frontiers in
NEUROROBOTICS

ORIGINAL RESEARCH ARTICLE
published: 03 June 2010
doi: 10.3389/fnbot.2010.00008

Linking language with embodied and teleological representations of action for humanoid cognition

Stephane Lallee, Carol Madden, Michel Hoen and Peter Ford Dominey*

Robot Cognition Laboratory, Integrative Neuroscience, Stem Cell and Brain Research Institute, Institut National de la Santé et de la Recherche Médicale U846, Bron, France

Edited by:
Angelo Cangelosi,
University of Plymouth, UK

Reviewed by:
José F. Fontanari,
University of Sao Paulo, Brazil
Claudia Scorolli,
University of Bologna, Italy

****Correspondence:***
Peter Ford Dominey, Robot Cognition Laboratory, Stem Cell and Brain Research Institute, Institut National de la Santé et de la Recherche Médicale, 18, avenue du doyen Jean Lépine, 69675 Bron Cedex, France.
e-mail: peter.dominey@inserm.fr

The current research extends our framework for embodied language and action comprehension to include a teleological representation that allows goal-based reasoning for novel actions. The objective of this work is to implement and demonstrate the advantages of a hybrid, embodied-teleological approach to action–language interaction, both from a theoretical perspective, and via results from human–robot interaction experiments with the iCub robot. We first demonstrate how a framework for embodied language comprehension allows the system to develop a baseline set of representations for processing goal-directed actions such as "take," "cover," and "give." Spoken language and visual perception are input modes for these representations, and the generation of spoken language is the output mode. Moving toward a teleological (goal-based reasoning) approach, a crucial component of the new system is the representation of the subcomponents of these actions, which includes relations between initial enabling states, and final resulting states for these actions. We demonstrate how grammatical categories including causal connectives (e.g., because, if–then) can allow spoken language to enrich the learned set of state-action-state (SAS) representations. We then examine how this enriched SAS inventory enhances the robot's ability to represent perceived actions in which the environment inhibits goal achievement. The paper addresses how language comes to reflect the structure of action, and how it can subsequently be used as an input and output vector for embodied and teleological aspects of action.

Keywords: human-robot interaction, action perception, language, cooperation

INTRODUCTION – A FRAMEWORK FOR LANGUAGE AND ACTION

One of the central functions of language is to coordinate cooperative activity (Tomasello, 2008). In this sense, much of language is about coordinating action. Indeed, language constructions themselves become linked to useful actions in our experience, as emphasized by Goldberg (1995, p. 5) "constructions involving basic argument structure are shown to be associated with dynamic scenes: experientially grounded gestalts, such as that of someone volitionally transferring something to someone else, someone causing something to move or change state…" Interestingly, this characterization is highly compatible with the embodied language comprehension framework, which holds that understanding language involves activation of experiential sensorimotor representations (Barsalou, 1999; Bergen and Chang, 2005; Zwaan and Madden, 2005; Fischer and Zwaan, 2008; Pulvermüller et al., 2009). We have pursued this approach in developing neurally inspired systems that make this link between language and action. We introduce this approach in the remainder of this section, describing the path we have taken to arrive at our present work.

In this context of linking language and action, we first developed an action recognition system that extracted simple perceptual primitives from the visual scene, including contact or collision (Kotovsky and Baillargeon, 1998), and composed these primitives into templates for recognizing events like give, take, touch and push. Siskind and colleagues (Fern et al., 2002) developed a related action learning capability in the context of force dynamics. A premise of this approach is that it is not so much the details of spatial trajectories of actions, but more their resulting states which characterize action in the context of perception and recognition (Bekkering et al., 2000). The resulting system provided predicate–argument representations of visually perceived events, which could then be used in order to learn the mapping between sentences and meaning. We demonstrated that naïve humans could narrate their actions which were perceived by the event recognition system, thus providing sentence-meaning inputs to the grammatical construction model, which was able to learn a set of grammatical constructions that could then be used to describe new instances of the same types of events (Dominey and Boucher, 2005).

We subsequently extended the grammatical construction framework to robot action control. We demonstrated that the robot could learn new behaviors (e.g., Give me the *object*, where *object* could be any one of a number of objects that the robot could see) by exploiting grammatical constructions that define the mapping from sentences to predicate–argument representations of action commands. This work also began to extend the language–action framework to multiple-action sequences, corresponding to more complex behaviors involved in cooperative activity (Dominey et al., 2009b). Cooperation – a hallmark of human cognition (see Tomasello et al., 2005) – crucially involves the construction of action plans that specify the respective contribution of both agents, and the representation of this shared plan by both agents. Dominey and Warneken (in press) provided the Cooperator – a 6DOF arm and monocular vision robot – with this capability, and demonstrated that the resulting system could engage in cooperative activity, help the human, and perform role reversal, indicating indeed that it had a "bird's eye view" of the cooperative activity. More recently, Lallee et al. (2009) extended this work so that the robot could acquire shared plans by observing two humans perform a cooperative activity.

An important aspect of this area of research is that the source of meaning in language is derived directly from sensory-motor experience, consistent with embodied language processing theories (Barsalou, 1999; Bergen and Chang, 2005; Zwaan and Madden, 2005). For instance, Fontanari et al. (2009) have demonstrated that artificial systems can learn to map word names to objects in a visual scene in a manner that is consistent with embodied theories. However, we also postulated that some aspects of language comprehension must rely on a form of "hybrid" system in which meaning might not be expanded completely into its sensory-motor manifestation (Madden et al., 2010). This would be particularly useful when performing goal-based inferencing and reasoning. Indeed, Hauser and Wood (2010) argue that understanding action likely involves goal-based teleological reasoning processes that are distinct from the embodied simulation mechanisms for action perception. These authors state that, "Integrating insights from both motor-rich (simulation, embodiment) and motor-poor (teleological) theories of action comprehension is attractive as they provide different angles on the same problem, a set of different predictions about the psychological components of action comprehension, and enable a broad comparative approach to understanding how organisms interpret and predict the actions of others" (Hauser and Wood, 2010, p. 4). This is consistent with a hybrid approach to action understanding that we have recently proposed (Madden et al., 2010; for other dual-representation approaches see: Barsalou et al., 2008; Dove, 2009). In that model, action perception and execution take place in an embodied sensorimotor context, while certain aspects of planning of cooperative activities are implemented in an amodal system that does not rely on embodied simulation.

A fundamental limitation of this approach to date is that the system has no sense of the underlying goals for the individual or joint actions. This is related to the emphasis that we have placed on recognition and performance of actions, and shared action sequences, without deeply addressing the enabling and resulting states linked to these actions. In the current research, we extend our hybrid comprehension model to address aspects of goal-based reasoning, thus taking a first step toward the type of teleological reasoning advocated by Hauser and Wood (2010). The following section describes how this new framework addresses the limitations of the current approach.

A NEW FRAMEWORK FOR ACTION AND LANGUAGE – COMBINING TELEOLOGICAL AND EMBODIED MECHANISMS

In Lallee et al. (2009) the iCub robot could observe two human agents perform a cooperative task, and then create a cooperative plan, which includes the interleaved temporal sequence of coordinated actions. It could then use that plan to take the role of either of the two agents in the learned cooperative task. This is illustrated in **Figure 1**. A limitation of this work is that the task is represented as a sequence of actions, but without explicit knowledge of the results of those actions, and the link between them. In the current work, this limitation is addressed by allowing the robot to learn for each action, what is the enabling state of the world which must hold for that action to be possible, and what is the resulting state that holds once the action has been performed. We will refer to this as the S_EAS_R state-action-state (SAS) representation of action. This is consistent with our knowledge that humans tend to represent actions in terms of goals – states that result from performance of the action (Woodward, 1998). Furthermore, neurophysiological evidence of such a goal specific encoding of actions has been observed in monkeys (Fogassi et al., 2005) whereby the same action (grasping) can be encoded in different manners according to intentions or goals (grasping for eating/grasping for placing).

