapping of these open class elements onto their thematic roles in the meaning will take place in the frontal cortical region including BA44, 46 and 6, corresponding to the SceneEventArray. Indeed, empirical investigations have localized the representation of event meaning in the pars opercularis and the fronto-parietal action system, both while visually observing events (Buccino et al., 2004), and listening to event descriptions (Tettamanti et al., 2005).

Because the frontal cortical region (including BA44/46/6) has been shown to play a role in both linguistic (grammatical) and non-linguistic structural mapping (Hoen, Pachot-Clouard, Segebarth, & Dominey, 2006), we speculate that meaning remains at a fairly abstract level at this point, corresponding to the propositional pole of the Hybrid (propositional-embodied) Comprehension Model. The filling of the PredictedReferentsArray, however, corresponds to a more language-specific ventral stream mechanism, culminating in the pars triangularis (BA45) of the ventral premotor area (see the declarative component of Ullman's 2004

model), as this area was activated by sentence processing but not non-linguistic sequencing (Hoen et al., 2006). The link with embodied meaning components is explained in Section 4.

3. Grammatical constructions for perception and action

We demonstrated that when provided with sentence-meaning pairs the model depicted in Fig. 1 could learn extended sets of constructions in English, French and Japanese (Dominey et al., 2006). These constructions could be used to understand new sentences not used in training, including extension to new grammatical forms. Up to this point, the training corpora were generated by hand. The success of the model suggested that it was ready to “go out into the world”, that is, to learn language based on its proper exposure to and interaction with the world.

The first step in this direction was to embody the model into a robot with perceptual inputs. In this setting, humans would perform actions that were seen by the robot, and would simultaneously narrate their actions with spoken sentences that were heard by the robot. This required providing the artificial cognitive system with audition in order to hear spoken language, and vision in order to see events. Audition was provided by “off the shelf” speech recognition software (IBM ViaVoice™). Low level vision was provided by a commercial system based on color segmentation (PanLab SmartVision). The vision system tracked the motion of objects in near real-time and provided the timing of contact events. Physical contact is a perceptual primitive that is so basic to events that even infants can reason about it (Kotovsky & Baillargeon, 1998). We, therefore, capitalized on the idea that events including touch, push, take, take-from, and give, can all be characterized in terms of patterns of physical contact, illustrated in Fig. 2A. For example, push is a contact between agent and object with a displacement. Touch is related, but with no displacement. Give involves multiple contacts: the agent contacts the object, the object contacts the recipient, and then the contact between agent and object is terminated. The generic contact sequence for a set of event types is illustrated in Fig. 2A, and the physical “event scenario” set-up is illustrated in Fig. 2B. Developmental issues in this context are addressed by Dominey and Boucher (2005). We thus implemented an embodied artificial cognitive system (robot) that could perceive physical events related to object manipulation based on a simplification of principal of force dynamics including contact (Siskind, 1996; Talmy, 1988). Importantly, the meaning of words and sentences was defined in terms of perceptual and physical properties of objects and their interactions. The simplifying assumption here is that the actions and language that the system is exposed to are constrained to these forms of discrete, concrete actions.

In this context, naïve human users performed actions with three objects, and narrated their actions, e.g. “The block gave the moon to the cylinder.” Based on the resulting corpora, the robot could learn the lexicon (corresponding to the limited number of actions and objects), and a diverse variety of grammatical constructions. Once the inventory of grammatical constructions had been acquired, it could be used both to understand new sentences (i.e. to generate the corresponding meaning), as well as to narrate new actions (i.e. to generate the corresponding sentence). It was of particular interest that grammatical focus could be used in order to select the appropriate construction that would allow the agent, object or recipient to be at the head of the sentence, depending on the discourse context (Dominey & Boucher, 2005).

Testing naïve users demonstrated the ability of the system to handle reference and perceptual errors. Errors could be due to speech transcription and perceptual mistakes (14%), and to errors in assigning the correct verb, as for example “push” and “touch”

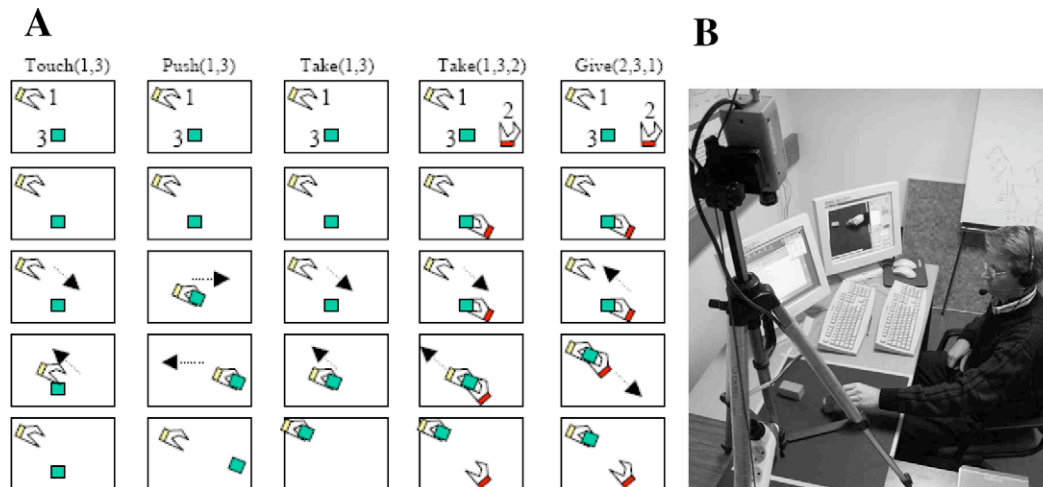


Fig. 2. (Modified from Dominey & Boucher, 2005) (A) Event recognition profiles. Each column corresponds to a sequence of “contact events” which can be used to characterize the specified action (touch, push, take and give), independent of many parameters including velocity, approach direction, etc. (B) The physical set-up with the robot that can learn language based on observing narrated actions performed by the human. Vision processing identifies objects, corresponding to 1–3 in event recognition profiles, to perform on-line event “meaning” extraction. User narrates her actions to thus provide < sentence, meaning > pairs for language learning. After learning, the system can narrate new observed events.

were confused (15%). Despite these errors in the training and testing input, the statistical regularity of the data set allowed the system to learn the grammatical constructions and perform thematic role assignment at 85% correct (Dominey & Boucher 2005).

