Robotics **2015**, *4*, 169-193; doi:10.3390/robotics4020169

OPEN ACCESS

robotics

ISSN 2218-6581

www.mdpi.com/journal/robotics

Article

How? Why? What? Where? When? Who? Grounding Ontology in the Actions of a Situated Social Agent

Stephane Lallee [1,2,*] and Paul F.M.J. Verschure [1,3]

[1] Laboratory of Synthetic Perceptive, Emotive and Cognitive Systems (SPECS), Center of Autonomous Systems and Neurorobotics, Universitat Pompeu Fabra, Roc Boronat 138, Barcelona 08018, Spain; E-Mail: paul.verschure@upf.edu

[2] Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis 138632, Singapore

[3] Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Llúis Companys 23, Barcelona 08010, Spain

***** Author to whom correspondence should be addressed; E-Mail: stephane-l@i2r.a-star.edu.sg; Tel.: (+65) 85 00 45 40.

Academic Editors: Nicola Bellotto, Nick Hawes, Mohan Sridharan and Daniele Nardi

Received: 28 February 2015 / Accepted: 8 May 2015 / Published: 10 June 2015

Abstract: Robotic agents are spreading, incarnated as embodied entities, exploring the tangible world and interacting with us, or as virtual agents crawling over the web, parsing and generating data. In both cases, they require: (i) processes to acquire information; (ii) structures to model and store information as usable knowledge; (iii) reasoning systems to interpret the information; and (iv) finally, ways to express their interpretations. The H5W (*How*, *Why*, *What*, *Where*, *When*, *Who*) framework is a conceptualization of the problems faced by any agent situated in a social environment, which has defined several robotic studies. We introduce the H5W framework, through a description of its underlying neuroscience and the psychological considerations it embodies, we then demonstrate a specific implementation of the framework. We will focus on the motivation and implication of the pragmatic decisions we have taken. We report the numerous studies that have relied upon this technical implementation as a proof of its robustness and polyvalence; moreover, we conduct an additional validation of its applicability to the natural language domain by designing an information exchange task as a benchmark.

Keywords: knowledge representation; human robot interaction; communication; natural language processing; perception action loop; artificial cognitive architecture

1. Introduction

We have reached the age of information. Inhabitants of developed countries are carrying with them devices that embed a large collection of sensors and which constantly send and receive information. For most people they currently take the form of smartphones and of wearable devices for early technology adopters. However, the Internet of Things community [1–3] forecasts an acceleration of the distribution of connected objects. These devices will sense the world, adding to the global data deluge, but at the same time interpreting this globally available information to provide end-users with the essential details they are looking for. One of the challenges of such an ecosystem lies in the analysis of this "big data" [4–6], and the mainstream approach that seems to be on the side of brute force, by joining large computational power and improved, but still not fully understood, connectionist learning algorithms, such as deep learning [7]. However, the recent success of this method, applied to multiple perceptual modalities (e.g., image recognition [8], speech recognition [9,10]), only demonstrate one thing: we do not have a generic, formal, and well understood architecture for representing knowledge in artificial systems. Indeed, among the extensive number of smart objects developed, the most advanced instances of those objects embed both sensors and actuators, they gather data, but they are also situated in the world and act on it with their bodies: They are robots.

The previously mentioned machine learning systems partially solve one of the large problems artificial intelligence and robotic research is yet to overcome, namely the symbol grounding or anchoring problem [11–13], which asks how to relate sensor data to symbolic representations. For example, displaying a rectangle and a label around a cat in an image, or generating a string representation from a sound wave are not enough for a robot to become an 'intelligent machine'. Being able to segment and classify sensory streams is a given, but is not enough for an agent to autonomously act in the world. Indeed, living beings monitor the world so that they can make decisions that will directly affect them, with the ultimate purpose of optimizing their own fitness. Robots are much like living machines in this respect: They need not only to perceive the world, but to understand it as a coherent scene that they are part of and that they can alter on the basis of their own goals [14]. In order to do so, a consideration of the dynamics of the world is required, through recognition of objects and agents in the scene, as well as how they act and interact. Formalizing the content and evolution of a scene requires the combination of perceptual, symbolic and rule based reasoning in a single unified framework and is a topic of increasing interest [15]. Such processes will generate information about who is acting, what they are doing, where and when it happens, and this will give cues about why it is happening. Those five interrogative words (who, what, where, when, why), together with "how", which describes the action itself, have been argued to be the main questions any conscious being has to answer in order to survive in the world [16,17]. They form the theoretic problem known as H5W, which is an elegant framework for modeling the knowledge generated by the scene-understanding processes. The Distributed Adaptive Control architecture (DAC) is a biologically and psychologically grounded cognitive architecture that intrinsically solves the H5W problem [14,18,19].

DAC, in a very specific way, organizes the different perceptual, emotional, cognitive and motor processes of a situated embodied agent interacting with the world; it provides a control system for these agents and, at the same time, gives insight in biological cognition using the principles of convergent validation and synthesis [20]. The former states that we gain validity of models by seeing

them as lenses through which data from different sources convergence, *i.e.*, anatomy, physiology and behavior, while the latter can be interpreted as seeing the embodied models we construct as theories in their own right [14].

DAC and H5W have been widely applied to a range of robotic modeling aspects, such as insects (e.g., ants, locusts and moths) and mammals (e.g., rodents and primates). The models interacts with the physical world by replicating robot foraging and the social world by looking at dyadic interaction. Here we look in particular at the question of whether the DAC architecture can be generalized to the control of the dyadic social interaction of a humanoid robot [21]. The core of this new implementation is a formalized knowledge representation, specifically tailored to answer the H5W problem. The whole architecture in this case is a complex distributed collection of modules dealing with the interpretation of multiple sensors, reasoning and motor control planning in an attempt to model the complete perception-action loop. By relying on this global machinery, we have been able to give a robot the ability to autonomously sense and interpret an environment and to be able to react and adapt to it accordingly. At the same time, the architecture provides useful insight into the robots internal state and motivations, which is due to the careful design the knowledge representation. Our implementation of the architecture has already been presented and validated in several previous studies [21–23]; In particular we have investigated the impact of how different robot behaviors are perceived by a human observer or partner.

While the H5W architecture itself is the well-defined collection and interaction of hardware and software components, the representation of knowledge in the system is a more vaporous and unreported concept. All the modules in the architecture rely on this representation of knowledge to communicate and exchange information and its role is therefore central in the success of the system. Despite the care taken in formalizing and implementing the ontology in our system (*i.e.*, the H5W framework), the relative lack of experimental methods available to test the efficiency of knowledge representations has led to an absence of reporting on this particular aspect. We believe that this is a common issue faced by robotic systems used for human robot interaction (HRI), as the focus is more about the interaction itself, *i.e.*, how it is generated, rather than about the representation which sustains this interaction. We demonstrate the importance of knowledge representation in robotic system and so we hope that this Special Issue will in part address the current lack of reporting on knowledge representation. In this paper, we firstly give a short overview of the DAC architecture and then we present an implementation of the H5W framework, to represent the situated knowledge of a robot in the context of social interaction. Additionally, we will consider and benchmark how the represented knowledge can be expressed in natural language form, therefore allowing it to be transferred between agents. We will focus not only on the knowledge representation *per se*, but also on how this representation can support the processes of gathering and sharing the information between the machine and a human.

2. DAC Overview

The Distributed Adaptive Control architecture (DAC) is the biologically and psychologically grounded cognitive architecture that we used as an engine to solve the H5W problem. This section aim is to provide an overview of the organization of DAC, which is graphically described in Figure 1.