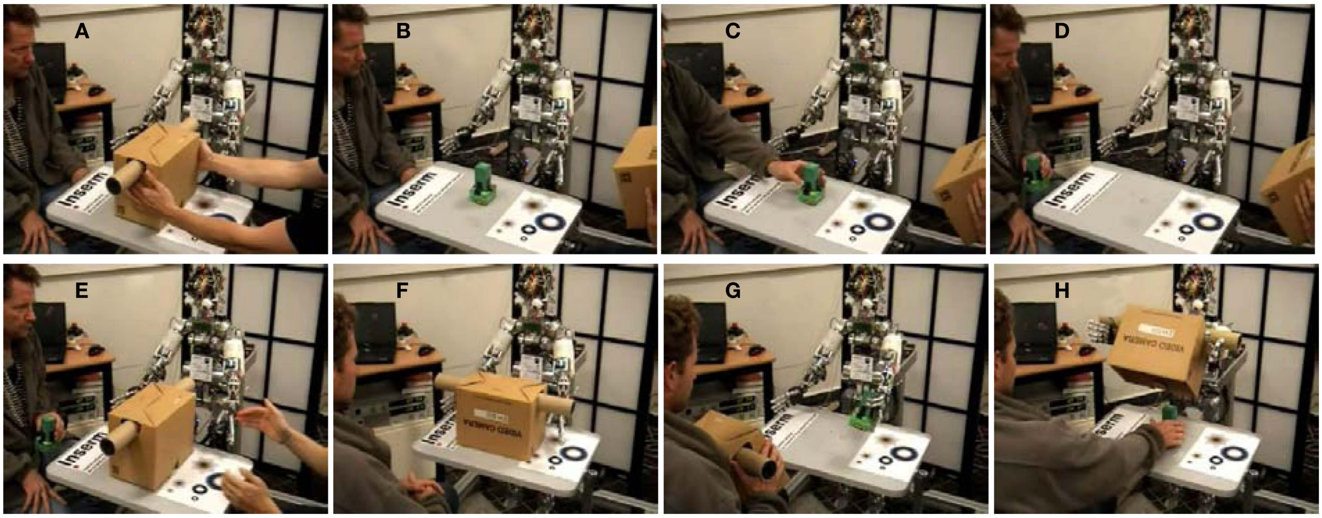


FIGURE 1 | On-line learning of a cooperative task. (A,B) Larry (left of robot) lifts the box that covers the toy. **(C,D)** This allows Robert (right of robot) to take the toy. **(E)** Larry replaces the box. **(F)** Robot now participates. **(G)** Human takes box, so Robot can take the toy. **(H)** Robot takes box so human can take the toy.

Interestingly, we quickly encountered limitations of the perceptual system, in the sense that when an action causes an object to be occluded, the visual disappearance of that object is quite different from the physical disappearance of the object, yet both result in a visual disappearance. The ability to keep track of objects when they are hidden during a perceived action, and the more general notion of object constancy is one of the signatures of core object cognition (Spelke, 1990; see Carey, 2009). This introduces the notion that human cognition is built around a limited set of "core systems" for representing objects, actions, number and space (Spelke and Kinzler, 2007). Robot cognition clearly provides a testing ground for debates in this domain, and the current study uses this platform to investigate the nature of the core system for agency. Embodied theories hold that actions are interpreted by mental simulation of the observed action, while teleological theories hold that this is not sufficient, and that a generative, rationality-based inferential process is also at work in action understanding (Gergely and Csibra, 2003). In our work, we employ both embodied learning of actions as well as a higher-level symbolic processing of these actions to yield a better understanding of the causes and consequences of events in the world. There are several research teams conducting very important and interesting work in scaling up from the primary perceptual layers (e.g., Fontanari et al., 2009; Tikhanoff et al., 2009). Our aim is to use the output of these layers in a more abstract and symbolic reasoning mainly driven by language, combining two approaches that are not antagonistic but rather complementary. This dual approach is consistent with Mandler's (2008) ideas of developmental concepts, as well as the role of amodal lexical associations in embodied language theories (e.g., Glaser, 1992; Kan et al., 2003), and several representational theories of meaning (Borghi and Cimatti, 2009; Dove, 2009).

As event understanding often involves inferences of links between intentions, actions, and outcomes, language can play an important role in helping children learn about relations between actions and their consequences (Bonawitz et al., 2009). The following section provides an overview of how language is used to enrich perceptual representations of action, and some of the corresponding neurophysiological mechanisms that provide some of these capabilities, based largely on data from humans. It is our belief that understanding these behavioral and neurophysiological mechanisms can provide strong guidelines in constructing a system for robot event cognition in the context of human–robot cooperation.

ASPECTS OF LANGUAGE AND CAUSALITY

One of the hallmarks of human cognition is the ability to understand goal-directed events. This ability surely entails the representation of events in terms of their causes and effects or goals (Bekkering et al., 2000; Sommerville and Woodward, 2005), but how does it work? Although some theorists have postulated that causality itself is a conceptual primitive, it has become evident that causality can be decomposed into constituent elements (see Carey, 2009 for discussion). According to physicalist models of causality, causes and effects are understood in terms of transfer or exchange of physical quantities in the world, such as energy, momentum, impact forces, chemical and electrical forces (Talmy, 1988; Wolff, 2007). Furthermore, nonphysical causation (e.g., forcing someone to decide) is understood by analogy to these physical forces. In this sense, physicalist models necessitate the ability to perceive kinematics, and dynamic forces, in order to represent causal relationships between entities. That is, to understand causality, one must have a body, and thus any implementation model of causal understanding necessitates an embodied system, to sense physical forces.

Dynamic forces are often invisible, such as the difference in the feeling of contact when an object is moving fast or slow, and how a pan feels when it is hot or cold. Because invisible dynamic forces map so well onto our experience of *kinematic* forces, or visual experience of forces (shape, size, position, direction, velocity, accelerations), humans often rely solely on visual information when attributing causal relationships in the world. In the same vein, causal understanding in non-human systems can be implemented through the use of kinematics as perceived via vision (e.g. Michotte, 1963). Thus, in our current work, we capitalize on this aspect of visual perception and restrict our representation of events to perceptual primitives that fall out of the visual input, leaving other perceptual modalities as well as motor actions for future implementation. Fern et al. (2002) and Siskind et al. (2001, 2003) have exploited the mapping of force dynamic properties into the visual domain, for primitives including contact, support and attachment. This results in robust systems in which event definitions are prespecified or learned, and then used for real-time event classification. Dominey and Boucher (2005) employed a related method for the recognition of events including give, take, push, touch in the context of grounded language acquisition.

In the context of development, once a toddler is able to sense and understand physical forces in the environment, he has the tools to understand causal relationships. Pioneering studies have shown that this understanding of causality and causal language is acquired very early in development, as infants may already perceive cause-effect relationships at only 27 weeks (Leslie and Keeble, 1987), and toddlers can already express many types of causal language by the age of 2–3 years (Bowerman, 1974; Hood et al., 1979). At this stage, exposure to language may help to accelerate the development of causal understanding. One study has shown that when toddlers are exposed to a causal relationship between two events accompanied by a causal description, they are more likely to initiate the first event to generate the second, and expect that the predictive relations will involve physical contact, compared to when they are exposed to the causal situation in the absence of causal language (Bonawitz et al., 2009). That is, though the toddler associates the two events in either case, this association might not be recognized as a causal link, and causal language, such as "the block makes the light turn on," can help to explicitly establish this link.

In this way, language is used as a tool to further conceptual understanding of goal-directed events and actions by helping toddlers more quickly integrate information about prediction, intervention, and contact causality. Thus, we can exploit language in our current system as a vector for establishing causal links between actions and their resulting states. In particular we are interested in the states that result from the "cover" and "give" actions which involve states related to the covered object being present, but invisible in the first case, and notions of change of possession in the second.

CORTICAL NETWORKS FOR LANGUAGE COMPREHENSION

In our effort to develop a system that can represent events and the state-transition relations between events, we can exploit knowledge of how language and event comprehension are implemented in the human nervous system. Language comprehension involves a cascade of computational operations starting from the decoding of speech in sensory areas to the emergence of embodied representations of the meaning of events corresponding to sensory-motor simulations (Barsalou, 1999; Bergen and Chang, 2005; Zwaan and Madden, 2005; see Rizzolatti and Fabbri-Destro, 2008 for review). These representations are triggered via: observation of others engaged in sensory-motor events; imagination of events and the evocation of these experiences through language. Therefore, we consider the existence of two parallel but interacting systems: one system for language processing, ultimately feeding information processes into a second system, dedicated to the processing of sensory-motor events. These systems are highly interconnected and their parallel and cooperative work can ultimately bootstrap meaning representations. The second system will also accommodate the representation of elaborated events that implicates processes derived from a system sometimes referred to as a "social perception" network (Decety and Grèzes, 2006; see Wible et al., 2009 for review). This second network is directly involved in teleological aspects of reasoning, including agency judgments, attributing goals and intentions to agents, inferring rationality about ongoing events and predicting outcomes of the ongoing simulation (Hauser and Wood, 2010). We will present these two systems and show how they interact to form complex meaning representations through language comprehension.