The next step towards embodiment corresponded to the transition from perception to action. This implied providing the robot with the ability to act, which we developed using several platforms, two of which will be focused here. From the perspective of language and grammatical constructions, the richer actions are those corresponding to predicates that take several arguments. Thus, “Give me the rose leg” is more interesting than “reach”, in the sense that it corresponds to a richer construction, and likewise to a more specific action, because it constrains the set of objects to those that can be “given”. Fig. 3 illustrates two robot platforms that have been used in the “embodiment” of the sentence processing system. In both cases the robotic systems are equipped with vision processing, and different issues are addressed. In the HRP2 humanoid robot (Fig. 3A) we explored such relations between “verb island” constructions (i.e. constructions in which the verb is a fixed part of the construction, see Tomasello, 2003), and corresponding robot behaviors. There we demonstrated how, once the robot learned the link between a command such as “Give me the rose leg”, it could then generalize this command to all graspable objects that it was capable of seeing (Dominey, Mallet, & Yoshida, 2007a – see video demonstration in Supplementary material).

In this context, one issue that is of particular interest is the development of cooperation. Warneken, Chen, and Tomasello (2006) have demonstrated that 18–24 months old children are capable of observing two adults participate in a cooperative task (game), and then immediately taking on the role of either of the participants, as if they had acquired a “bird’s eye view” of the interaction. This implies that the children are capable of constructing a representation of the shared plan, the interacting actions, that comprise the cooperative task. This in turn suggests quite interesting representations: shared plans that indicate the distinct actions of different “players” and the coordination between these actions.

Fig. 3C represents the architecture that we developed in this context. A central and key feature of this architecture is the concept that action representation is in a common format, independent of whether that representation derives from vision of that action, from a spoken language command requesting that action, whether it is being used by the robot to describe an action or to

produce an action. It is noteworthy, though we will not go into the details here, that this is consistent with a number of findings in humans demonstrating such a common representation for actions in Brodmann’s area BA44 (see Buccino et al., 2004).

We thus demonstrated that indeed such a representation could allow the robot in Fig. 3B to describe an observed action, imitate that action, receive action commands via spoken language and execute those actions all based on a common representation of action (see video demonstration in Supplementary material). We also implemented an action sequence learning capability. As illustrated in Fig. 3 the cooperative sequencing made the distinction between who did what, and thus allows the robot to help the human user, and to actually switch roles with the user (Dominey & Warneken, in press).

4. Introduction of the simulator

Up until this point, our robot’s meaning representation has been an impoverished predicate-argument format. Thus, while we can interrogate the robot about the order of events, the robot does not use its own body configuration during the execution of a particular action to better understand the event or predict upcoming events, for example. This is thus inconsistent with results from the field of embodied cognition, which proposes that world knowledge is encoded in “embodied simulations” of our experiences. The following subsection reviews this growing body of evidence for the idea that the multimodal neurophysiological processes that are involved in perception and action in the environment are themselves reactivated in the later mental representation of related concepts (Barsalou, 1999; Damasio, 1989; Glenberg, 1997). Following this review of evidence, we describe how such simulations can be implemented in our robot model.

4.1. Neurophysiological evidence for simulations

Even single word comprehension relies on embodied representations. That is, the representation of semantic content associated with single words is realized by activating, at least partially, the same neural networks that are engaged in the sensorimotor representation of the real-world objects or actions symbolized by these words. Initial experimental evidence supporting this claim was ob-