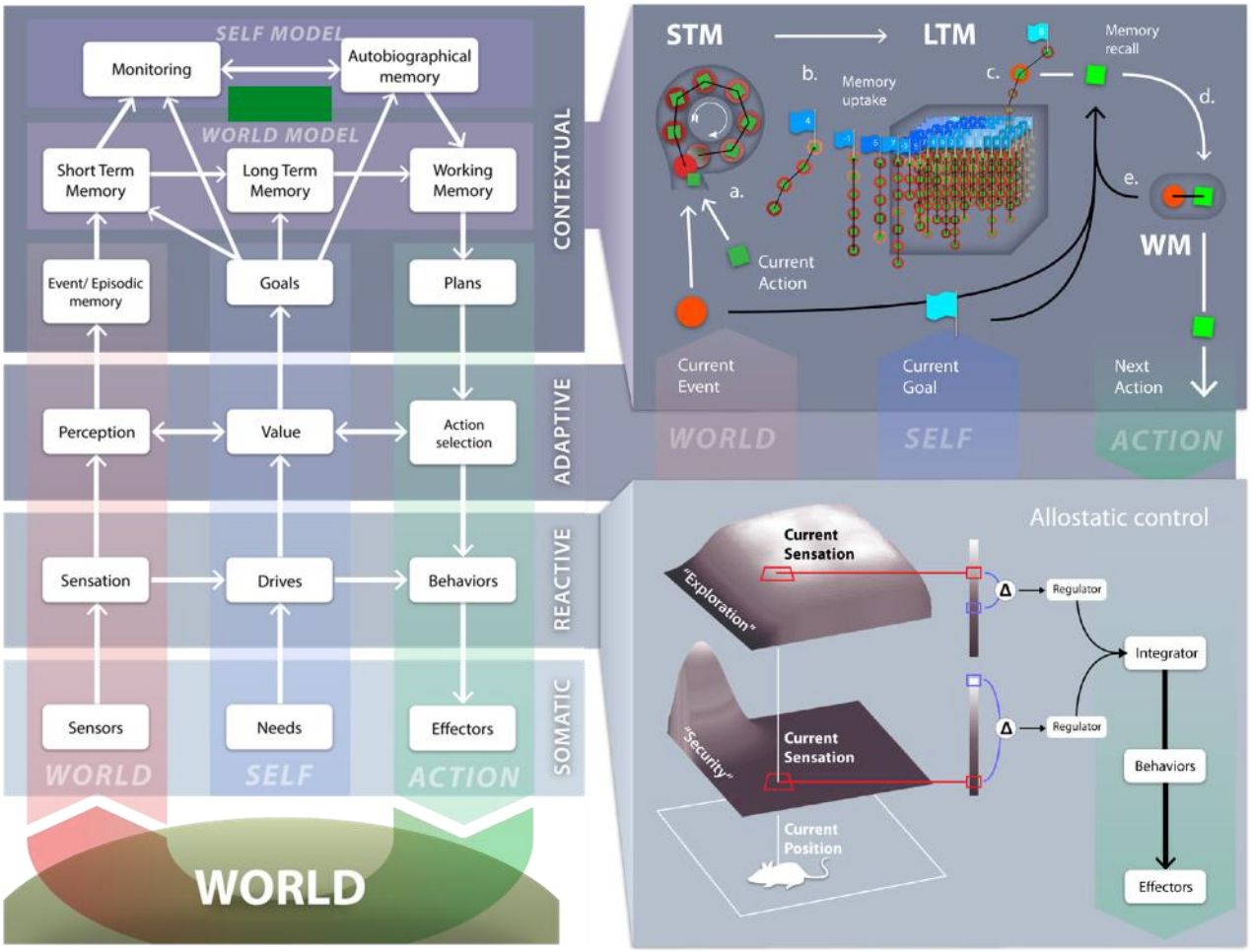

Figure 1. Graphical representation of the Distributed Adaptive Control (DAC) theory of mind and brain. See text for further information. (Adapted from [24]).

Figure 1 Left: Highly abstract representation of the DAC architecture. DAC proposes that the brain is organized as a three-layered control structure with tight coupling within and between these layers distinguishing: The Soma (SL) and the Reactive (RL), Adaptive (AL) and Contextual (CL) layers. Across these layers a columnar organization exists, that deals with the processing of states of the world or exteroception (left, red column), the self or interoception (middle, blue) and action (right, green column) that mediates between the first two. The soma designates the body with its sensors, organs and actuators. The Reactive Layer (RL) comprises dedicated Behavior Systems (BS) that combine predefined sensorimotor mappings with drive reduction mechanisms that are predicated on the needs of the body (SL).

Figure 1 Right lower panel: Each BS follows homeostatic principles supporting the Self Essential Functions (SEF) of the body (SL). In order to map needs into behaviors, the essential variables served by the BSs have a specific distribution in space called an 'affordance gradient'. In this example we consider the (internally represented) "attractive force" of the home position supporting the Security SEF or of open space defining an Exploration SEF. The values of the respective SEFs are defined by the difference between the sensed value of the affordance gradient (red) and its desired value given the

prevailing needs (blue). The regulator of each BS defines the next action as to perform a gradient ascent on the SEF. An integration and action selection process across the different BSs and forces a strict winner-take-all decision selection that defines the specific behavior emitted. The allostatic controller of the RL regulates the internal homeostatic dynamics of the BSs to set priorities defined by needs and environmental opportunities through the modulation of the affordance gradients, desired values of SEFs and/or the integration process. The Adaptive Layer (AL) acquires a state space of the agent-environment interaction and shapes action. The learning dynamics of AL is constrained by the SEFs of the RL that define value. The AL crucially contributes to exosensing by allowing the processing of states of distal sensors, e.g., vision and audition, which are not predefined but rather are tuned in somatic time to properties of the interaction with the environment. Acquired sensor and motor states are in turn associated through the valence states signaled by the RL. The AL is modeled after the paradigm of classical conditioning and the acquisition of the sensory-motor state space is based on predictive mechanisms, in order to optimize encoding and counteract biases due to behavioral feedback. The AL has been mapped to the cerebellum, amygdala, cortex and hippocampus. The contextual layer (CL) is divided between a world and a self-model. The CL expands the time horizon in which the agent can operate through the use of sequential short and long-term memory systems (STM and LTM respectively). These memory systems operate on the integrated sensorimotor representations that are generated by the AL and acquire, retain and express goal-oriented action regulated by the RL.

Figure 1 Right upper panel: The CL comprises a number of interlocked processes: a: When the error between predicted and encountered sensory states falls below a STM acquisition threshold, the perceptual predictions (red circle) and motor activity (green rectangle) generated by AL are stored in STM as (**a**), so called, segment. The STM acquisition threshold is defined by the time averaged reconstruction error of the perceptual learning system of AL. (**b**): If a goal state (blue flag) is reached, e.g., reward or punishment, the content of STM is retained in LTM as a sequence conserving its order, goal state and valence marker, e.g., aversive or appetitive, and STM is reset so new sequences can be acquired. Every sequence is thus defined through sensorimotor states and labeled with respect to the specific goal it pertains to and its valence marker. (**c**): If the outputs generated by the RL and AL to Action Selection are sub-threshold, the CL realizes its executive control and perceptual predictions generated by AL are matched against those stored in LTM. (**d**): The CL selected action is defined as a weighted sum over the segments of LTM. (**e**): The contribution of LTM segments to decision-making depends on four factors: Perceptual evidence, memory chaining, the distance to the goal state, and valence. The working memory (WM) of the CL is defined by the memory dynamics that represents these factors. Active segments in WM that contributed to the selected action are associated with those that were previously active establishing rules for future chaining. The core features of CL have been mapped to the prefrontal cortex. The self-model component of the CL monitors task performance and develops (re)descriptions of task dynamics anchored in the Self. In this way the system generates meta-representational knowledge that forms autobiographical memory.

3. Ontology of H5W

3.1. H5W Definition: How, Who, What, When, Where, Why?

Humans have always been elaborating ways to represent, save and share their mental constructs. Through drawing, then through spoken language and writing we have been trying to crystalize our experience of the world as to communicate it with others. Philosophers back to Plato and Aristotle have been studying our knowledge, how it is created and how it links to the world, in an effort to understand the nature of reality. This philosophical branch, known as ontology, led to one of the main ways of formalizing the human knowledge through the establishment of entities and relations among them. More concretely, the World Wide Web Consortium (W3C) issued a recommendation (Web Ontology Language (OWL) [25]) and several databases representing various domains of knowledge are available [26–29]. As a result, ontologies appears to be a representation of choice for robots as they are systems which should reason about facts and using some knowledge base, take decisions. This approach has been quite successful in certain cases such as symbolic reasoning and planning [30,31], or in the generation of a limited dialog between a machine and a human [32]. However, in most cases ontology is considered as a tool, a collection of recipes that a robot can browse in order to achieve a given task. In the context of a community of robots connected to the same ontology, this is a very efficient way of representing a common knowledge base that can grow fast as each robot can contribute its own experience [33]. The main concern with the traditional ontological approach lies in its claims to solve the knowledge representation problem from a pure symbolic standpoint: While being a very efficient tool to represent symbolic knowledge, it is far from solving the anchoring problem.

Robots which operate in the physical world do not need only to reason about symbols and express statements about them: They need to perceive the sensory representation that those symbols represent, and they need to be able to act upon them. Providing this full variety of skills, from anchoring perception, to generating actions, is the purpose of the many cognitive architectures that have been developed for robotics over the last decades [34]. A pure ontological representation could lead to the illusion of world understanding, but it would still suffer the caveat pointed by Searle in his famous Chinese room example [35]. However, if a robot was to build its own ontological representation, grounded in the sensorimotor experience of reality, it would approach a more human-like representation of the world. Doing so requires that the machine experiences the physical world. It needs to perceive the world, to act and, through time, to understand the consequences of its actions so it can start to reason about its goals and how to achieve them [36]. Moreover, it requires an intrinsic motivation to act, some primary force that would drive its actions in a reactive way even before any knowledge is acquired from experience. Different communities have different, but not incompatible, views on what could be the source of agents motivation ranging from the innate desire to learn [37] to hierarchies of drives inherited from survival and evolution [38,39]. For an overview of the different approaches, refer to a recent survey on the specific topic of motivation in artificial systems [40]. Once given some motivation and basic exploratory behaviors, the agent will start experiencing the world and its sensory-motor contingencies. As a result, it will witness a scene of agents, including itself, and objects

interacting in various manners, times and places. It will start to ask and answer six fundamental questions about its surrounding environment:

- Who is there? (subject)
- What is there? (object)
- How they behave? (action/verb)
- When it happens? (time)
- When it take place? (place)
- Why do they behave like this? (motivation/causality)

Taken together, those six questions form the H5W problem which has been introduced and discussed by Verschure as the main concerns that live beings are facing when trying to survive in a world filled with other agents [41]. Indeed, the immediate knowledge of agents situated in the world gravitates around those questions as they provide a generic way of describing any event in a scene. From a pragmatic point of view, they also form a constrained type of ontology by linking several symbols through multiple semantic relationships. Additionally, they have a direct mapping to language through the obvious link with the grammar of an English sentence. For those reasons, the H5W problem is well suited for defining the knowledge representation of a robot by grounding a powerful symbolic formalism (ontology) within the reality faced by an embodied machine. In the following section, we will demonstrate a simple way of representing the entities which populate the world as well as the relations among them, therefore providing the core representation of a much larger cognitive architecture.