One central view in the recent models of the cortical processing of language is that it occurs along two main pathways, mostly lateralized to the left cortical hemisphere (Ullman, 2004; Hickok and Poeppel, 2007; see also Saur et al., 2008). The first route is referred to as the ventral-stream. It is dedicated to the recognition of complex auditory (or visual) objects involving different locations along the temporal lobe and the ventralmost part of the prefrontal cortex (BA 45/46). The second one is named the dorsal-stream and is dedicated to the connection between the language system and the sensory-motor system, that is both implicated in the transformation of phonetic codes into speech gestures for speech production, but also in the temporal and structural decoding of complex sentences (Hoen et al., 2006; Meltzer et al., in press). It implicates regions in the posterior part of the temporo-parietal junction (TPJ), parietal and premotor regions and reaches the dorsal part of the prefrontal cortex (BA 44).

In the ventral pathway, speech sounds are decoded in or nearby primary auditory regions of the dorsal superior temporal gyrus (BA 41/42), before phonological codes can be retrieved from the middle posterior superior temporal sulcus (mp-STS – BA 22), and words recognized in regions located in the posterior middle temporal gyrus (pMTG – BA 22/37; see Hickok and Poeppel, 2007 for review; Scott et al., 2006; Obleser et al., 2007). Then, these lexical symbols can trigger the reactivation of long-term stored sensory-motor experiences, either via implications of long-term autobiographic memory systems in the middle temporal gyrus or in long-term sensory-motor memories, with a widespread storage inside the sensory-motor system. Therefore, complex meaning representation can actually engage locations from the ventral pathway but also memories stored inside the dorsal pathway (Hauk et al., 2008; e.g., Tettamanti et al., 2005). This primary network feeds representation into a secondary-extended cortical network, whenever language leads to complex mental representations of complex events. Our initial computational models predicted dual structure-content pathway distinction (Dominey et al., 2003), which was subsequently confirmed in neuroimaging studies demonstrating the existence and functional implication of these two systems (Hoen et al., 2006), leading to further specification of the model (Dominey et al., 2006, 2009a).

TOWARD A NEUROPHYSIOLOGICAL MODEL OF EMBODIED AND TELEOLOGICAL EVENT COMPREHENSION

More recently, we extended this to a hybrid system in which sentence processing interacts both with a widespread embodied sensory-motor system, and with a more amodal system to account for complex event representation and scenario constructions operating on symbolic information (Madden et al., 2010). This second network, seems to engage bilateral parietal–prefrontal connections including bilateral activations in the parietal lobule for the perception and monitoring of event boundaries (Speer et al., 2007) as well as dorsal prefrontal regions seemingly implicated in the global coherence monitoring of the ongoing mental representation elaboration (Mason and Just, 2006). The monitoring of complex event representation includes the ability of deciding if ongoing linguistic information can be inserted in the current representation and how it modifies the global meaning of this representation. These aspects rely on information and knowledge that are not primary characteristics of the language system *per se* but rather include general knowledge about causal relations between events, intentionality and agency judgments etc. These properties are sometimes called teleological reasoning and different authors have now shown that processes involving teleological reasoning are sustained by a distributed neural network, referred to as a "social perception" cognitive network that is closely related to the language system (Wible et al., 2009).

This social perception network is implicated in teleological reasoning as determining agency or intentionality relations and involves regions as the right inferior parietal lobule (IP), the superior temporal sulcus (STS) and ventral premotor regions. All these regions are part of the well-known mirror system (Decety and Grèzes, 2006). The TPJ or IP and STS regions, in addition to being part of the mirror system, are also heavily involved in other social cognition functions. Decety and Grèzes (2006), in an extensive review, have designated the right TPJ as the "social" brain region. Theory of mind is the ability to attribute and represent other's mental states or beliefs and intentions or to "read their mind" ("predict the goal of the observed action and, thus, to "read" the intention of the acting individual" – from Decety and Grèzes, 2006, p. 6). Therefore, it seems that regions that are implicated in social-cognition, that is to say regions implicated in agency, intentionality judgments on others are also implicated in the same judgments on a simulation/representation of mental simulations triggered by language.

Figure 2 illustrates a summary representation of the cortical areas involved in the hybrid, embodied-teleological model of language and event processing. The language circuit involves the frontal language system including BA 44 and 45 with a link to embodied representations in the premotor areas, and in the more posterior parietal areas – both of which include mirror neuron

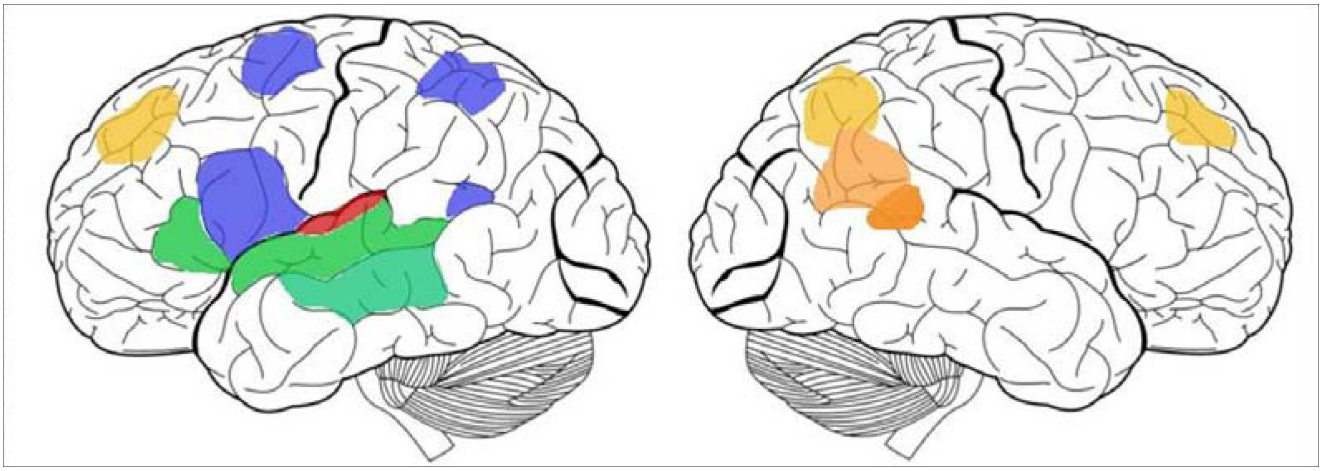


FIGURE 2 | Cortical networks for language processing (simplified). Ventral stream areas (green) are part of a first network dedicated to speech decoding and phonological/lexical processing along the superior temporal sulcus (STS), middle temporal gyrus (MTG) and ventral prefrontal cortex (Pfc). Dorsal stream areas (blue) constitute a sensory-motor interface implicated both in the transcription of phonological codes into articulatory codes (adapted from Hickok and Poeppel, 2007) but also in the temporal/structural organization of complex sentence comprehension, and engage the left temporo-parietal junction, the parietal lobule and dorsal prefrontal regions (Hoen et al., 2006; Meltzer et al., in press). The social perception or teleological cognition network (oranges) is implicated in complex event representation and the attribution of agency, theory of mind in the right TPG (orange, from Decety and Lamm, 2007), causality and intentionality in the posterior STS (dark orange, from Saxe et al., 2004; Brass et al., 2007), and also comprises areas implicated in the global monitoring of the coherence of event representation (light orange, from Mason and Just, 2006). Networks are shown in their specialized hemispheres but most contributions are bilateral.

activity in the context of action representation. This corresponds to the embodied component of the hybrid system. The teleological reasoning functions are implemented in a complimentary network that includes STS and TPJ/IP. In the current research, while we do not model this hybrid system directly in terms of neural networks, we directly incorporate this hybrid architecture into the cognitive system for the robot.

MATERIALS AND METHODS

This section will present in three parts the physical platform, the behavioral scenarios, and the system architecture.