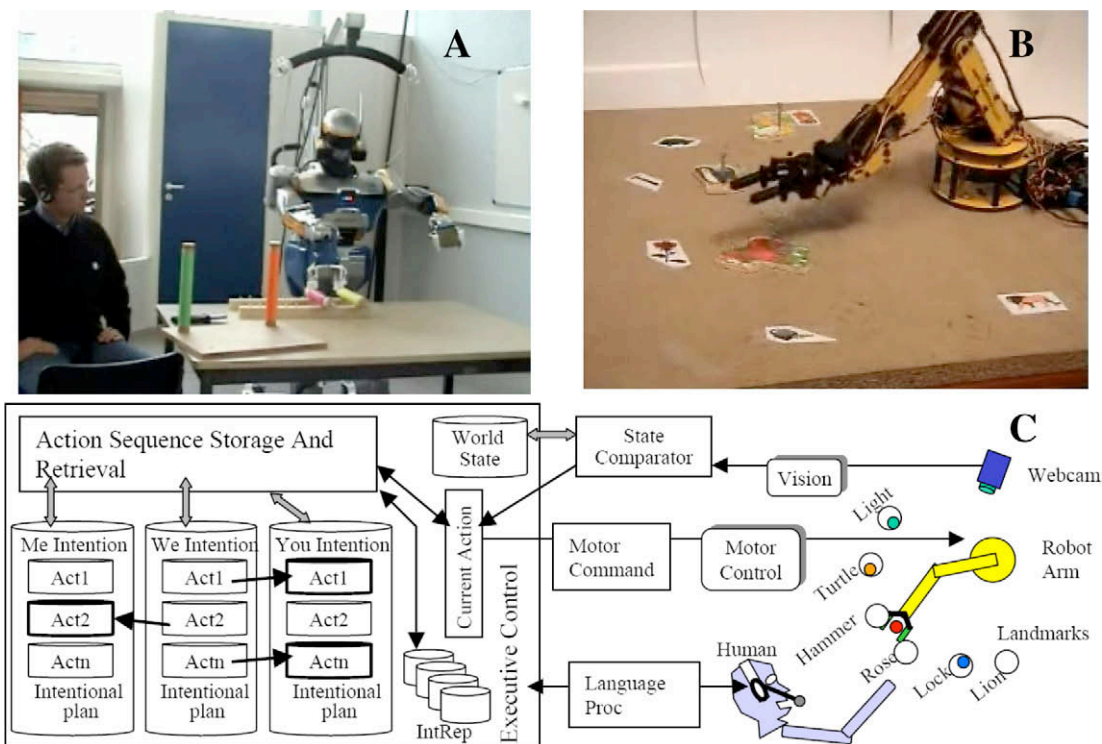


Fig. 3. (A) Human–robot cooperation (Dominey, Mallet, & Yoshida, 2007b). (B and C) Cooperator robot and control architecture (Dominey and Warneken, in press). In a shared work-space, human and robot manipulate objects (green, yellow, red and blue circles corresponding to dog, horse, pig and duck), placing them next to the fixed landmarks (light, turtle, hammer, etc.). *Action*: Spoken commands interpreted as individual words or grammatical constructions, and the command and possible arguments are extracted using grammatical constructions in Language Proc. The resulting Action (Agent, Object, Recipient) representation is the Current Action. This is converted into robot command primitives (Motor Command) and joint angles (Motor Control) for the robot. *Perception*: Vision provides object location input, allowing action to be perceived as changes in World State (State Comparator). Resulting Current Action used for action description, imitation, and cooperative action sequences. *Imitation*: The user performed action is perceived and encoded in Current Action, which is then used to control the robot under the supervision of Executive Control. *Cooperative Games*. During observations, individual actions are perceived, and attributed to the agent or the other player (Me or You). The action sequence is stored in the We Intention structure, that can then be used to separately represent self vs. other actions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

tained from lesion studies showing that patients with a cortical lesion covering the motor and premotor areas experienced major difficulties producing action-verbs specifically, compared to nouns. This observation suggested the existence of a shared representation between the action-execution system, located in motor areas and the semantic representation of the action, activated by the processing of the corresponding action-word (Damasio & Tranel, 1993). Later, pioneering neuroimaging studies on specific word-category activation-patterns confirmed these former neuropsychological observations, action-words being associated with more significant cortical activity in motor or premotor areas compared to words depicting colors (Martin, Haxby, Lalonde, Wiggs, & Ungerleider, 1995) or persons or animals (Grabowski, Damasio, & Damasio, 1998).

More recent neuroimaging studies confirmed the existence of a clear overlap in the cortical activation patterns associated with the processing of words related to action knowledge and the cortical motor areas implicated in these actions (Assmus, Giessing, Weiss, & Fink, 2007; Chao & Martin, 2000; Noppeney, Josephs, Kiebel, Friston, & Price, 2005). Hauk et al. (2004) even showed that these activation patterns followed the natural somatotopic organization of sensory-motor cortical areas, the representation of words related to actions from specific body-parts such as the hand, leg or mouth, being associated with activations in the somatotopic areas corresponding to the hand, leg or mouth representation, respectively. Tettamanti et al. (2005) extended this approach to simple action sentence processing, observing similar topographic activation in sensory-motor cortical areas.

Subsequent fMRI studies have shown that the link between the cortical representation of knowledge and sensory systems could be extended beyond the example of action planning and execution, demonstrating that this was also the case for all other sensory modalities. For example, Goldberg, Perfetti, and Schneider (2006) asked participants to determine if concrete words possessed a certain sensory property along various modalities (gustatory, tactile, auditory or visual). The activation of each knowledge modality was associated with activation patterns corresponding to the brain loci of each sensory system. Goldberg, Perfetti, Fiez, and Schneider (2007) presented volunteers with animal names and asked them about abstract ('has bones') or sensory ('has fur') knowledge about these animals. Results from this study showed that abstract knowledge retrieval compared to sensory knowledge is associated with increased activity in cortical areas of the left prefrontal cortex. This growing body of results demonstrates the modality-specific relation between knowledge and sensory-motor systems.

Another very interesting observation in this context was that words referring to manipulable- compared to non-manipulable-objects were associated to increased activity in motor areas, as if the representation of a manipulable tool (a hammer) required or at least was often associated to the mental simulation of the corresponding manipulation (using the hammer with the hand) (Martin, Wiggs, Ungerleider, & Haxby, 1996). This observation suggests that beyond the direct and somatotopic overlap between single action-words and real-action representations, even single nouns can trigger representations as complex simulation of complete events implicating these objects. These nouns are able to trigger embodied

representation of the larger situations in which they occur, which in turn implicates the action related to the usage of this object. Extending this observation to general semantic representation suggests that single words can trigger general, yet complex event simulations that may then be further constrained by surrounding context such as neighboring words and syntactic constructions in which they may be embedded. Therefore, simulation in sentence comprehension can rely on the same architecture used for the comprehension of single words using the constraints and specifications of syntax to constrain event simulations as described above. This issue of grammatical constraint on simulations will be further addressed in Section 5, but first we now turn to the implementation of a simulator in our robot model.