3.2. The H5W Data-Structures

As stated before, any ontology tries to establish the links (relations) between a set of terms, or entities. Entities are the possible concepts being manipulated and therefore represent the core material of the knowledge, while the relations define how those concepts relate to each other. The most common relation in traditional ontologies is "is-a", which defines an inheritance between two entities (e.g., "water is-a liquid") and therefore split the terms into a hierarchy of classes. Our representation takes a similar approach by considering that all concepts are "entities", and that a specific situation or fact is described by a set of relations between those entities. We adopted the substrate of a software library (C++) for modeling the H5W framework; we therefore took advantage of the object oriented aspect of the language to incorporate directly the ontology representation into the programing method (see the class diagram in Figure 2). Concretely, this means that the user of the library can define objects of the class "Entity" which represent the concepts that the artificial can manipulate. Another class, the "Relation", allows linking several instances of "Entity" together, therefore implementing the traditional ontology model. However, as we are to model the knowledge representation of a situated agent, we included additional constraints given by the structure of the world. We therefore defined *a priori* strongly subtyped entities (Figure 2) which hold noticeable properties and ground the knowledge domain in the context of a cognitive agent development. The following sections will give details about the "Relation" data structure and the different subclasses of "Entity" we chose to impose. We will focus on their specifics and the reasons behind our choices.

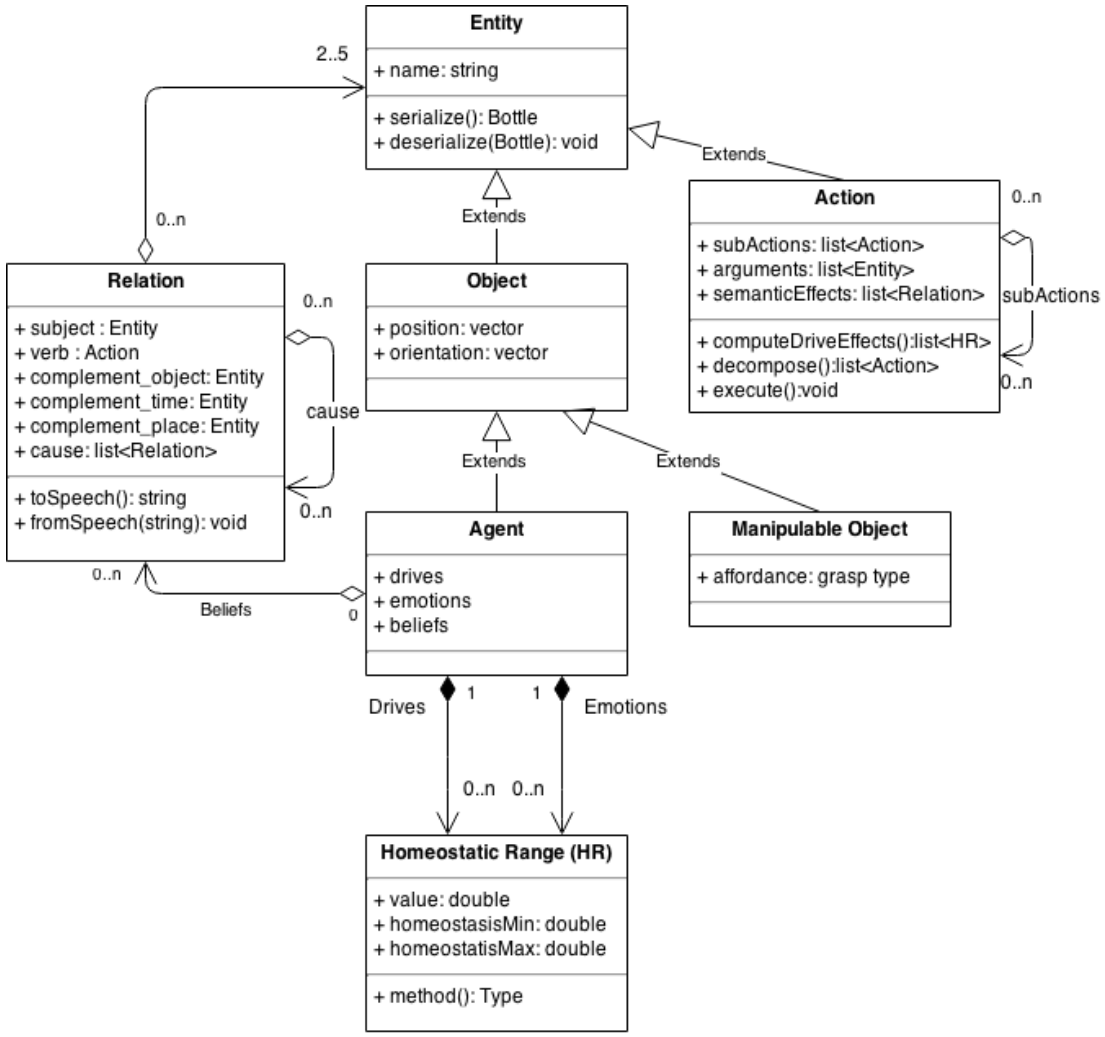

Figure 2. Class diagram of the H5W implementation. The H5W per se is defined by the class Relation, which aggregate a set of entities. Those entities can be of specific types, including the Agent type which in turn can aggregate a list of Relation to model its beliefs.

3.2.1. Relation

The Relation idea is the core of the H5W framework. It links together up to 5 entities and assigns them with semantic roles to form a solution to the H5W problem. The common way of representing relations in ontology is an edge linking two terms/nodes; it allows consideration of ontology as a graph, therefore allows the powerful algorithms of graph theory to be applied. However, it is rather limiting as more complex relations involving multiple nodes would require additional meta-data such as edge labeling. An alternative to the simple graph-oriented approach is to consider multiplex graphs [42] that allows us to define a relation as a set of edges with roles. If we consider the knowledge of our agent as a set of entities representing the concepts it knows (*i.e.*, the object, agents, actions and more generally any word), then those entities can be modeled as the nodes of a graph. We define a Relation as a set of 5 edges connecting those nodes in a directed and labeled manner. The labels of those edges are chosen so that the Relation models a typical sentence from the English grammar of the form: Relation → Subject Verb [Object] [Place] [Time]

The brackets indicate that the complements are facultative; the minimal relation is therefore composed of two entities representing a subject and a verb. In this respect, our framework assumes that verbs have a strictly positive valency (*i.e.*, they have a least one argument which is their subject) and does not account for languages with avalent verbs such as Mandarin Chinese. Considering that we want to represent the interaction between agents and their environment and that our focus is more on the representation of meaning than on the linguistic expression of it, we consider this assumption as acceptable, moreover the reason of this choice is also pragmatic: We consider a relation without a subject as an imperative form, it therefore describes an order (see Section 4. The H5W acquisition and transfer through dialog). We also impose a constraint on the specific types of entity (defined below) which can fill each role: The verb has to be an Action while the other roles have to be anything but an Action.

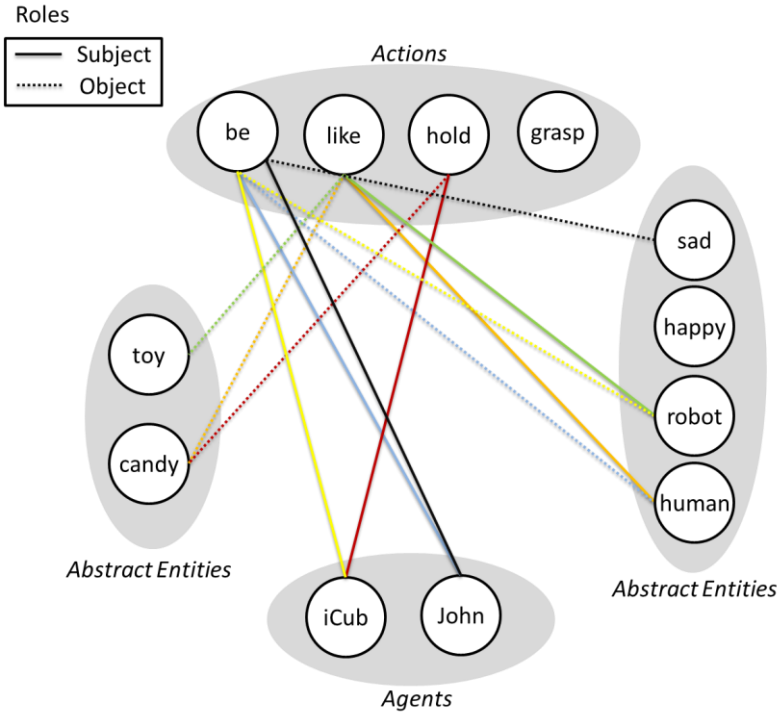

Figure 3. Graphical view of the multiplex graph representation of the knowledge. Each node is an Entity, edges of different colors represent different relations, the type of line represent the semantic role of connected entities in the relation. We used only simple relations of the form (subject, verb, object) for clarity. Although multiplex graph layout is quickly messy, they can be used to capture complex semantic links between concepts. In this example the relations are *R (iCub, be, robot), R (John, be, human), R (robot, like, toy), R (human, like, candy*), *R (iCub, hold, candy*), *R (John, be, sad*). Such a description of the situation can easily be used by any expert system or planner (e.g in this case the robot could decide to make John happy by giving him the candy).