THE iCub HUMANOID AND SYSTEM INFRASTRUCTURE

The current research is performed with the iCub, a humanoid robot developed as part of the RobotCub project (Tsagarakis et al., 2007). The iCub is approximately 1 m tall, roughly the size and shape of a 3-year-old child, and its kinematic structure has a total of 53 degrees of freedom controlled by electric motors, primarily located in the upper torso. The robot hands are extremely dexterous and allow manipulation of objects thanks to their 18 degrees of freedom in total. The robot head is equipped with cameras, microphones, gyroscopes and linear accelerometers. The iCub is illustrated in **Figures 1 and 4**.

Our research focuses on cognitive functions that operate on refined sensory data. We use off the shelf systems for both visual object and word recognition because they handle this raw sensory information quite well. Spoken language processing and overall system coordination is implemented in the CSLU Rad toolkit. The system is provided with an "innate" recognition vocabulary including a set of action names (give, take, touch, cover, uncover), derived predicates (on, has), object names (block, star, sign), and causal language connectives (if–then, because). That is, a list of words is given to the system in advance, so it will be able to recognize them from speech. It is possible to have the speech recognition behaviors emerge (e.g., Fontanari et al., 2009), but as mentioned above, that was not the goal of this work. The ability to recognize this innate vocabulary and use it in recognition grammars is provided by the CSLU toolkit which deals with HHM processing of the sound signal. The grammars that parse the speech signal both for input and output are hard coded into the system, however the system learns to associate a parsed sentence (verb, subject, object) with the visual perception of the corresponding action. Vision is provided by a template-matching system (Spikenet™) based on large spiking neurons networks, here again we use this tool to make a bridge between the raw sensory images, and the symbols of recognized objects. We developed state and action management in C#. Interprocess communication is realized via the YARP protocol.

EXPERIMENTAL SCENARIOS

In this section we describe the experimental human–robot interaction scenarios that define the functional requirements for the system. The current scenarios concentrate on action representation in the embodied and teleological frameworks. They demonstrate how language can be used (1) to enrich the representation of action and its consequences, and (2) to provide access to the structured representation of action definitions, and current knowledge of the robot. An embodied artificial system should incorporate both perceptual and motor representations in action comprehension, and current work is underway on this issue. However, in the current demonstrations we focus solely on perceptual (visual) representations of actions.

First we put the emphasis on the robot's ability to learn from the human when the human performs physical actions with a set of visible objects in the robot's field of view. Typical actions include covering (and uncovering) one object with another, putting one object next to another, and briefly touching one object with another. For actions that the robot has not seen before, the robot should ask the human to describe the action. The robot should learn the action description (e.g., "The block covered the star"), and be capable of generalizing this knowledge to examples of the same action performed on different objects. For learned actions, the robot should be able to report on what it has seen. This should take place in a real-time, on-line manner. Knowledge thus acquired should be available for future use, and in future work, the robot will also be able to learn its own motor representations of actions.

Another element that has to be learned is the causal relation between an action and the resulting state, which is not always trivial. When one object covers another, the second object "disappears" but is still physically present, beneath the covering object. In this scenario actions are performed that cause state changes, in terms of the appearance and disappearance of objects. The robot should detect these changes and attempt to determine their cause. The cause may be known, based on prior experience. If not, then the robot should ask the human, who will use speech for clarification about this causal relation.

The links between actions and their enabling and resulting states correspond directly to grammatical expressions with the if–then construction. The sentence "If you want to take the block then the block must be visible" expresses an enabling relation, where the state "block visible" enables the action "take the block." In contrast, the sentence "If you cover the star with the block, then the star is under the block," or "If you cover the star with the block then the star is not visible" expresses a causal relation. This scenario should demonstrate how by using these forms of grammatical constructions, we can interrogate the system related to these enabling and causal relations.

Once the robot has learned about new actions in one context, we want it to use this knowledge in another context. Concretely, in the cooperative task where Larry uncovers the toy so that Robot can pick it up, the robot should be able to begin to make the link between the resulting state of the "uncover" action as the enabling state of the subsequent "take" action. In this experiment, through a process of interrogation we will demonstrate that the robot has the knowledge necessary to form a plan for getting access to a covered object, by linking goals with resulting states of actions, and then establishing the enabling state as a new goal. After each learning session, the robot knowledge is stored in a long-term memory which we call Knowledge Base. It stores all the action definitions and their causes and consequences in term of states in an XML file that can be loaded on the robot.

We monitor the evolution of the Knowledge Base in order to analyze the performance of the recognition capabilities of the system under extended use. We start with a naïve system (i.e., an empty Knowledge Base), and then for the five actions *cover*, *uncover*, *give*, *take*, and *touch*, we expose the robot to each action with the block and the sign, and then in the transfer condition test the ability to recognize these actions with a new configuration (i.e., with the block and the star). We repeat this exhaustive exposure five times (one for each action). The dependant measure will be the number of presentations required for the five actions to be recognized in the training configuration, and transfer configuration, in each of the five phases. This experiment is detailed in Section "Usage Study" and in **Figure 5**.

COGNITIVE SYSTEM ARCHITECTURE

We developed a cognitive system architecture to respond to the requirements implied in Section "Experimental Scenarios," guided by knowledge of the cognitive linguistic mechanism in humans and their functional neurophysiology, and by our previous work in this area (See **Figure 3**). The resulting system is not neuro-mimetic, but its architecture is consistent with and inspired by our knowledge of the corresponding human system and on neural correlates found in the monkey (Fogassi et al., 2005). We describe the architecture in the context of processing a new action, and illustrated in **Figure 4**.

The human picks up the block and places it on the sign. Vision provides the front end of the perceptual system. Video data from the eyes of the iCub are processed by the Spikenet vision software which provides robust recognition for pretrained templates that recognize all objects in the scene. Each template is associated with a name and the camera coordinates of the recognized location. One to four templates were required per object.

Based on our previous work, inspired by human developmental studies, we identified three perceptual primitives to be extracted from the object recognition, which would form the basis for generic action recognition – these are *visible(object, true/false)*, *moving(object, true/false)*, and *contact(obj1,obj2, true/false)*. These primitives are easily extracted from the Spikenet output based on position and its first derivative, and are provided as input to Temporal Segmentation. The temporal segmentation function returns the most recent set of segmented primitives that occurred within specified time window. This corresponds to our hypothesis that a given complex action will be constituted by a pattern of primitives that occur in a limited time window, separated in time by periods with no action. The resulting pattern of primitives for contact is illustrated in **Figure 4C**.

When the robot detects changes in the visual scene, the above processing is initiated. The Action Management function matches the resulting segmented perceptual primitives with currently defined action in the Knowledge Base. Each action in the Knowledge Base is defined by its pattern of action primitives, its name, the arguments it takes, any preconditions (i.e., the enabling state S_E in the S_EAS_R representation), and the resulting state. Thus, during action recognition, the Action Management function compares this set of segmented primitives with existing action patterns in the Knowledge Base. If no match is found then the system prompts the human to specify the action and its arguments, e.g., "I cover the sign with the block."

The State Management determines that as a result of the action, the World State has changed, and interrogates the user about this. The user then has the opportunity to describe any new relations that result from this action but that are not directly perceptible. When the block covers the sign, the sign is no longer visible, but still present. The State Management asks "Why is the sign no longer visible?" Thus the human can explain this loss of vision by saying "Because the block is on the sign." The action manager binds this relation in a generic way (i.e., it generalizes to new objects when the event "cover" is perceived) to the definition of "cover" (see **Figure 4D**).

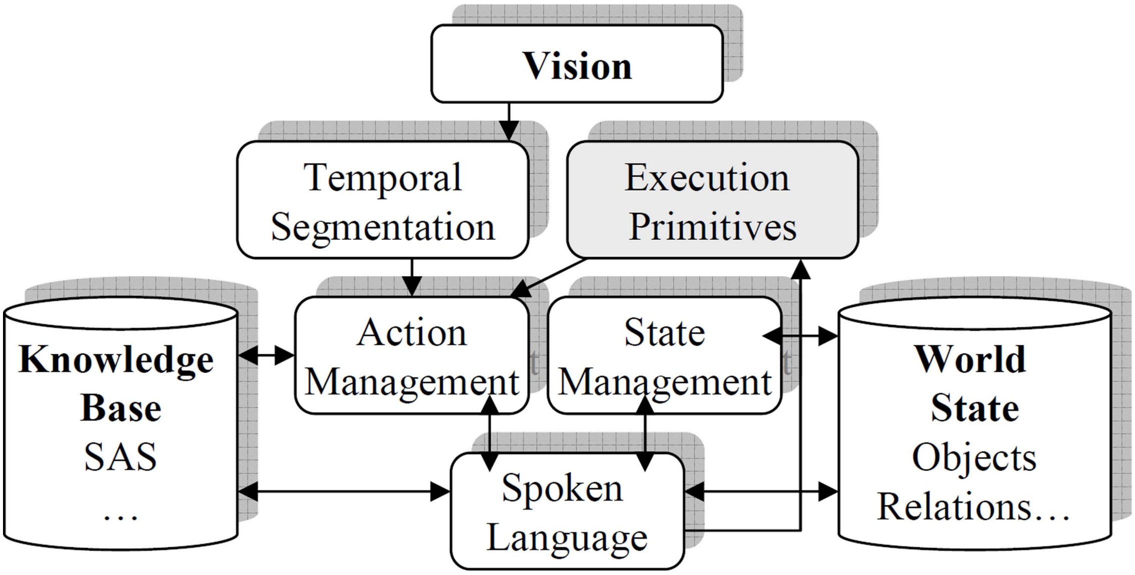


FIGURE 3 | Cognitive system architecture. See text for description.