4.2. Toward situated simulations in the robot model

In the proposed system, amodal/propositional referents in the SceneEventArray are linked to modality-specific sensorimotor “simulations” encoding visual, tactile, and auditory inputs (Mavridis & Roy 2006), as well as motor outputs. Grammatical constructions (closed class words in the ConstructionIndex) will mediate the links between objects/actions/relations in the SceneEventArray and their referents within sensorimotor simulations.

While the current implementation of the model codes for very few situational aspects, any feature that is potentially codable within the ConstructionIndex and OpenClassArray can be selectively linked to sensorimotor simulations. For example, once verb tense is coded in the ConstructionIndex, “will push” would activate the set of sensorimotor pushing simulations in which the onset has not yet occurred, whereas “has pushed” would activate the resultant states of these same sensorimotor pushing simulations (see Section 5). These mappings would be learned through principles of cross-situational statistics (Siskind 1996, and described in Dominey & Boucher, 2005, Section 3.2), capitalizing on the idea that the future tense more often co-occurs with experiences in which the onset of the action has not yet occurred. Our challenge is to identify

the tense, mode and aspect features that should be included, and to establish their encoding in the SceneEventArray, and the corresponding simulations.

Preliminary versions of all components of Perception/Action and Predicate-argument components (Fig. 5) exist. The novelty here is the linking of these elements to a conceptual system based on a situated simulation model. With the resulting hybrid system, the robot can run simulations to understand and monitor ongoing events, predict future events, test plans, understand past events, and negotiate “how we can do this” with the human user. The result is a compact propositional representation of cooperative events which can then be expanded to the required level of detail through use of the simulator.

The current model was originally designed to handle thematic role assignment, and it should be stated that it currently exhibits several limitations. The vocabulary size is restricted to 25 elements, because of the “one bit per word” coding scheme. Our 25 binary element vector allows a 2^{25} vocabulary size, and scaling will rely on the quality of the associative WordToReferent memory. The ConstructionInventory has been demonstrated with a number of English, French and Japanese constructions, and its scalability is linked to the viability of the cue competition model (Bates et al., 1982; MacWhinney, 1982). With the augmentation of vocabulary and the construction inventory and the coding of additional features of language, further instantiations of the model promise to scale up quite nicely. Already, the model is robust to a certain degree of noise in the input (e.g., incorrect assignment of open and closed class words) as well as ambiguous referents in a scene. The model initially deals with such ambiguous references by using cross-situational statistics over multiple scene/narration pairs. Then, once grammatical mappings are learned, syntactic bootstrapping is used to disambiguate novel referents within a known grammatical construction (see Dominey & Boucher 2005). As the ConstructionIndex is expanded to accommodate a larger set of grammatical constructions and morphology, more situational features will be represented in the SceneEventArray, activating a more

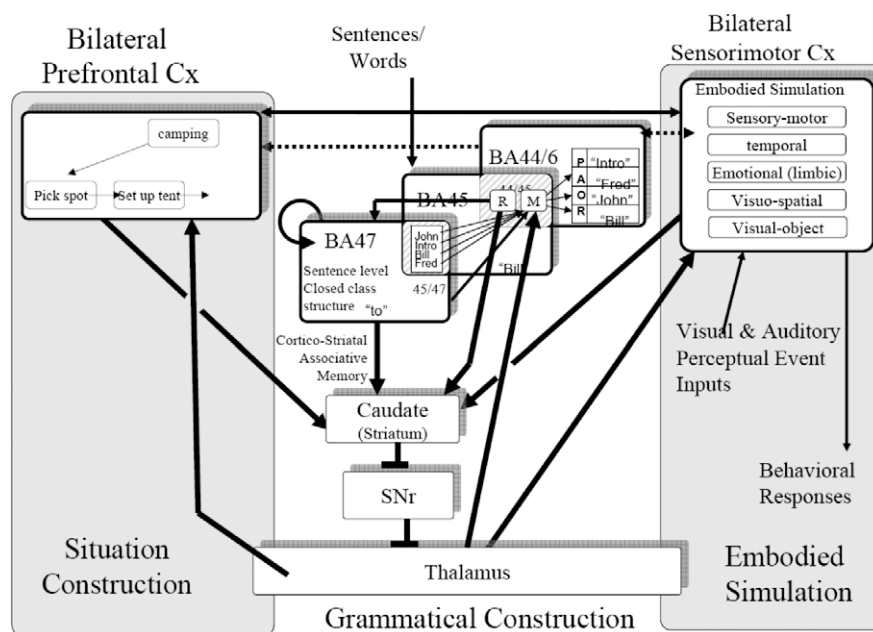


Fig. 4. Towards the Neurophysiological Grounding of the Hybrid Comprehension Model. Overview – The Cortex – Striatum Thalamus-Cortex (CSTC) system provides a general capability for using current cortical activation to retrieve future patterns of activation (prediction) from the Cortico-Striatal associative memory. This has been demonstrated and implemented for Grammatical Constructions (Dominey et al., 2003, 2009). Here we extend the same CSTC processing capability to Situation Constructions and Embodied simulation processing. Grammatical Construction processing (center) is based on Dominey et al. (2009). Situation Construction Processing (left) will extend the notion of CSTC function for grammatical constructions in sentences to situation constructions in discourse. At the lowest level of abstraction, Embodied Simulation (right) corresponds to the classical CSTC sensory-motor loops.