By linking those entities and binding them with roles within a relation, we create a powerful way to define a fact which describes the world state or its evolution (e.g., a state: *"iCub, be, robot";* an event: *"iCub, push, ball";* a more complex event: *"iCub crash, null, laboratory, yesterday").* By defining

several relations over the same set of entities, we build the knowledge of the agent as a multiplex graph which represents the various semantic links among entities (see Figure 3). The case of the "Why" question is a bit more complex, as the causality of a given statement can refer to a complex collection of conditions, the approach we have taken so far is to consider that the "why" of a relation, if specified, is a pointer to set of other relations (e.g., "*John, be, sad*" is a consequence of "*John, like, candy*" and "*iCub, hold, candy*"). It is harder to conceptualize within graph theory, even considering multiplex graphs as it would mean that a relation (which is a set of edges) would point to other set of edges. Instead, we can build up another graph representation in which case relations act as nodes, and the "why" part of the relations are edges within this graph (see Figure 4). Using this standpoint, forward and backward chaining processes of planning algorithms generate a path of connected relations along this graph.

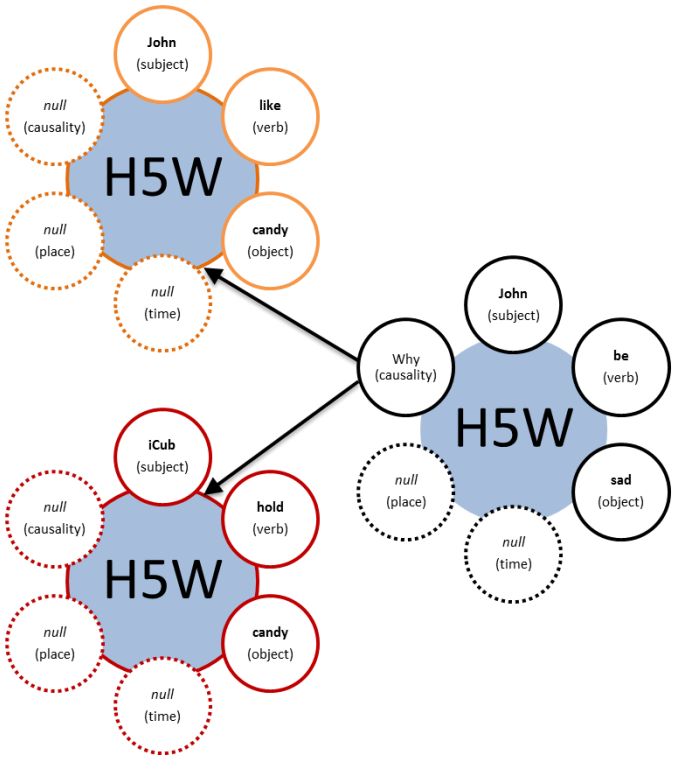

Figure 4. Graphical expression of the causality among relations. The "why" problem is answered by having relations refering to other relations. In this case the example presented refers to the knowledge represented in Figure 3: John is sad because he likes candy and the holder of the candy is the iCub.

3.2.2. Object

As human beings and visually dominated animals, we tend to split the world into independent elements based on their fundamental physical properties such as their unity through motion, their relative non deformability and the behavior they have when put in contact among each other [43]. Our first specification of entities are those which hold a physical reality, which are units of matter positioned and oriented in the physical world. Those are objects, their parts cannot move independently; they can be experienced through the haptic sense and perceived as a whole. While the

abstract "Entity type" is defined only by a name, the "Object type" holds vectors defining a position, an orientation and some dimensions (*i.e.*, bounding box). Moreover, at the sensor level, an object is also by its sensory description (e.g., visual template, point cloud, 3D model, *etc.*), this link allows us to solve the anchoring problem by grounding directly the sensory-specific percept with the symbolic representation. As the same object can have very different descriptions for different sensors, it is more ecologic to have this description defined by the sensor (or its driver) itself and to increment it with every new sensor.

3.2.3. Manipulable Objects

Modeling software frequently describes the world in terms of named and positioned boxes or meshes. In the case of an embodied agent, whose goal it is to interact with the world, it is useful to distinguish among those objects the ones that are static and the ones that could be interacted with and displaced. In particular, some objects can be manipulated by the agent and therefore can be augmented with properties that define how to handle them. This type of information, named affordance [44,45], is computed using visual and/or tactile information during an exploratory process and thereafter are used in manipulation tasks. It can refer to various aspect of the manipulation process including the type of object (e.g., configuration of the hand and the contact points) and how the object can change the kinematic chain, for example to be used as a tool. Formalizing and extracting this information from the sensory stream in robots is a research topic on its own [46–48], but it is clear that it is a crucial property that some objects should be embedded with. In our implementation, we adopted the simple yet generic approach of a unique identifier, assigned and used by the motor control subsystems of the architecture, to define the affordances.

3.2.4. Agent

Some objects demonstrate self-motion compared to the passivity of most of the world, we attribute them with agency [49] and eventually as goal motivated entities [50]. Agents are the active elements in a scene, they are the subjects which act on the other entities and their actions are motivated not only by external causes but also by an invisible internal state which give them some goal. The way humans represent others is deeply influenced by the way they represent themselves, leading to the attribution of human traits to animals, physical or virtual artifacts [51]. Despite abilities of perspective taking, we tend to reason that the dynamics of others mental states are similar to ours, and that environmental stimuli would affect them the same way they affect us. The crucial aspect when representing an agent, is to represent everything that is hidden from the sensors: Its internal state, which is the engine driving its behavior. The representation of agent we adopt here is based on the cognitive architecture model we have used, therefore making the representation of self and others similar in many points. While describing the whole architecture is not in the scope of this paper (refer to [21] for an extensive description), we can extract two crucial aspects: (1) the agent's behavior is guided by a hierarchy of drives and emotions; (2) the representation of the world maintained by the agent is the collection of beliefs formalized following the H5W framework. The resulting structure contains all the necessary information to represent the self of the cognitive architecture as well as to represent others.

As stated before, an agent needs an intrinsic motivation to act. Actions can be formed at different levels, such as uncontrollable reactions to a sensory signal (reflexes), adapted to a given situation and internal needs (e.g., foraging) or planned in the long term based on past experience (e.g., storing food for winter). The DAC architecture [14,18], through its different layers, captures all the described levels, especially that which is of interest to us here: The immediate satisfaction of the internal homeostasis. Part of the internal model of the agent is defined by a set of needs. Those are abstract models of the effects of biochemical variables such as leptin, which is a hormone inhibiting the desire to consume food [52], or complex physiological equilibriums aiming at homeostasis such as the temperature regulatory system. Each need represents a variable that the agent aims at maintaining in a given range. When the variable's value becomes too low or too high, corresponding actions are taken in order to bring it back within its ideal regime. This phenomenon is called homeostasis and can be observed at the levels of physiology, behavior or at the interface between the two (*i.e.*, the animal engages a specific behavior that will lead to the regulation of a physiological state). It has been a relatively popular and successful approach in the design of autonomous artificial systems, especially robots [22,39,53] as it produces ethologically realist behaviors. A tightly linked concept is the notion of emotion: Although they are not consensually modeled, it is agreed that humans are subject to internal state which shape and affect globally the way they behave [54–56]. Emotions are reflected through facial expressions [57], which act as cues for others to perceive our internal state. Where the drives are responsible for triggering a specific behavior, the emotions seem to have a shaping role, defining the stance of this behavior. Taken as a whole, the couple of drives and emotions gives a relatively complete structure that acts as a motivational engine, pushing an organism towards actions that should maintain a stable state (e.g., to survive). Those considerations led us to implement the notions of drives and emotions within the "Agent" type as two sets of homeostatic variables.

The second addition to the "Agent" entity models the beliefs about the world held by the agent. Depending on prior knowledge and on the events witnessed, different agents may represent the same reality differently. The ability to represent beliefs of others, to update them based on the situations they witness and to reason about how people should reason based on their own knowledge, is often considered to be a major component of theory of mind (ToM). The ability to represent this discrepancy among one's and others' beliefs has long been adopted as a standard test (Sally-Anne) for testing for ToM, especially in relation with autism [58,59]. However, it is now argued that the ability to represent false beliefs and the ToM are two different things [60]. In an effort to embed robots with a ToM, the Sally-Anne Test has been replicated and successfully passed by several robots or artificial intelligence systems [61,62]. In our implementation, the beliefs employ again the "Relation concept": Each belief is a specific relation and the whole agent's knowledge is simply a list of beliefs. Interestingly, representing the beliefs of another agent is technically the same thing as representing the beliefs of the self; we therefore embed every agent data structure with a distinct ontology, which represent this agent's beliefs, attributed by the observer agent. The case of the "Self" is just the specific "Agent" instance that is containing all other representations. It is the agents list of beliefs that represents its perceived state of the world (SOW), and it may in turn include other agents, which will have their own SOW. To avoid the recursivity issue (*i.e.*, "You believe that I believe that you believe…") the cognitive architecture is responsible for the level of beliefs propagation, although the data structure can theoretically deal with any number of recursions. Consequently this means that whenever an event is

observed, the architecture updates its self-beliefs, but also updates the beliefs of all the agents' able witness the same event.