If a match is found, then the system maps the concrete arguments in the current action segment with the abstract arguments in the action pattern. It can then describe what happened. For a recognized action, State Management updates the World State with any resulting states associated with that action. In the case of cover, this includes encoding of the derived predicate on (block, star).

RESULTS

LEARNING NEW ACTIONS AND THEIR DERIVED CONSEQUENCES

Here we present results from an interaction scenario in which the user teaches the robot four new actions: cover, uncover, give and take. In order to explain the system level functionality, details for learning are illustrated in **Figure 4** for the action "cover." The corresponding dialog is presented in **Table 1**.

For new actions (that have not yet been defined in the Knowledge Base) the system uses the set of observed primitives from Temporal Segmentation to generate a generic pattern of primitives to define the action (**Figure 4C**). If any unexpected perceptual changes occur, the system asks the human why this is the case, and the human can respond by describing any new relation that holds. For example, when the block covers the sign, the sign becomes not visible. The system asks the human why, and the human responds that this is "because the block is on the sign." This new relation on (block, sign) is added as part of the generic definition of the cover action, illustrated in **Figure 4D**.

Table 1 provides a record of the interaction in which the robot learns the meaning of "cover" and then displays this knowledge by recognizing cover in a new example. We observed that executing a given action like cover may sometimes lead to a different ordering of the segmented primitive events, e.g., detecting of the end of the block's movement may occur before or after the sign being visually obstructed. This is accommodated by encoding multiple patterns for a given action in the database. This redundant coding captures the physical redundancy that is expressed in the observations made by the system. The result is that when any of the appropriate patterns for an action are recognized, the action is recognized.

A total of five distinct actions were learned and validated in this manner. The resulting definitions are summarized in **Table 2**. **Figure 5** provides some performance statistics for learning these actions and then using the learned definitions to recognize new actions.

USE OF CAUSAL CONSTRUCTIONS TO INTERROGATE S_EAS_R REPRESENTATIONS

This experiment demonstrates how the "if–then" construction can be used to extract the link between actions, the required enabling states, and the resulting states. Results are presented in **Table 3**.

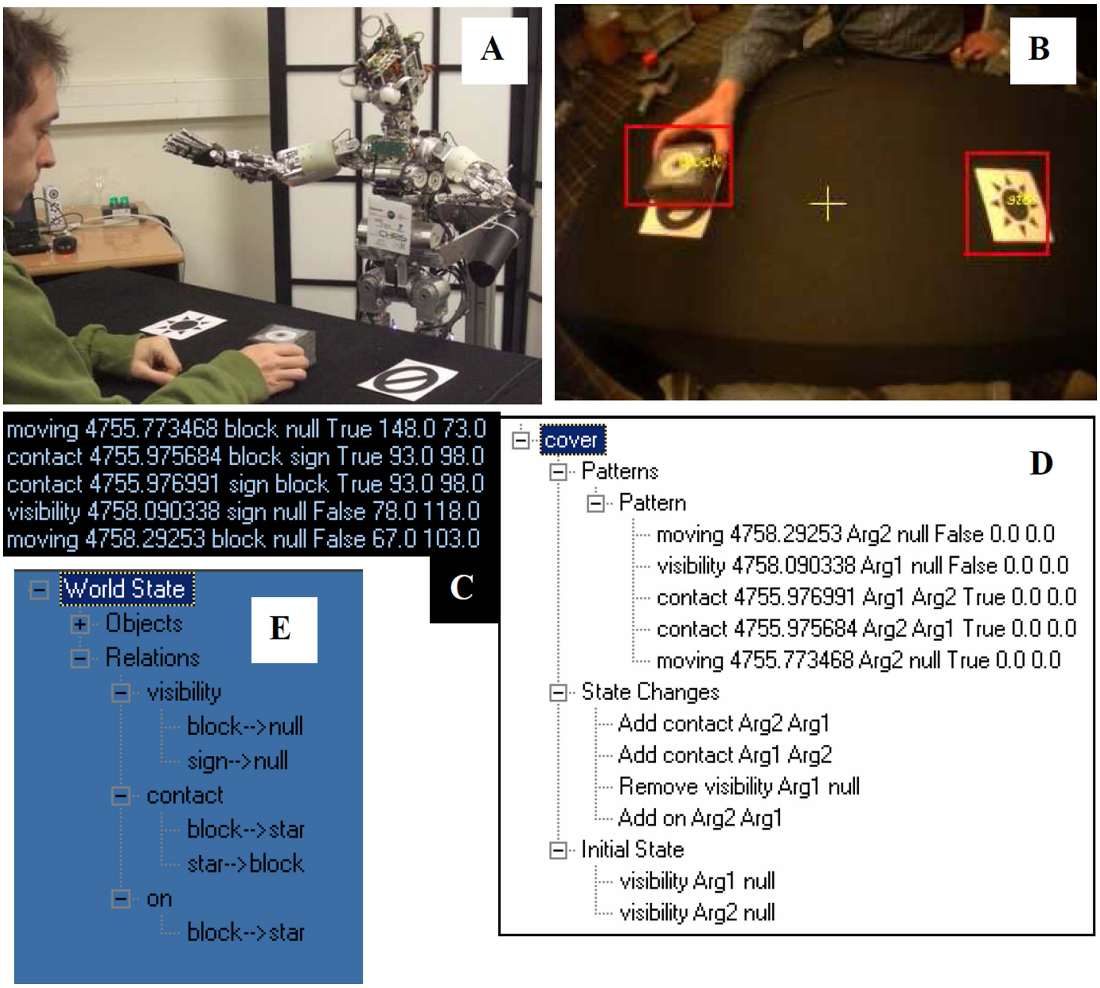
USE OF CAUSAL KNOWLEDGE IN TELEOLOGICAL REASONING

Here we consider a scenario similar to "uncover the block" scenario introduced in Section "Introduction – A Framework for Language and Action," and **Figure 1**. In this context, an object is covered by another, and the user's goal is to use the first object in a new task. The goal then is to find out how to gain access to the first object that is currently covered. The robot observes one human put the toy on the table, and another human cover the toy with the box. The objective is to begin to perform teleological reasoning about action sequences that have never been observed. Results are presented in Table 4.

This experiment demonstrates how the SAS (S_EAS_R) representation provides the required information for goal-based reasoning.

USAGE STUDY

We performed six additional experiments, which involved processing of 111 separate actions, to begin to evaluate the robustness of the system. Experiments 1–4 each started with an empty Knowledge Base, and examined the ability to learn the five actions, and then transfer this knowledge to new object configurations.



```
moving 4755.773468 block null True 148.0 73.0
contact 4755.975684 block sign True 93.0 98.0
contact 4755.976991 sign block True 93.0 98.0
visibility 4758.090338 sign null False 78.0 118.0
moving 4758.29253 block null False 67.0 103.0
```

```
⊟ cover
  ⊟ Patterns
    ⊟ Pattern
        moving 4758.29253 Arg2 null False 0.0 0.0
        visibility 4758.090338 Arg1 null False 0.0 0.0
        contact 4755.976991 Arg1 Arg2 True 0.0 0.0
        contact 4755.975684 Arg2 Arg1 True 0.0 0.0
        moving 4755.773468 Arg2 null True 0.0 0.0
  ⊟ State Changes
      Add contact Arg2 Arg1
      Add contact Arg1 Arg2
      Remove visibility Arg1 null
      Add on Arg2 Arg1
  ⊟ Initial State
      visibility Arg1 null
      visibility Arg2 null
```

```
⊟ World State
  ⊞ Objects
  ⊟ Relations
    ⊟ visibility
        block-->null
        sign-->null
    ⊟ contact
        block-->star
        star-->block
    ⊟ on
        block-->star
```

FIGURE 4 | Learning and generalizing "cover Arg1 with Arg2." (A) Robot setup and visual scene before the action. **(B)** Vision: Robot's view of scene after the block is put on the sign. **(C)** Temporal Segmentation: Time ordered sequence of perceptual events observed during the action. **(D)** Knowledge Base: Abstract pattern template for cover, along with the resulting state changes, and required initial state. **(E)** Relations encoded in the World State after the "cover" is recognized in a new action "covers the star with the block."