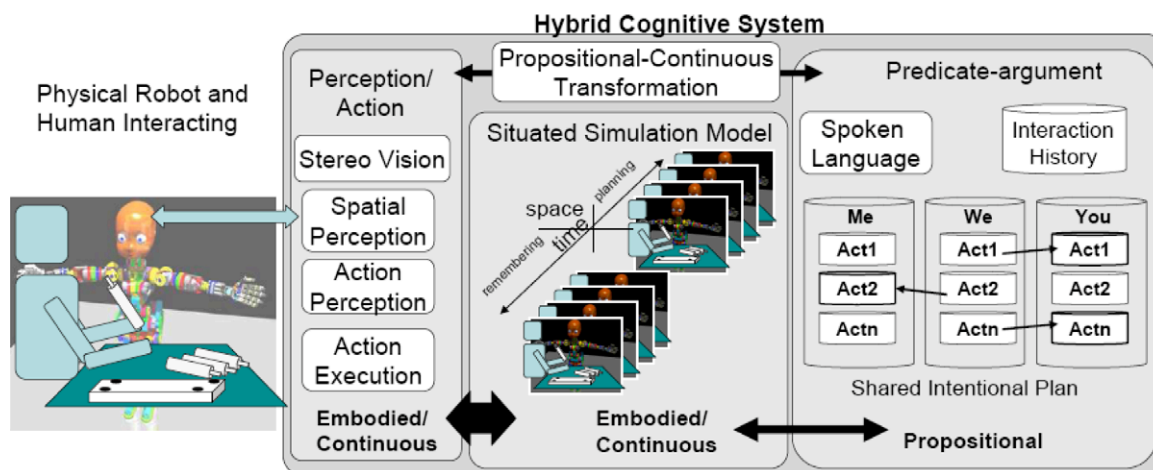


Fig. 5. Proposed Hybrid Cognitive System. A robotic implementation of the Hybrid Comprehension Model. Right panel corresponds to highest level situation constructions from Fig. 4, with Spoken Language corresponding to grammatical constructions. Middle panel Situated Simulations corresponds to embodied simulations from Fig. 4, here implemented as an actual simulator of the robot in its environment, Embodied/Continuous Perception/Action components interface with Propositional Predicate-argument components for language and planning.

specific set of sensorimotor memories to yield more precise understanding.

4.3. Towards a Hybrid Comprehension Model

The reviewed data clearly argue for the reality of situated simulations, yet the use of predicate-argument representations seems necessary in order to escape the temporal constraints of full blown simulation. In order to account for predicate-argument representations of sentences and discourse on the one hand, and embodied situated simulations on the other, we propose the Hybrid Comprehension Model framework. We introduce this framework from two complimentary perspectives: the “purist” perspective of neurophysiologically based simulation models (illustrated in Fig. 4), and the more pragmatic, robotic engineering implementation perspective (illustrated in Fig. 5). From the neurophysiological perspective the Grammatical Construction System (Fig. 4, central panel) implemented in left hemisphere perisylvian cortico-striato-thalamo-cortical (CSTC) system as described in Section 2.2 will activate perceptual-motor traces of experience, which constitute the representational currency for constructing meaningful situation models and thus comprehending described events.

The current framework proposes that representations are implicated in two facets of a hybrid system. First, the perceptual-motor traces are activated within an Embodied Simulation (implemented in the bilateral sensorimotor CSTC, Fig. 4 right panel) to create representations of a given event and thus give rise to a meaningful understanding of that event grounded in our prior experience. Second, these representations can be accessed and manipulated as whole units in the higher-level Situation Construction System (implemented in the bilateral prefrontal CSTC, Fig. 4 left panel). In this manner, multiple events or scenes can be compiled with limited processing effort, and more complex sequences of events can be represented and understood largely through the relations between the events rather than the specific features within an event. Though each event is grounded in perceptual-motor traces of experience (which can be instantaneously accessed with increased attentional resources), this knowledge can be collapsed into processing chunks that are accessed and manipulated without access to their individual features. Such a hierarchical representation of behavior has been proposed in the domain of executive control of action (Koechlin & Summerfield, 2007).

To extend our grammatical construction model from single sentences to discourse comprehension, neural networks responsible for the representation of successions of complex events should be identified. Recent neuroimaging studies are beginning to tackle this issue by analyzing neural activation patterns observed for transitions between different events in a narrative. Speer, Zacks, and Reynolds (2007), for example, recorded neural activity associated with transitions between events and identified bilateral parietal activations. Other networks have been associated with the monitoring of temporal or causal coherence between sequences of events in narratives, situated in the dorsal prefrontal cortices whereas text integration would be monitored at a ventral prefrontal level (Mason & Just, 2006). These observations begin to offer the possibility to provide a broad model of language comprehension from single word to discourse comprehension in an embodied framework. Part of the objective here is to provide a coherent neurophysiological framework for the management of representations at these distinct levels (see Dominey, 2007). Indeed, this approach has proved useful in the study of the neurophysiological mechanisms that underlie grammatical construction processing (Dominey et al., 2003, 2006) and thus bodes well for the future.

A crucial question that arises concerns how these simulations can be controlled – initiated, paused, terminated, etc. Interestingly, we can consider that one of the important functions of language is to allow the speaker to control the unfolding of these simulations in the listener.