To summarize, the "Agent" data structure we define is an object (it has spatial location) which holds a set of drives, a set of emotions and a list of beliefs. A specific "root" instance of "Agent" is used to model the "Self" of the cognitive architecture, its drives and emotions, which are used to fuel the behavioral engine, while its list of beliefs is a representation of the perceived state of the world. Others are represented as elements in this list.

3.2.5. Action

The world is defined both in space and time; it is a dynamic collection of elements that are constantly modifying their spatial properties as a result of self-propulsion or of passive physical reaction to another element contact. This dynamic can be segmented as it is creating a transition between two relatively stable states and humans tend to label it, resulting in the concept of action. The theory of action is vast, spread across several fields of study and the definition of an action may vary greatly. The original study of the concept goes back to Aristotle and it refers to the description of the process ultimately leading a biological being to move (for review and broader considerations see [63]). In our framework, we consider a more general definition of action as a label which describes the transition from one state to another, a state referring to a stable dynamic of objects properties. This pragmatic definition is close to the conceptual spaces approach of Gardenfors which considers actions as patterns of forces [64]. Moreover, this definition is very useful in the domain of planning and teleological reasoning as it can be described as a state-action-state triplet. Another important property of action is the compositionality: An action is made of smaller actions, chained together in a continuous manner. Such composite constructs are sometime called plans, recipes or hierarchical actions [65–69], and all are based on the notion that a sequence of symbols can be folded to create another symbol. While some common approaches use a tree-like structure, we opt for a recursive data-structure. Any action embeds a list of sub-actions (which can be empty, in this case the action is called "atomic"); it also defines a list of pre-conditions and post-conditions. Pre-conditions refer to facts about the world state that must be verified before an action is considered executable; post-conditions are changes that will occur in the world state as a result of this action's execution (*i.e.*, a list of relations to be removed/added to the world state). In the composite case, the pre-conditions of an action are the union of the composite action itself and the pre-conditions of the first sub-action. The global post-conditions can be computed by the successive application of the post-conditions of every sub-action and those of the composite action itself. The action data structure implements a recursive unfolding method which linearizes it as a sequence of atomic executable actions; it also provides methods for calculating the global effects (post-conditions) by simulating the evolution of the world state throughout the execution. Within an action definition, the pre and post conditions refer to relations involving abstract arguments which are substituted to the effective ones during the execution (see Figure 5 for an example and [36,70,71] for an extensive description of the implementation).

Action name

- *take-from*

Sub-actions

1. $subject *open* $complement_place
2. $subject *grasp* $complement_object
3. $subject *close* $complement_place

Pre-Conditions

1. $complement_place *is* present
2. $complement_place *contains* $complement_object

Post-Conditions

- $complement_object *is* present

Figure 5. The formal description of the composite action "take from". Italic words refer to other actions, elements starting with a $ symbol means they are abstract arguments that will be substituted by the real ones during the action execution. An action is called through together with a relation that defines those arguments, for example (*iCub, take-from, pen, drawer*).

Atomic actions when executed are interpreted by the cognitive architecture which typically calls the corresponding motor programs, although the representation presented can also be used to model non motor actions. Indeed our representation of action is as generic as possible so that it can have a mapping with the language concept of verb.

Taken together, the classes we define allow representing properties and facts of the physical world by providing a careful specification of properties of objects, agents and actions. Those constraints ensure that a situated agent (*i.e.*, a robot) will be able to take full advantage of the ontology in order to not only reason about a world state, but also to build it from perceptions and to act on it. However, through inheriting an abstract representation composed only of a single word (the "Entity") we keep our model open to more generic ontology modeling and allow it to represent nonphysical concepts (e.g., "fatigue", "blue", *etc.*). Lastly, the "Relation" data-structure has direct mapping to sentences constructed in the English language, which is particularly of use in the context of human-robot interaction, as this is the most common way for humans to exchange knowledge, through spoken communication.

4. The H5W Acquisition and Transfer through Dialog

4.1. Mapping between Natural Language and H5W Semantic

Any knowledge representation is devoid of sense if it cannot be scaled or consulted. The point in the H5W representation is to define and link concepts to enable a machine to interact with its environment.

The hook between the representations and the sensory data presented in earlier work [22,72], allows the agent to make use of a set of heterogeneous sensors to fill in the information regarding the spatial behavior and identity of the various physical elements in the world. However, one may want to teach the system about facts that are beyond the spatial reach of its sensors (e.g., "It is cold outside"); moreover, when it comes to more abstract or qualitative concepts the sensing approach quickly shows limits (e.g., "Freedom is good."). An obvious source for this type of knowledge would be the endless collection of sensors, text corpus and ontologies available online; however our focus is the case of natural interaction with the environment. This environment includes agents who are potential sources of abstract knowledge, as is the case when considering the interaction of a toddler with his caregivers. In this context, the direct mapping of the H5W representation with speech turns the spoken dialog into a channel of choice for expanding or interrogating the knowledge of the artificial agent. Part of the platform we have developed around the H5W framework aimed at studying the social interaction between a humanoid robot (iCub) and a human partner therefore investigating the issue of spoken communication. As a result, we implemented a dialog interface based on the H5W representation and which uses speech recognition and synthesis to offer natural interaction.

Despite much progression in the domain, open speech recognition (*i.e.*, dictation) results are still very variable: The state of the art in this domain is provided by Google and is open for testing [73]. This variability is especially problematic in the case of a robot dealing with symbolic representations and the usual solution in robotic platforms is to use pre-specified grammar, pre-defining the possible commands available to the user. The fixed grammar approach enhances a lot the recognition quality but has two major drawbacks: (1) it needs the user to be aware of the grammar as it does not allow much variations on the user side; (2) it often requires the vocabulary to be defined before all the interactions. The H5W framework helps to overcome those limitations through the tight link it has with language. We developed an open source grammar based engine [74] that allows us to create, run and modify grammars at runtime as well as to keep track of vocabulary categories that can expand dynamically (the interlocutor can add a new word to vocabulary by switching to open recognition or spelling modes). Although this engine is independent from the H5W framework, it fits very well with the current representation as the different "Entity" categories can be mapped to given vocabularies (*i.e.*, agents, objects, verbs, and concepts). In the case of situated interaction, the world state perceived by the robot directly defines the vocabularies by giving priority to the present physical entities (*i.e.*, it is more likely to speak about the direct environment). Regarding the grammar aspect, we considered a relatively simple yet very generic form of dialog which allows an agent either to inform, question or give orders to another agent, about a H5W representation (Figure 6).

By defining relations with some additional specifications, the major components of dialog can be formulated and interpreted by the system. Expressing a full relation represent an affirmation, therefore increasing the knowledge of the communication recipient. A partial relation with a special symbol "?" in place of one or more of the roles represent a question in which the filled roles represent some constraints to be verified (e.g., "(*iCub, like, ?)*" means "What does iCub like?"). The role being questioned directly defines the interrogative word that will start the sentence formulation (*i.e.*, Who = subject, What = object, When = complement of time, Where = complement of place, How = verb, Why = causal relations), therefore setting a specific path in the grammar. Finally an order is an "invalid" Relation that has no subject, in this condition the corresponding sentence is in the

imperative tense, therefore prompting the execution of the rest of the relation, with the interlocutor as the missing subject. The grammars used to transform a sentence into a H5W relation are relatively simple and available online as part of the WYSIWYD project [75]. The language system can be used bidirectionally: Speech recognition can be used as a way for the user to modify or query the robots knowledge, as well as to command it; and speech synthesis used by the robot to express in a spoken manner, all the concepts and relations that it is manipulating. While the H5W representation has been extensively used in the robot control domain for representing perceptual concepts, the direct mapping to natural language has never been tested in a systematic manner. In the following section, we assess the performance of the H5W representation as a way to support information exchange through natural language interaction.