The key performance indices are (1) how many trials are required to learn an action with one set of objects, and (2) how well does this learning transfer to recognition of the same actions with different objects? Over the four experiments, a given action required 1.35 demonstrations to be learned accurately. This learning then transferred to new conditions on 70% of the new trials. Thirty percent of the trials required additional learning before the actions were recognized under new conditions. To quantify learning we performed an ANOVA on this data with three independent variables corresponding to Experiments (1–4), Experience (Naive, Experienced), and Condition (Training, Transfer). The dependant variable was the number of trials performed. The main observation was that once the system had learned the actions with one set of objects, it could transfer this knowledge to recognize the same actions with new object with little or no additional training. This was revealed by the significant interaction between Experience and Condition, independent of Experiment. Thus there is a significant effect of training on the ability to recognize new instances of learned action patterns ($p < 0.05$). Closer investigation revealed that in Experiment 2 the vision system was generating false movement recognition which lead to a number of irrelevant patterns being learned. When only Experiments 1, 3 and 4 are considered, an average of 1.13 trials are required for learning, and the knowledge transfers to 100% of the new trials with no additional learning.

In Experiments 5 and 6 we retained the Knowledge Base from Experiment 4, and then tested it with a new user, and examined the evolution over two complete tests with the five actions and the two object configurations. In Experiment 5, a total of six additional demonstrations were required to recognize the five actions in the two different object configurations. In Experiment 6 only one additional demonstration was required during the recognition of the 10 distinct actions. Overall these tests indicate that when the vision system is properly calibrated, the system is quite robust in the ability to learn generalized action recognition patterns.

Table 1 | Interaction fragment as robot learns meaning of "cover." Completion time (seconds).

LEARNING "COVER" (0:42)

```
12:22:22 User : watch <user then places the block on the sign>
12:22:31 Robot : Need to Learn this one
12:22:58 User : cover the sign with the block
12:23:00 Robot : Did you say cover sign block ?
12:23:02 User : yes
12:23:04 Robot : Adding cover sign block
```

LEARNING RESULTING PREDICATE "ON" (0:17)

```
12:23:08 Robot : Why are the sign and the block close one each other ?
12:23:12 User : because the block is on the sign
12:23:15 Robot : Did you say because block on sign ?
12:23:17 User : yes
12:23:25 Robot : What should I do now ?
```

TRANSFER "COVER" TO NEW SITUATION (0:16)

```
12:24:07 Robot : What should I do now ?
12:24:09 User : watch
12:24:21 Robot : Recognized cover star block
12:24:23 Robot : What should I do now ?
```

USER CAN INTERROGATE THE WORLD STATE:

```
12:24:26 User : Tell about star
12:24:28 Robot : I will query info about the star
12:24:30 Robot : star contact block
12:24:33 Robot : block on star
```

Table 2 | Enabling State – Action – Resulting State (SEASR) representation for the learned actions.

Enabling state	Action	Resulting state
Visible Arg1 Visible Arg2	Cover Arg1 with Arg2	Contact Arg1 Arg2 Visible Arg2 On Arg2 Arg1
Visible Arg2 True On Arg2 Arg1	Uncover Arg1 with Arg2	Visible Arg1 Visible Arg2
Visible Arg1 Visible Arg2	Give Arg1 to Arg2	Contact Arg1 Arg2 Visible Arg1 Visible Arg2 Has Arg2 Arg1
Contact Arg1 Arg2 Visible Arg1 Visible Arg2 On Arg2 Arg1	Take Arg1 from Arg2	Visible Arg1 Visible Arg2

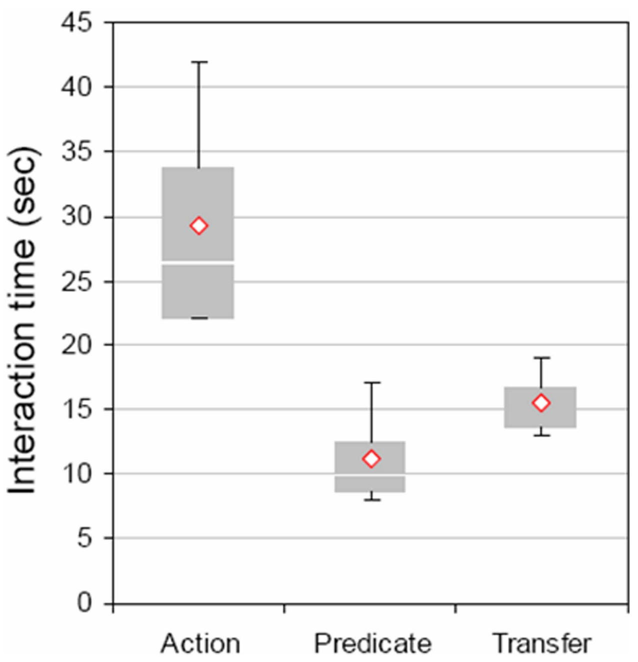


FIGURE 5 | Performance effects for learning and transfer. Interaction times were recorded for the learning of four novel actions, learning the derived predicates for those actions, and then using the resulting definitions to recognize new instances of the five actions (Action, Predicate, and Transfer, respectively). We compared interaction time for these three conditions in a single factor ANOVA. Recognition in transfer to new actions (15.5 s) was significantly faster than initial action learning time (29.25), $p < 0.05$.

Table 3 | Interaction fragment as robot explains the preconditions and results of the cover action.

```
13:43:10 User : What are the conditions for cover the sign with the block?
13:43:12 Robot : If you want to cover the block with the sign then
13:43:14 Robot : sign need to be visible
13:43:16 Robot : block need to be visible
13:43:23 User : What are the consequences of cover the sign with the block?
13:43:26 Robot : If you cover the sign with the block then
13:43:28 Robot : block will be contact with sign
13:43:30 Robot : sign will be contact with block
13:43:32 Robot : sign will no longer be visible
13:43:34 Robot : block will be on sign
```

THE BLOCK IS CURRENTLY ON THE SIGN.

```
13:43:40 User : How can I make the sign visible?
13:43:43 Robot : If you want to make the sign visible then
13:43:45 Robot : You can uncover the sign
```

DISCUSSION

Part of the stated objective of this work has been to implement, and demonstrate the advantages of, a hybrid embodied-teleological approach to action–language interaction, both from a theoretical perspective and via results from human–robot interaction experiments with the iCub robot. This objective was motivated by our observation that true cooperation requires not only that the robot can learn shared action sequences, but that it represents how those actions are linked in a chain of state changes that lead to the goal. This means that the robot must be able to represent actions in terms of the states that allow them to be performed, the states that result from their performance including the "unseen" predicates, for example, related to object permanence.

We developed a perceptual system that extracts patterns of spatio-temporal visual properties in order to encode actions in terms of these patterns. We re-discovered that action and meaning are not purely perceptual (Carey, 2009), and that additional properties related to object permanence and physical possession also form part of the meaning of action. Based on studies indicating that language can be used by toddlers to accelerate the acquisition of such knowledge (Bonawitz et al., 2009), when our cognitive system encounters unexpected results from an action, it interrogates the user, much like a developing child (Hood et al., 1979). This allows the user to explain, for example, that when the block covers the star, the star is not visible (but still there) *because* the block is *on* the star. We refer to these additional predicates (*on*, *has*) as derived predicates. This demonstrates that language can play an essential role in refining the representation of the meaning of action which is first approximated purely from the perceptual stream, by introducing derived predicates that become part of the meaning of the action. These predicates are encoded in the state changes that are to be introduced whenever the action is recognized. Thus, when the *give* and *take* actions are recognized, the derived predicate *has* (indicating possession) will be appropriately updated.