5. Language-based control of the simulator

The preceding section reviews evidence that when humans read or hear language, the linguistic input acts as cues to guide simulations of described situations. Grammar is one such component of the linguistic input that has been shown to play a particularly important role in regulating focal and temporal aspects of these simulations. The current section identifies specific behavioral effects of grammar's constraint on simulations in humans, with the present challenge being to incorporate a simulation system in our robot model such that it will reproduce these same effects. Some of the effects are already compatible with the present language capabilities of the robot model, some can be obtained with straightforward adjustments to the model, and some remain outside of the scope of the current model, at least for the near future. This exercise is helpful in determining what aspects of the robot

model are sound, where modifications need to be made, and where the model falls short, with valuable insights on future directions.

There are various ways in which grammar is used to constrain the focus of simulations. For instance, grammatical subjects have been shown to be the preferred antecedents of ambiguous pronouns (Crawley, Stevenson, & Kleinman, 1990; Fredriksen, 1981), suggesting that the subject of a sentence usually weighs in as the most focused entity in a representation of a described situation. These effects can also be observed in the current robot model, even without the aid of a simulator. During sentence production, when given a visual scene event and a focus noun, the model will chose the grammatical construction that describes the scene with the focus noun as the subject or head noun phrase of the sentence (Dominey & Boucher, 2005). Going in the reverse (comprehension) direction, the grammatical subject can be tracked through the `OpenClassArray`, `PredictedReferentsArray` to the `SceneEventArray` where it can then be used to guide future comprehension or production.

The temporal phase of a described situation (onset, middle, offset) can be differentially focused through the grammatical aspect of the verb. The imperfective aspect (was running) has been shown to produce a representation of an ongoing situation at a middle stage of completion, whereas the perfect (had run) and perfective aspects (ran) more often denote the endpoint of a described situation (Madden & Theriault, 2009; Madden & Zwaan, 2003; Magliano & Schleich, 2000; Morrow, 1985). While this type of effect is more difficult to draw out of our previous predicate-argument models, it becomes natural with the introduction of the simulation system. Verbs from the predicate-argument component of the model will not only activate the lexical meaning of an action in the `SceneEventArray`, but within the situated simulation model they will also activate previous experiences of seeing or performing that action. When the imperfective aspect is used, the previous experience will be simulated at an intermediate stage, whereas the perfect(ive) aspect will simulate the endstate of that action.

In the same vein, this feature of the simulator might also be able to accommodate effects of grammatical verb aspect on the focus of simulations. Research has shown that certain features of a situation are more focused by the imperfective aspect than the perfect(ive) aspect, such as the characters involved in a situation (Carreiras, Carriedo, Alonso, & Fernández, 1997), typical locations in which situations can occur (Ferretti, Kutas, & McRae, 2007), and instruments, such as hammers and spoons, that are typically used in situations (Truitt & Zwaan, 1997). Because the robot simulates perfective verbs in an intermediate stage, the instruments as they are in use and the other features of the situation are available for inspection in the simulation. The perfect(ive) aspect simulates the resultant state of the action, and, therefore, the features of the ongoing event will be less available. This use of verb aspect is not yet implemented in the robot model, but would be feasible in the near future.

There is also initial evidence supporting the idea that different components of the memory system are invoked depending on whether a situation is represented as completed or ongoing. For instance, retrieval across situations has been shown to activate areas in the medial temporal lobes used for long term memory systems to a greater degree than ongoing situations, which should rather be maintained in the working memory system (Swallow, Zacks, & Abrams, 2007). Within the current framework, this finding is perhaps compatible with the idea of activating a simulation with many features already active in the perceptual inputs, versus activating a simulation solely through the `SceneEventArray` pathway. While plausible, this idea remains just a hypothesis, and its implementation within the robot model is not yet warranted.

6. Discussion

Language-based control of the simulator may find its most useful expression as it allows us to manage and negotiate cooperation. Cooperative tasks that require coordination during ongoing action could evoke sentences like “OK, while I am lifting this thing up, you put the roller underneath, and then I’ll put it back down.” Such sentences define coordinated actions that will take place during the execution of another action. Indeed, they seem to allow the speaker to control the execution of a simulation in the hearer, in order to specify the cooperation. Trying to understand how to implement this in a robot can help us to understand how it works in humans. In particular, our analyses have led us to the position that a purely symbolic or propositional encoding of events and actions will likely not be sufficient in such situations. In understanding the sentence, if the robot simply invokes the symbolic “lift(x)” action representation, then there will be no point of temporal reference in the representation – nothing that corresponds to the ongoing event “while I am lifting this thing up”. That is, the specified action of putting the roller underneath does not have a clear anchor point. However, if the robot can actually simulate its own perception corresponding to the lift action, then he can examine the unfolding of that action, and even determine when the optimal time to insert the roller will take place.

This leads us to propose a hybrid system in which compact representations of actions, action sequences, coordinated action sequence and shared plans, etc., could be stored in a propositional format as “situation constructions”. But, when necessary, they could be “expanded” via the situated simulations (Mavridis & Roy 2006), providing the hearer with the desired level of detail. Further we hold that language allows the speaker to “direct the film”, to precisely control the initiation, unfolding and termination of appropriate simulations in the mind of the listener, through precise grammatical mechanisms that have evolved for this purpose. Bergen and Chang (2005) had extensively developed related ideas in their embodied construction grammar framework. Similarly, the notion of hybrid systems has a long history in the related domains of artificial intelligence and neural networks, with proposed systems that combine the benefits of symbolic AI systems with the distributed representations of neural networks in hybrid systems, analogous to what we propose here (e.g. Wermter & Sun, 2000). Our future research will examine more closely the relations between these approaches.