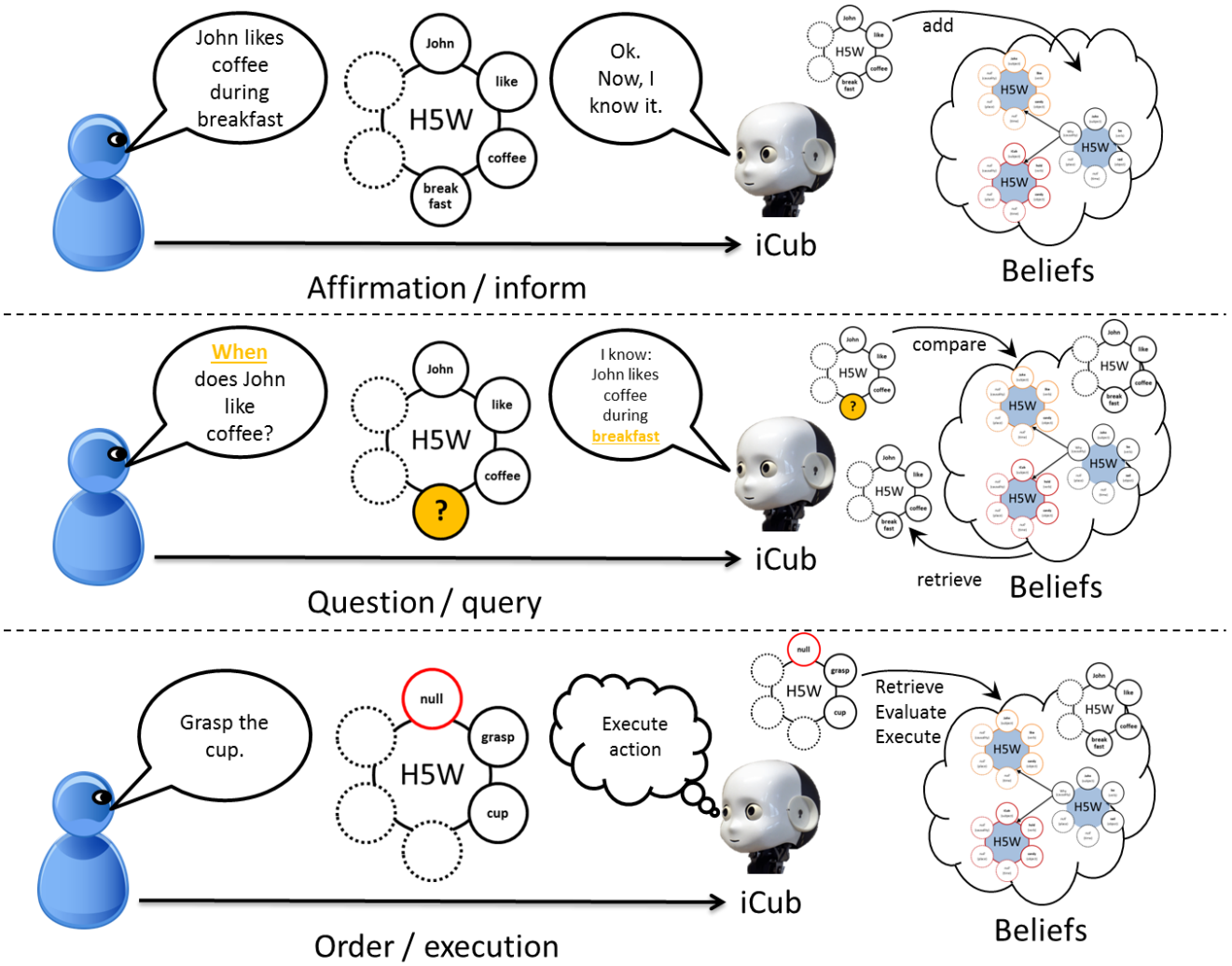

Figure 6. The dialog system. Through simple grammar, affirmative, interrogative and imperative sentences can be mapped to a semantic representation, using the H5W framework. The agent can either build this representation from speech recognition or use an existing representation and generate a sentence description.

4.2. Benchmarking H5W for Information Exchange

The H5W framework offers a direct mapping between natural language and semantic content. In order to evaluate the quality and applicability of this link in the context of a dialog between a human and a machine, we designed a simple experiment about information exchange. The scenario is inspired by the popular board game "Cluedo", in which characters are interacting in a house to solve a murder. The original game goal consists of uncovering the crime by finding the right combination of a suspect, weapon and place, through the exchange of information between the players. In our version, we have two players and we extend the representation to include an action and a time. The elements composing the situation are picked up from 6 characters, 5 actions, 6 objects, 9 places and 4 time periods, therefore allowing for 6480 combinations. Each player possesses a collection of 5 facts that are complete (e.g., *"Miss Peacock hid the knife in the living room during the morning."*) as well as 5 other facts, which have some missing parts (e.g., *"Colonel Mustard took _?_ in the kitchen during _?_"*), taken together the two players share the same collection of 10 complete facts. In turns, each player asks a question to the other player in order to fill his missing knowledge. The goal is to fill all the missing gaps in the statements and the game ends when both players have completely filled in the missing parts.

The setup consists of an audio headset (Microsoft LifeChat LX3000) connected to a computer with the screen turned off. One player is the subject wearing the headset and the second player is the computer. The only modality available for interaction is the audio communication, so the user can only send information to the system through the microphone and receive feedback through the headset. The subject is untrained and instructed to speak naturally to the system. Each subject performed the game twice, with different statements for the first and second game, as we wanted to assess the adaptability of the human user to the system. The subjects were also asked to fill in a questionnaire regarding demographic information, their familiarity with synthetic speech engine and their opinion about the interaction. The measures of familiarity and assessment of the interaction were self-reported on 5 point Likert scales.

The sample consisted of 10 subjects (age: M = 24.1, SD = 5.8) from various nationalities and having various native languages, all of them could speak and understand professional English, albeit with strong accents. None of the subjects reported global expertise with synthetic speech systems (e.g., those used in GPS systems, the personal assistants like "Siri" or "Cortana"), with a mean reported dialog frequency of 1.8 (1 being "never", 5 being "everyday"). We evaluated each subject in terms of the time taken to fill the knowledge gaps in the game and the number of errors of the human and system. We counted as an error any completed fact that differed from the original fact, known by the other player. We report (Figure 7) an outstanding performance (>90% of correct answers) when the system is answering a question from the user, and an above average performance (>63%) when the user is answering the questions of the system. The difference between the two conditions is explained by two factors. First, when the system answers a question (e.g., *"Where did Miss Scarlet hide the knife?"*) it formulates the answer using a complete sentence (e.g., *"Miss Scarlet did hide the knife in the Kitchen."*). This allowed the user to spot any false recognition, while still using the statement of the system to fill his partial facts. Interestingly, some users did also adopt this way of answering after

several interactions with the system, while the system was perfectly able (and less error prone) to catch direct statements such as *"In the kitchen"*.

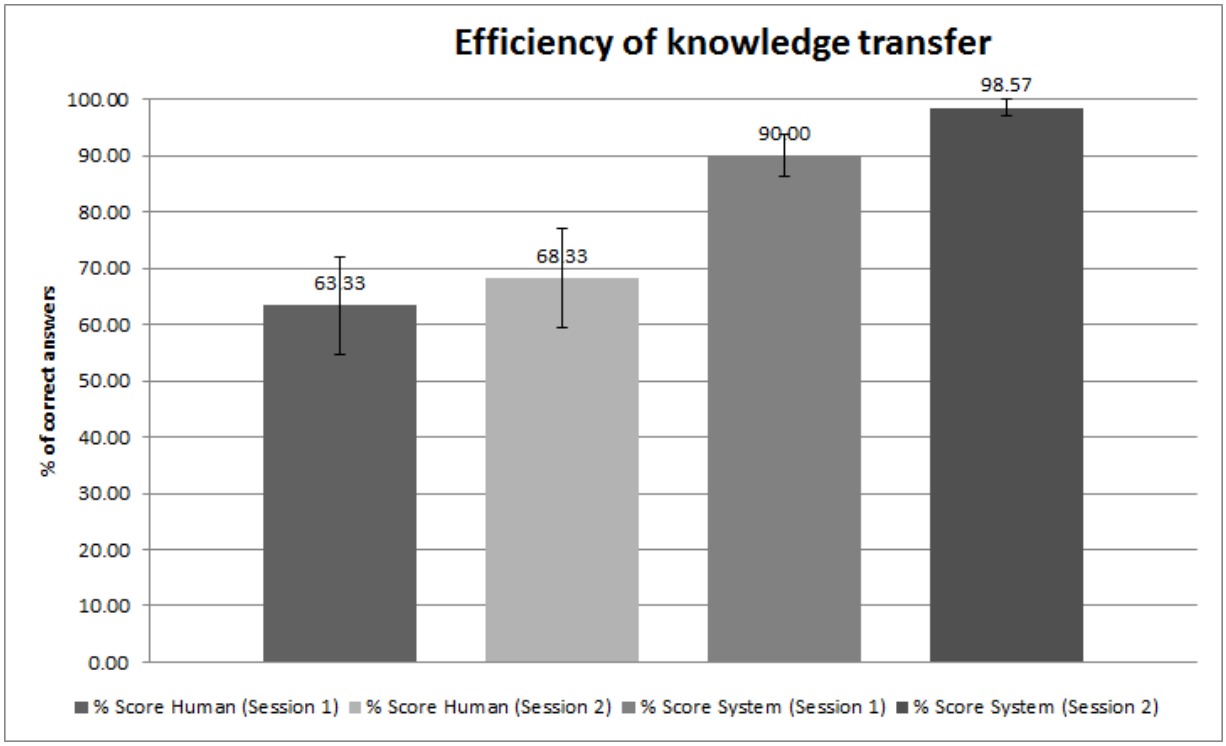

Figure 7. Efficiency of the knowledge transfer, as percentage of correct answers retrieved from the partner. "Score Human" refers to the answers retrieved by the system asking the subject a question. "Score system" refers to the answers retrieved by the human, asking the system a question.