We believe that this is a fundamental development in the link between language and action, because it goes beyond a pure identity mapping between sentences and meaning, and instead uses language to change and enrich forever the meaning of action as part of a developmental/learning process. In this way, the use of language by the iCub to transfer knowledge to new trials is similar to the causal learning of toddlers observed by Bonawitz et al. (2009), as in both cases language is a symbolic processing tool for memory and cognition. This is consistent with theories of language in which words are not only considered as markers for referents in the world, but also as tools that allow us to reason and operate in the world (Borghi and Cimatti, 2009; Mirolli and Parisi, in press) as well as current ideas of how language evolved in humans through sensory-motor and social interaction, as well as possible simulations of these ideas in artificial systems (see Parisi, 2006; Parisi and Mirolli, 2006). These theories explain that language is used not only within the individual for reasoning and memory but also within a broader social network for communicative purposes. Therefore, our ongoing research on cooperative action (Dominey et al., 2009b; Dominey and Warneken, in press) is an important step in better understanding how language acts as a tool to facilitate goal-directed action between two or more agents.

A crucial component of the new system is the representation of actions which includes the link to initial enabling states, and final resulting states. The resulting system produces a Knowledge Base that encodes the representation of action meanings, and a World State that encodes the current state of the world. As mentioned above, we demonstrate how grammatical constructions that exploit causal connectives (e.g., because) can allow spoken language to enrich the learned set of SAS representations, by inserting derived predicates into the action definition. We also demonstrated how the causal connective "if–then" can be employed by the robot to inform the user about the links between enabling states and actions, and between actions and resulting states. Again, this extends the language–action interface beyond veridical action descriptions (or commands) to transmit more subtle knowledge about enabling and resulting states of actions, how to reach goals etc.

Indeed, in the context of the "hybrid" embodied and teleological system, we demonstrated how representations of enabling and resulting states provides the system with the knowledge necessary to make the link between goals as the resulting states of actions, and the intervening actions that are required. This is part of the basis of a teleological reasoning capability (Csibra, 2003). In the current system, we have not implemented a full blown reasoning capability, that can perform forward and backward chaining on the states and action representations. This is part of our ongoing research.

In Foundations of Language, Jackendoff (2002) indicates that while languages may vary in their surface structure, the organization of the conceptual structure that they express appears more universal (see Fadiga et al., 2009 for a review of common syntactic structure in language, action and music). We extended this notion to consider that indeed, the compositional structure of syntax is derived from that of the conceptual system (Dominey, 2003), and Jackendoff (2003) agreed. In this context, one of the most promising results of the current research is the continued observation that language reflects the structure of conceptual representations. We have previously demonstrated this in situations where multiple actions are linked by shared states, resulting in descriptions such as "Larry took the toy that Robert uncovered with the box" (Dominey and Boucher, 2005). The current work extends this to include functional and causal links between elements in the SAS representations (e.g., the if–then constructions in **Tables 3 and 4**).

The experiments described here have focused on perception of events. We realize that action is crucial to event understanding, and we are currently working to integrate this $\mathrm{S_EAS_R}$ framework into our existing cooperative action framework (Dominey et al., 2009b; Dominey and Warneken, in press). We will first demonstrate that the mechanism presented here for learning

Table 4 | Interaction fragment where the robot performs basic goal-based reasoning about how to solve the task of getting the star when it is covered by the block, and then uses it to cover the sign.

```
18:03:50 Robot : Recognized cover star with block
18:04:05 User : What are the conditions to cover the sign with star
18:04:07 Robot : If you want to cover the sign with the star then
18:04:09 Robot : sign needs to be visible
18:04:11 Robot : star needs to be visible
The sign is visible, but the star is covered with the block, so the user asks the robot how to make the sign visible:
18:04:18 User : How can I see the star
18:04:23 Robot : You can uncover the star
The user can now uncover the star by removing the block, and then achieve the final goal which is to cover the sign with the star
```

the perceptual patterns associated with perceived actions can be applied to learning motor patterns associated with executed actions. This will result in further enriched action representations that include the enabling and resulting states, the perceptual primitive patterns, and the action primitive patterns. We will then use these representations in the context of learning cooperative tasks by observation. This will yield a situation in which the robot can represent the trajectory from initial state to final goal state via coordinated action sequence, and will thus provide the basis for intentional reasoning, and the extension of the teleological reasoning to cooperative activity.

ACKNOWLEDGMENTS

This research is supported by the European Commission FP7 ICT Projects CHRIS and Organic, and French ANR Projects Amorces and Comprendre.

REFERENCES

Barsalou, L. W. (1999). Perceptual symbol systems. *Behav. Brain Sci.* 22, 577–660.

Barsalou, L. W., Santos, A., Simmons, W. K., and Wilson, C. D. (2008). "Language and simulation in conceptual processing," in *Symbols, Embodiment, and Meaning*, eds M. De Vega, A. M. Glenberg and A. C. Graesser (Oxford, UK: Oxford University Press), 245–284.

Bekkering, H., Wohlschlager, A., and Gattis, M. (2000). Imitation of gestures in children is goal-directed. *Q. J. Exp. Psychol.* 53, 153–164.

Bergen, B., and Chang, N. (2005). "Embodied construction grammar in simulation-based language understanding," in *Construction Grammar(s): Cognitive Grounding and Theoretical Extensions*, eds J.-O. Östman and M. Fried (Amsterdam: John Benjamins), 147–156.

Bonawitz, E. B., Horowitz, A., Ferranti, D., and Schulz, L. (2009). "The block makes it go: causal language helps toddlers integrate prediction, action, and expectations about contact relations," in *Proceedings of the Thirty-first Cognitive Science Society*, eds N. Taatgen and H. van Rijn (Amsterdam: Cognitive Science Society), 81–86.

Borghi, A. M., and Cimatti, F. (2009). "Words as tools and the problem of abstract word meanings," in *Proceedings of the Thirty-first Cognitive Science Society*, eds N. Taatgen and H. van Rijn (Amsterdam: Cognitive Science Society).

Bowerman, M. (1974). Learning the structure of causative verbs: a study in the relationship of cognitive, semantic, and syntactic development. *Proc. Res. Child Lang. Dev.* 8, 142–178.

Brass, M., Schmitt, R. M., Spengler, S., and Gergely, G. (2007). Investigating action understanding: inferential processes versus action simulation. *Curr. Biol.* 17, 2117–2121.

Carey, S. (2009). *The Origin of Concepts*. New York: Oxford University Press.

Csibra, G. (2003). Teleological and referential understanding of action in infancy. *Phil. Trans. R. Soc. Lond. B.* 358, 447–458.

Decety, J., and Grèzes, J. (2006). The power of simulation: imagining one's own and other's behavior. *Brain Res.* 1079, 4–14.

Decety, J., and Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: how low-level computational processes contribute to meta-cognition. *Neuroscientist* 13, 580–593.

Dominey, P. F. (2003). A conceptuocentric shift in the characterization of language. *Comment on Jackendoff, BBS*, 674–675.

Dominey, P. F., and Boucher, J. D. (2005). Learning to talk about events from narrated video in a construction grammar framework. *Artif. Intell.* 167, 31–61.

Dominey, P. F., Hoen, M., Blanc, J. M., and Lelekov-Boissard, T. (2003). Neurological basis of language and sequential cognition: evidence from simulation, aphasia, and ERP studies. *Brain Lang.* 86, 207–225.

Dominey, P. F., Inui, T., and Hoen, M. (2009a). Neural network processing of natural language: II. Towards a unified model of corticostriatal function in learning sentence comprehension and non-linguistic sequencing. *Brain Lang.* 109, 80–92.

Dominey, P. F., Mallet, A., and Yoshida, E. (2009b). Real-time spoken-language programming for cooperative interaction with a humanoid apprentice. *Int. J. HR* 6, 147–171.

Dominey, P. F., and Warneken, F. (in press). The origin of shared intentions in human–robot cooperation. *New Ideas Psychol.*

Dove, G. (2009). Beyond conceptual symbols. A call for representational pluralism. *Cognition* 110, 412–431.

Fadiga, L., Craighero, L., and D'Ausilio, A. (2009). Broca's area in language, action, and music. The neurosciences and music III – disorders and plasticity. *Ann. N. Y. Acad. Sci.* 1169, 448–458.

Fern, A., Givan, R., and Siskind, J. M. (2002). Specific-to-general learning for temporal events with application to learning event definitions from video. *J. Artif. Intell. Res.* 17, 379–449.