In beginning to attempt to put this all together, we continue the computational neuroscience tradition of looking at the functional neuroanatomy data, in order to gain insight on how these mechanisms are implemented. This provides insight that indeed, perception and action of related events activate largely overlapping brain areas specific to the appropriate type of event in parietal, premotor and motor cortical areas. Likewise, the multi-sentence control of these simulations in discourse settings appears to recruit frontal areas that are well situated to orchestrate the unfolding of simulation activity in the sensorimotor areas. We believe that the exploration of embodied sentence processing through the implementation of neurophysiologically grounded robotic systems will provide a new and potent tool in the future development of this research domain.

Acknowledgments

This work is supported by the European Projects *Cooperative Human Robot Interaction Systems* (CHRIS), and *Self-Organized recurrent Neural Learning for Language Processing* (Organic) under the FP7-ICT Cognitive Systems and Robotics Program, and the French ANR projects *Comprendre* and *AMORCES*.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bandl.2009.07.001.

References

- Anderson, M. L. (2003). Embodied cognition: A field guide. *Artificial Intelligence*, 149, 91–130.
- Assmus, A., Giessing, C., Weiss, P. H., & Fink, G. R. (2007). Functional interactions during the retrieval of conceptual action knowledge: An fMRI study. *Journal of Cognitive Neuroscience*, 19, 1004–1012.
- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 577–660.
- Bates, E., McNew, S., MacWhinney, B., Devescovi, A., & Smith, S. (1982). Functional constraints on sentence processing: A cross-linguistic study. *Cognition*, 11, 245–299.
- Bergen, B., & Chang, N. (2005). Embodied construction grammar in simulation-based language understanding. In Jan-Ola Östman & Miriam Fried (Eds.), *Construction grammar(s): Cognitive grounding and theoretical extensions*. Amsterdam: John Benjamins.
- Blanc, J. M., Dodane, C., & Dominey, P. F. (2003). Temporal processing for syntax acquisition: A simulation study. In *Proceedings of the 25th annual meeting of the cognitive science society*. Boston, MA.
- Buccino, G., Vogt, S., Ritzl, A., Fink, G. R., Zilles, K., Freund, H. J., et al. (2004). Neural circuits underlying imitation learning of hand actions: An event-related fMRI study. *Neuron*, 42, 323–334.
- Calabresi, P., Picconi, B., Tozzi, A., & Di Filippo, M. (2007). Dopamine-mediated regulation of corticostriatal synaptic plasticity. *Trends in Neurosciences*, 30, 211–219.
- Carreiras, M., Carriedo, N., Alonso, M. A., & Fernández, A. (1997). The role of verb tense and verb aspect in the foregrounding of information during reading. *Memory & Cognition*, 25, 438–446.
- Chao, L. L., & Martin, A. (2000). Representation of manipulable man-made objects in the dorsal stream. *Neuroimage*, 12, 478–484.
- Crawley, R., Stevenson, R., & Kleinman, D. (1990). The use of heuristic strategies in the interpretation of pronouns. *Journal of Psycholinguistic Research*, 4, 245–264.
- Damasio, A. R. (1989). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition*, 33, 25–62.
- Damasio, A. R., & Tranel, D. (1993). Nouns and verbs are retrieved with differently distributed neural systems. *Proceedings of the National Academy of Sciences*, 90, 4957–4960.
- Dominey, P. F. (1995). Complex sensory-motor sequence learning based on recurrent state-representation and reinforcement learning. *Biological Cybernetics*, 73, 265–274.
- Dominey, P. F. (2005). Emergence of grammatical constructions: Evidence from simulation and grounded agent experiments. *Connection Science*, 17, 289–306.
- Dominey, P. F. (2007). Spoken language and vision for adaptive human-robot cooperation. In Matthias Hackel (Ed.), *Humanoid robotics*. Vienna: ARS International.
- Dominey, P. F., & Boucher, J. D. (2005). Learning to talk about events from narrated video in the construction grammar framework. *Artificial Intelligence*, 167, 31–61.
- Dominey, P. F., Hoen, M., Blanc, J. M., & Lelekov-Boissard, T. (2003). Neurological basis of language and sequential cognition: Evidence from simulation, aphasia, and ERP studies. *Brain and Language*, 86, 207–225.
- Dominey, P. F., Hoen, M., & Inui, T. (2006). A neurolinguistic model of grammatical construction processing. *Journal of Cognitive Neuroscience*, 18, 2088–2107.
- Dominey, P. F., Inui, T., & Hoen, M. (2009). Neural network processing of natural language: II. Towards a unified model of cortico-striatal function in learning sentence comprehension and non-linguistic sequencing. *Brain and Language*, 109(2–3), 80–92.
- Dominey, P. F., Lelekov, T., Ventre-Dominey, J., & Jeannerod, M. (1998). Dissociable processes for learning the surface and abstract structure sensorimotor sequences. *Journal of Cognitive Neuroscience*, 10, 734–751.
- Dominey, P. F., Mallet, A., & Yoshida, E. (2007a). Progress in programming the HRP-2 humanoid using spoken language. In *Proceedings of ICRA 2007*, Rome.
- Dominey, P. F., Mallet, A., & Yoshida, E. (2007b). Real-time cooperative behavior acquisition by a humanoid apprentice. In *Proceedings of the IEEE conference on humanoid robotics*.
- Dominey, P. F., & Warneken, F. (in press). The basis of shared intentions in human and robot cognition. *New Issues in Psychology, Special Issue on Cognitive Robotics & Theoretical Psychology*.
- Ferretti, T. R., Kutas, M., & McRae, K. (2007). Verb aspect and the activation of event knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 182–196.
- Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.
- Fredriksen, J. R. (1981). Understanding anaphora: Rules used by readers in assigning pronominal referents. *Discourse Processes*, 4, 323–347.
- Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences*, 20, 1–55.
- Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, 9, 558–565.
- Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.
- Goldberg, R. F., Perfetti, C. A., Fiez, J. A., & Schneider, W. (2007). Selective retrieval of abstract semantic knowledge in left prefrontal cortex. *Journal of Cognitive Neuroscience*, 27, 3790–3798.
- Goldberg, R. F., Perfetti, C. A., & Schneider, W. (2006). Perceptual knowledge retrieval activates sensory brain regions. *Journal of Cognitive Neuroscience*, 26, 4917–4921.
- Gorniak, P., & Roy, D. (2007). Situated language understanding as filtering perceived affordances. *Cognitive Science*, 31, 197–231.
- Grabowski, T. J., Damasio, H., & Damasio, A. R. (1998). Premotor and prefrontal correlates of category-related lexical retrieval. *Neuroimage*, 7, 232–243.
- Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, 41, 301–307.
- Hoen, M., Pachot-Clouard, M., Segebarth, C., & Dominey, P. F. (2006). When Broca experiences the Janus syndrome: An ER-fMRI study comparing sentence comprehension and cognitive sequence processing. *Cortex*, 42, 605–623.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.
- Koechlin, E., & Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences*, 11, 229–235.
- Kotovskiy, L., & Baillargeon, R. (1998). The development of calibration-based reasoning about collision events in young infants. *Cognition*, 67, 311–351.
- Lungarella, M., Metta, G., Pfeifer, R., & Sandini, G. (2003). Developmental robotics: A survey. *Connection Science*, 15, 151–190.
- MacWhinney, B. (1982). Basic syntactic processes. In *Language development*. In S. Kuczaj (Ed.), *Syntax and semantics* (Vol. 1). Hillsdale, NJ: Lawrence Erlbaum.
- Madden, C. J., & Theriault, D. J. (2009). Verb aspect and perceptual simulation. *Quarterly Journal of Experimental Psychology*, 62(7), 1294–1302.
- Madden, C. J., & Zwaan, R. A. (2003). How does verb aspect constrain event representations? *Memory & Cognition*, 31, 663–672.
- Magliano, J. P., & Schleich, M. C. (2000). Verb aspect and situation models. *Discourse Processes*, 29, 83–112.
- Martin, A., Haxby, J. V., Lalonde, F. M., Wiggs, C. L., & Ungerleider, L. G. (1995). Discrete cortical regions associated with knowledge of color and knowledge of action. *Science*, 270, 102–105.
- Martin, A., Wiggs, C. L., Ungerleider, L. G., & Haxby, J. V. (1996). Neural correlates of category-specific knowledge. *Nature*, 379, 649–652.
- Mason, R. A., & Just, M. A. (2006). Neuroimaging contributions to the understanding of discourse processes. In M. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics*. Amsterdam: Elsevier.
- Mavridis, N., & Roy, D. (2006). Grounded situation models for robots: Where words and percepts meet. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS)*.
- Morrow, D. G. (1985). Prepositions and verb aspect in narrative understanding. *Journal of Memory & Language*, 24, 390–404.
- Noppeney, U., Josephs, O., Kiebel, S., Friston, K. J., & Price, C. J. (2005). Action selectivity in parietal and temporal cortex. *Cognitive Brain Research*, 25, 641–649.
- Pfeifer, R., & Gómez, G. (2005). Interacting with the real world: Design principles for intelligent systems. *Artificial Life and Robotics*, 9, 1–6.
- Pylyshyn, Z. W. (1986). *Computation and cognition: Toward a foundation for cognitive science*. Cambridge, MA: MIT Press.
- Roy, D., Hsiao, K. Y., & Mavridis, N. (2004). Mental imagery for a conversational robot. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 34, 1374–1383.
- Shi, R., Werker, J. F., & Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition*, 72, B11–B21.
- Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61, 39–91.
- Speer, N. K., Zacks, J. M., & Reynolds, J. R. (2007). Human brain activity time-locked to narrative event boundaries. *Psychological Science*, 18, 449–455.
- Swallow, K. M., Zacks, J. M., & Abrams, R. A. (2007). Perceptual events may be the “Episodes” in episodic memory. In *Proceedings of the 48th annual meeting of the psychonomics society*.
- Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, 10, 117–149.
- Tettamanti, M., Buccino, G., Saccuman, M. C., Gallese, V., Danna, M., Scifo, P., et al. (2005). Listening to action-related sentences activates fronto-parietal motor circuits. *Journal of Cognitive Neuroscience*, 17, 273–281.
- Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.
- Truitt, T. P., & Zwaan, R. A. (1997). Verb aspect affects the generation of instrument inferences. In *Proceedings of the 38th annual meeting of the psychonomic society*, Philadelphia.
- Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, 92, 231–270.
- Vernon, D., Metta, G., & Sandini, G. (2007). A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents. *IEEE Transactions on Evolutionary Computation*, 11, 151–180.
- Warneken, F., Chen, F., & Tomasello, M. (2006). Cooperative activities in young children and chimpanzees. *Child Development*, 77, 640–663.
- Wermter, S., & Sun, R. (2000). *Hybrid neural systems*. Heidelberg: Springer.
- Zwaan, R. A., & Madden, C. J. (2005). Embodied sentence comprehension. In Diane Pecher & R. A. Zwaan (Eds.), *The grounding of cognition: The role of perception and action in memory, language, and thinking* (pp. 224–245). Cambridge, UK: Cambridge University Press.