The second reason that explains the difference between the two conditions (user's question vs system's question) is related to the synchronicity of the speaker and the recognition system. When the subjects asked a question, they generally had a relatively large mute time during which they were looking at their form and formulating the question silently, therefore allowing the recognition system to purge any previous noise and get steady for recognition. In the other condition (human answering questions from the system), the subjects were more prompt to answer, sometimes even while the system was still asking the question. This problem also produced a more chaotic interaction, as the system could not react quickly enough, often due to false recognition of some noise, therefore saying things like "*Sorry I did not understand*" before the user spoke. On the other hand, the system could also have discarded the users answer as prior noise and therefore would wait for an answer, making the subject believe that the system was not responsive. The issue of synchronization of speakers and more generally of turn taking is a standard problem in the domain of human-robot interaction and one solution is to use non-verbal clues to enforce the turn taking behavior [76]. Finally, we can also point to some weak correlation between the global level of stress reported by the subjects (1 = no stress, 5 = stress) and their scores on instructing the system ($r = 0.68$). The score obtained while listening to the system is not correlated with

the stress level (r = 0.4). Overall the subjects did not report being stressed neither while listening to the system (M = 2.4, STD = 1.26) nor while talking to the system (M = 2.8, STD = 1.55).

Furthermore, we asked the subjects to rate several qualitative aspects of the interaction. Although this data is self-reported, the subjective perception of interactive artificial systems is a central point to assess before any deployment as a product. Our questionnaire pinpointed several important aspects of the spoken interaction. Overall the results are in favor of the system (Figure 8): The subjects felt that the system was listening to them, that it was using the information they provided and that it was answering adequately. Ultimately they reported an exchange of information with the system and more importantly they found that the system was easy to use. Together, those results allow use to claim the validity of the H5W framework as a way to represent, query and transfer semantic content through natural language.

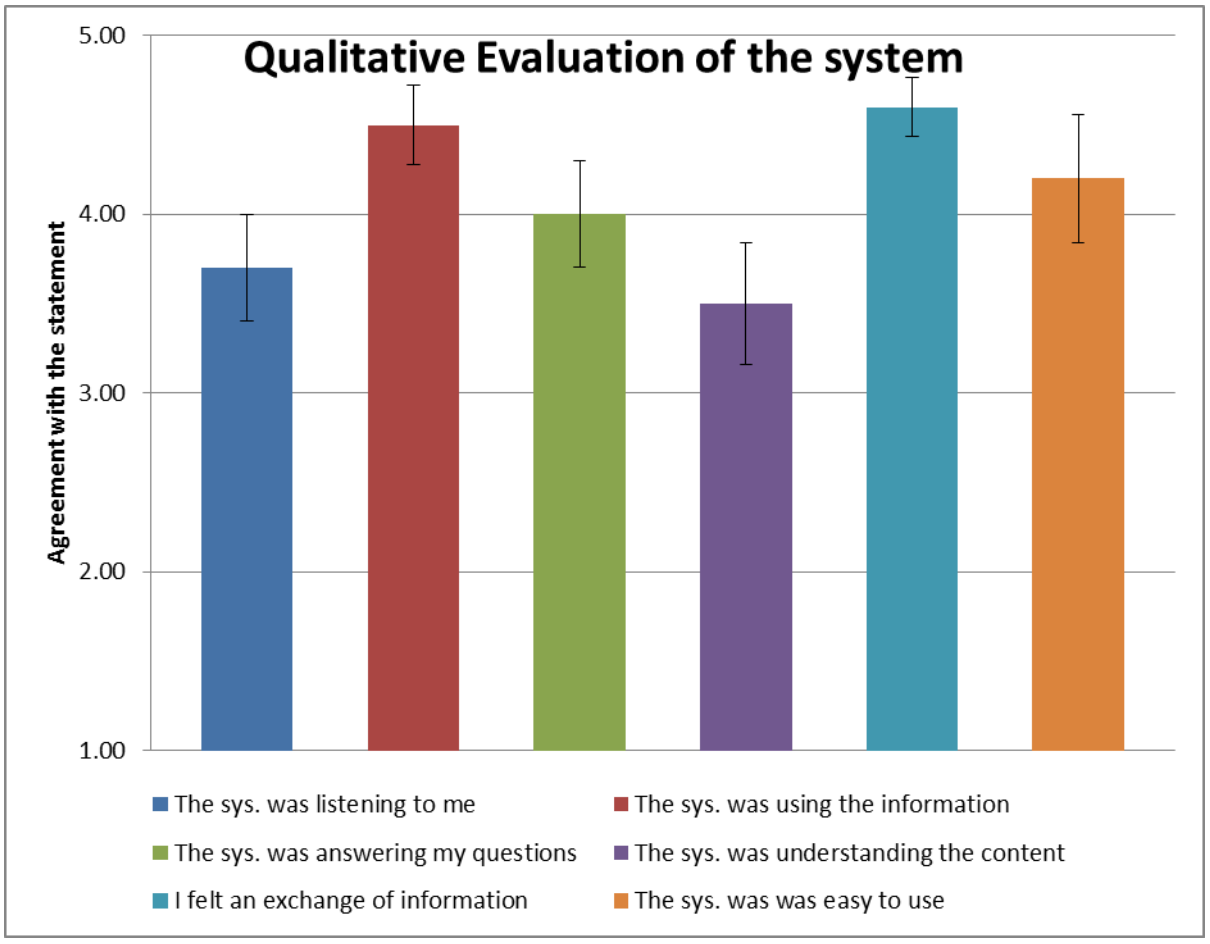

Figure 8. Self-reported appreciation of the system by the subjects using a Likert scale. A score of 5 means the subject agreed with the statement, a 1 means he disagreed. This data confirms that the successful exchange of information between the user and the artificial system was also perceived as a successful spoken interaction by the user. Please refer to the supplementary material for the exact formulation of the questionnaire.

5. Discussion

Since its implementation as part of the DAC architecture, the H5W library has been intensively used as a tool to support other studies. In particular, it is the core implementation of multiple HRI scenarios such as the exploration of the effects of allostatic control on the human perception of the robot [22,23], or several studies about the effect of other's representation in terms of empathy and ToM [21]. In developmental robotics, it has also been used as a way to maintain the working memory of the robot and served as a base for the development of long term autobiographical memory [77–79]. It has also supported perceptual representations in the context of language acquisition [80–82] and in this work we have strengthened the validity of the H5W framework in the context of natural language by assessing its ability to represent, query and share semantic information in a purely language based scenario. Through the definition of a few concepts (*i.e.*, agents, objects, actions) which are the main elements of any interaction scenario, we manage to cover numerous situations and do not limit our knowledge representation to the scope of a single study. In addition, we kept the possibility of a complete abstract representation of entities which, together with an extended relational structure, allows compatibility with more traditional ontological models. An important contribution is the careful attention we have placed have taken in defining the problem from a software engineering perspective and providing a concrete usable implementation of the notions defined. With the current report, we have intended to give a comprehensive insight into these definitions, which are at the core of the architecture and are used for other studies. We used concepts coming from ontology and software engineering domains in order to create a representation that is simple, yet generic and powerful enough to allow psychologists to implement genuine scenarios or roboticists to scale up existing cognitive architectures. By correctly conceptualizing the problems that are facing situated agents and turning this conceptualization into a robust software implementation, we have built the foundations for a long lasting representation of knowledge.

Acknowledgments

This work is supported by the EU FP7 project WYSIWYD (FP7-ICT-612139) and by A*STAR JCO grant #1335H00098. We would like to thank Luke Boorman for his contribution to the final version of this manuscript.