Fischer, M. H., and Zwaan, R. A. (2008). Embodied language: a review of the role of the motor system in language comprehension. *Q. J. Exp. Psychol.* 61, 825–850.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., and Rizzolatti, G. (2005). Parietal lobe: from action organization to intention understanding. *Science* 308, 662–667.

Fontanari, J. F., Tikhanoff, V., Cangelosi, A., Ilin, R., and Perlovsky, L. I. (2009). Cross-situational learning of object–word mapping using Neural Modeling Fields. *Neural Netw.* 22, 579–585.

Gergely, G., and Csibra, G. (2003). Teleological reasoning in infancy: the naïve theory of rational action. *Trends Cogn. Sci.* 7, 287–292.

Glaser, W. R. (1992). Picture naming. *Cognition* 42, 61–105.

Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Hauk, O., Shtyrov, Y., and Pulvermüller, F. (2008). The time course of action and action-word comprehension in the human brain as revealed by neurophysiology. *J. Physiol. Paris* 102, 50–58.

Hauser, M., and Wood, J. (2010). Evolving the capacity to understand actions, intentions, and goals. *Annu. Rev. Psychol.* 61, 303–324.

Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. *Nat. Rev. Neurosci.* 8, 393–402.

Hoen, M., Pachot-Clouard, M., Segebarth, C., and 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.

Hood, L., Bloom, L., and Brainerd, C. J. (1979). What, when, and how about why: a longitudinal study of early expressions of causality. *Monogr. Soc. Res. Child Dev.* 44, 1–47.

Jackendoff, R. (2002). Foundations of Language: brain, Meaning, Grammar, Evolution. Oxford University Press.

Jackendoff, R. (2003). Précis of Foundations of Language: brain, Meaning, Grammar, Evolution, BBS 26, 651–707.

Kan, I. P., Barsalou, L. W., Solomon, K. O., Minor, J. K., and Thompson-Schill, S. L. (2003). Role of mental imagery in a property verification task: fMRI evidence for perceptual representations of conceptual knowledge. *Cogn. Neuropsychol.* 20, 525–540.

Kotovsky, L., and Baillargeon, R. (1998). The development of calibration-based reasoning about collision events in young infants. *Cognition* 67, 311–351.

Lallee, S., Warkeken, F., and Dominey, P. F. (2009). "Learning to collaborate by observation and spoken language," in *Ninth International Conference on Epigenetic Robotics*, Venice.

Leslie, A. M., and Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition* 25, 265–288.

Madden, C. J., Hoen, M., and Dominey, P. F. (2010). A cognitive neuroscience perspective on embodied language for human–robot cooperation. *Brain Lang.* 112, 180–188.

Mandler, J. M. (2008). On the birth and growth of concepts. *Philos. Psychol.* 21, 207–230.

Mason, R. A., and Just, M. A. (2006). "Neuroimaging contributions to the understanding of discourse processes," in *Handbook of Psycholinguistics*, eds M. Traxler and M. A. Gernsbacher (Amsterdam: Elsevier), 765–799.

Meltzer, J. A., McArdle, J. J., Schafer, R. J., and Braun, A. R. (in press). Neural aspects of sentence comprehension: syntactic complexity, reversibility, and reanalysis. *Cereb. Cortex.*

Michotte, A. (1963). *The Perception of Causality*. New York: Basic Books.

Mirolli, M., and Parisi, D. (in press). Towards a Vygotskyan cognitive robotics: the role of language as a cognitive tool. *New Ideas Psychol.*

Obleser, J., Zimmermann, J., Van Meter, J., and Rauschecker, J. P. (2007). Multiple stages of auditory speech perception reflected in event-related FMRI. *Cereb. Cortex* 17, 2251–2257.

Parisi, D. (2006). "Simulating the evolutionary emergence of language: a research agenda," in *The Evolution of Language*, eds A. Cangelosi, A. D. M.

Smith and K. Smith (Singapore: World Scientific), 230–238.

Parisi D., and Mirolli M. (2006). "The emergence of language: how to simulate it," in *Emergence of Communication and Language*, eds C. Lyon, C. Nehaniv and A. Cangelosi (London: Springer), 269–285.

Pulvermüller, F., Shtyrov, Y., and Hauk, O. (2009). Understanding in an instant: neurophysiological evidence for mechanistic language circuits in the brain. *Brain Lang.* 110, 81–94.

Rizzolatti, G., and Fabbri-Destro, M. (2008). The mirror system and its role in social cognition. *Curr. Opin. Neurobiol.* 18, 179–184.

Saur, D., Kreher, B. W., and Schnell, S., et. (2008). Ventral and dorsal pathways for language. *Proc. Natl. Acad. Sci. U. S. A.* 105, 18035–18040.

Saxe, R., Xiao, D. K., Kovacs, G., Perrett, D. I., and Kanwisher, N. (2004). A region of right posterior superior temporal sulcus responds to observed intentional actions. *Neuropsychologia* 42, 1435–1446.

Scott, S. K., Rosen, S., Lang, H., and Wise, R. J. (2006). Neural correlates of intelligibility in speech investigated with noise vocoded speech – a positron emission tomography study. *J. Acoust. Soc. Am.* 120, 1075–1083.

Siskind, J. M. (2001). Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *J. Artif. Intell. Res.* 15, 31–90.

Siskind, J. M. (2003). Reconstructing force-dynamic models from video sequences. *Artif. Intell.* 151, 91–154.

Sommerville, A., and Woodward, A. L. (2005). Pulling out the intentional structure of action: the relation between action processing and action production in infancy. *Cognition* 95, 1–30.

Speer, N. K., Zacks, J. M., and Reynolds, J. R. (2007). Human brain activity time-locked to narrative event boundaries. *Psychol. Sci.* 18, 449–455.

Spelke, E. S. (1990). Principles of object perception. *Cogn. Sci.* 14, 29–56.

Spelke, E. S., and Kinzler, K. D. (2007). Core knowledge. *Dev. Sci.* 10, 89–96.

Talmy, L. (1988). Force dynamics in language and cognition. *Cogn. Sci.* 12, 49–100.

Tettamanti, M., Buccino, G., Saccuman, M. C., Gallese, V., Danna, M., Scifo, P., Fazio, F., Rizzolatti, G., Cappa, S. F., and Perani, D. (2005). Listening to action-related sentences activates fronto-parietal motor circuits. *J. Cogn. Neurosci.* 17, 273–281.

Tomasello, M. (2008). *Origins of Human Communication*. Cambridge: MIT Press.

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: the origins of cultural cognition. *Behav. Brain Sci.* 28, 675–691.

Tikhanoff, V., Cangelosi, A., and Metta, G. (2009). "Language understanding in humanoid robots: simulation experiments with iCub platform", in *Proceedings of 2009 International Conference on Integration of Knowledge Intensive Multi-Agent Systems (KIMAS'09)*, St. Louis, MA.

Tsagarakis, N. G., Metta, G., Sandini, G., Vernon, D., Beira, R., Becchi, F., Righetti, L., Victor, J. S., Ijspeert, A. J., Carrozza, M. C., and Caldwell, D. G. (2007). iCub – the design and realization of an open humanoid platform for cognitive and neuroscience research. *Adv. Robot.* 21, 1151–1175.

Ullman, M. T. (2004). Contributions of memory circuits to language: the declarative/procedural model. *Cognition* 92, 231–270.

Wible, C. G., Preus, A. P., and Hashimoto, R. (2009). A cognitive neuroscience view of schizophrenic symptoms: abnormal activation of a system for social perception and communication. *Brain Imaging Behav.* 3, 85–110.

Wolff, P. (2007). Representing causation. *J. Exp. Psychol. Gen.* 136, 82–111.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition* 69, 1–34.

Zwaan, R. A., and Madden, C. J. (2005). "Embodied sentence comprehension," in *Grounding cognition: The Role of Perception and Action in Memory, Language, and Thinking*, eds D. Pecher and R. A. Zwaan (Cambridge, UK: Cambridge University Press), 224–245.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 December 2009; paper pending published: 20 February 2010; accepted: 11 May 2010; published online: 03 June 2010.

Citation: Lallee S, Madden C, Hoen M and Dominey PF (2010) Linking language with embodied and teleological representations of action for humanoid cognition. Front. Neurorobot. ***4****:8. doi: 10.3389/fnbot.2010.00008*