Author Contributions

Stephane Lallee is the main software developer of the H5W framework and author of this paper. He also implemented numerous modules composing the WR-DAC architecture used in the WYSIWYD project. Paul Verschure is the original author of the DAC theory and the H5W problem.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Atzori, L.; Iera, A.; Morabito, G. The internet of things: A survey. *Comput. Netw.* **2010**, *54*, 2787–2805.
2. Sundmaeker, H.; Guillemin, P.; Friess, P.; Woelfflé, S. *Vision and Challenges for Realising the Internet of Things*; *Publications Office of the European Union: Luxembourg, Luxembourg,* 2010.
3. Ashton, K. That 'Internet Of Things' thing. *RFID J.* **2009**, *22*, 97–114.
4. McAfee, A.; Brynjolfsson, E. Big data: The management revolution. *Harv. Bus. Rev.* **2012**, *90*, 60–68.
5. Manyika, J.; Chui, M.; Brown, B.; Bughin, J. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*; McKinsey Global Institute: Summit, NJ, USA, 2011.
6. Lohr, S. The age of big data. Available online: http://wolfweb.unr.edu/homepage/ania/NYTFeb12.pdf (accessed on 18 February 1015).
7. Bengio, Y. *Learning Deep Architectures for AI*; Now Publishers Inc.: Norwell, MA, USA, 2009.
8. Krizhevsky, A.; Sutskever, I.; Hinton, G. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* 25; The MIT Press: Cambridge, MA, USA, 2012; pp. 1097–1105.
9. Hinton, G.; Deng, L.; Yu, D. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.* **2012**, *29*, 82–97.
10. Collobert, R.; Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008; pp. 160–167.
11. Coradeschi, S.; Saffiotti, A. An introduction to the anchoring problem. *Rob. Auton. Syst.* **2003**, 43, 85–96.
12. Coradeschi, S.; Saffiotti, A. Anchoring symbols to sensor data: Preliminary report. In *AAAI/IAAI*; Association for the Advancement of Artificial Intelligence: Palo Alto, CA, USA, 2000.
13. Harnad, S. The symbol grounding problem. *Phys. D* **1990**, *42*, 335–346.
14. Verschure, P. Distributed adaptive control: A theory of the mind, brain, body nexus. *Biol. Inspired Cogn. Archit.* **2012**, *1*, 55–72.
15. Beeson, P.; Kortenkamp, D.; Bonasso, R.P.; Persson, A.; Loutfi, A.; Bona, J.P. An ontology-based symbol grounding system for human-robot interaction. In Proceedings of the 2014 AAAI Fall Symposium Series, Arlington, MA, USA, 13–15 November 2014.
16. Prescott, T.J. Lepora, N.F.; Verschure, P.F.M.J. A future of living machines?: International trends and prospects in biomimetic and biohybrid systems. In Proceedings of the SPIE 9055, Bioinspiration, Biometrics and Bioreplication, San Diego, CA, USA, 9–12 March 2014.
17. Verschure, P. Formal minds and biological brains II: From the mirage of intelligence to a science and engineering of consciousness. *IEEE Intell. Syst. Trends Controv.* **2013**, 7–10.
18. Pfeifer, R.; Verschure, P. Distributed adaptive control: A paradigm for designing autonomous agents. In *Toward a Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life*; MIT Press: Cambridge, MA, USA, 1992; pp. 21–30.
19. Verschure, P.; Kröse, B.; Pfeifer, R. Distributed adaptive control: The self-organization of structured behavior. *Rob. Auton. Syst.* **1992**, *9*, 181–196.

20. Verschure, P. Connectionist explanation: Taking positions in the Mind-Brain dilemma. In *Neural Networks and a New Artificial Intelligence*; International Thomson Computer Press: Boston, MA, USA, 1997.
21. Lallée, S.; Vouloutsi, V.; Blancas, M.; Grechuta, K.; Puigbo, J.; Sarda, M.; Verschure, P.F.M. Towards the synthetic self: Making others perceive me as an other. *Paladyn J. Behav. Robot.* 2015, submit.
22. Vouloutsi, V.; Lallée, S.; Verschure, P. Modulating behaviors using allostatic control. In *Biomimetic and Biohybrid Systems*; Springer: Berlin, Germany, 2013, pp. 287–298.
23. Vouloutsi, V.; Grechuta, K.; Lallée, S.; Verschure, P. The Influence of behavioral complexity on robot perception. In *Biomimetic and Biohybrid Systems*; Springer: Berlin, Germany, 2014, pp. 332–343.
24. Verschure, P.F.; Pennartz, C.M.; Pezzulo, G. The why, what, where, when and how of goal-directed choice: neuronal and computational principles. *Philos. Trans. R. Soc. London B Biol. Sci*. **2014**, *369*, 20130483.
25. Antoniou, G.; Harmelen, F. Web ontology language: Owl. In *Handbook on Ontologies*; Springer: Berlin, Germany, 2009.
26. Avraham, S.; Tung, C.; Ilic, K. The Plant Ontology Database: A community resource for plant structure and developmental stages controlled vocabulary and annotations. *Nucleic Acids Res.* **2008**, *36*, D449–D454.
27. Consortium, G.O. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **2004**, *32*, D258–D261.
28. Soldatova, L.; Clare, A.; Sparkes, A.; King, R. An ontology for a Robot Scientist. *Bioinformatics* **2006**, *22*, e464–e471.
29. Tenorth, M.; Clifford Perzylo, A.; Lafrenz, R.; Beetz, M.; Perzylo, A. The RoboEarth language: Representing and exchanging knowledge about actions, objects, and environments. In Proceedings of the Twenty-Third international joint conference on Artificial Intelligence, Beijing, China, 3–9 August 2013; pp. 1284–1289.
30. Lemaignan, S.; Ros, R. ORO, a knowledge management platform for cognitive architectures in robotics. In Proceedings of the 2010 IEEE/RSJ International Conference on, Intelligent Robots and Systems (IROS), Taipei, Taiwan, 18–22 October 2010; pp. 3548–3553.
31. Tenorth, M.; Beetz, M. KnowRob: A knowledge processing infrastructure for cognition-enabled robots. *Int. J. Robot. Res.* **2013**, *32*, 566–590.
32. Ros, R.; Lemaignan, S.; Sisbot, E. Which one? Grounding the referent based on efficient human-robot interaction. In Proceedings of the 2010 IEEE RO-MAN, Viareggio, Italy, 13–15 September 2010; pp. 570–575.
33. Zweigle, O.; Andrea, R.; Häussermann, K. RoboEarth—Connecting robots worldwide. In Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, Seoul, Korea, 24–26 November 2009; pp. 184–191.
34. Chella, A.; Kurup, U.; Laird, J.; Trafton, G.; Vinokurov, J.; Chandrasekaran, B. The challenge of robotics for cognitive architectures. In Proceedings of the International Conference on Cognitive Modeling, Ottawa, ON, Canada, 11–14 July 2013; pp. 287–290.
35. Searle, J. Minds, brains, and programs. *Behav. Brain Sci.* **1980**, *3*, 417–424.

36. Lallee, S.; Madden, C.; Hoen, M.; Dominey, P.F. Linking language with embodied and teleological representations of action for humanoid cognition. *Front. Neurorobot.* **2010**, *4*, 12.
37. Oudeyer, P. Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evol. Comput.* **2007**, *11*, 265–286.
38. Maslow, A. A theory of human motivation. *Psychol. Rev.* **1943**, *50*, 370–396.
39. Breazeal, C. A Motivational system for regulating human-robot interaction. In *AAAI/IAAI*; Association for the Advancement of Artificial Intelligence: Palo Alto, CA, USA, 1998.
40. Hawes, N. A survey of motivation frameworks for intelligent systems. *Artif. Intell.* **2011**, *175*, 1020–1036.
41. Verschure, P. F. M. J. Formal Minds and Biological Brains II: From the Mirage of Intelligence to a Science and Engineering of Consciousness. *IEEE Intell. Syst. Trends Controv.* **2013**, *28*, 7–10.
42. White, D.R.D.; Reitz, K.P.K. Graph and semigroup homomorphisms on networks of relations. *Soc. Netw.* **1983**, *5*, 193–234.
43. Spelke, E. Principles of object perception. *Cogn. Sci.* **1990**, *14*, 29–56.
44. Gibson, J. *The Theory of Affordances*; Lawrence Erlbaum Assoociates, Inc.: Hilldale, NJ, USA, 1977.
45. Lockman, J.; McHale, J. Object manipulation in infancy. In *Action in Social Context; S*pringer Publishing: New York, NY, USA, 1989; pp. 129–167.
46. Tikhanoff, V.; Pattacini, U.; Natale, L.; Metta, G. Exploring affordances and tool use on the iCub. In IEEE-RAS International Conference on Humanoid Robots (Humanoids), Atlanta, GA, USA, 15–17 October 2013.
47. Moldovan, B.; Moreno, P. Learning relational affordance models for robots in multi-object manipulation tasks. In Proceedings of the 2012 IEEE International Conference on Robotics and Automation (ICRA), Saint Paul, MN, USA, 14–18 May 2012; pp. 4373–4378.
48. Moldovan, B.; van Otterlo, M. Statistical relational learning of object affordances for robotic manipulation. In Proceedings of the International Conference on Inductive Logic Programming, Dubrovnik, Croatia, 17–19 September 2012.
49. Leslie, A. A theory of agency. In *Causal cognition: A multidisciplinary debate*; Sperber, D.; Premack, D.; Premack, A. J., Eds.; Oxford University Press: New York, NY, 1995; pp. 121–141.
50. Csibra, G.; Gergely, G.; Bíró, S.; Koós, O.; Brockbank, M. Goal attribution without agency cues: The perception of 'pure reason' in infancy. *Cognition* **1999**, *72*, 237–267.
51. Reeves, B.; Nass, C. *How People Treat Computers, Television, and New Media Like Real People and Places*; Cambridge University Press: Cambridge, UK, 1996.
52. Brennan, A.; Mantzoros, C. Drug insight: The role of leptin in human physiology and pathophysiology—Emerging clinical applications. *Nat. Clin. Pract. Endocrinol. Metab.* **2006**, *2*, 318–327.
53. Sanchez-Fibla, M. Allostatic control for robot behavior regulation: A comparative rodent-robot study. *Adv. Complex Syst.* **2010**, *13*, 377–403.
54. Frijda, N. *The Emotions*; Cambridge University Press: Cambridge, UK, 1987.
55. Fellous, J.; LeDoux, J.; Arbib, M. Toward basic principles for emotional processing: What the fearful brain tells the robot. In *Who Needs Emotion Brain Meets Robot*; Oxford University Press: Oxford, UK, 2005; pp. 245–270.

56. Adolphs, R.; Tranel, D.; Damasio, A.R. Dissociable neural systems for recognizing emotions. *Brain Cogn.* **2003**, *52*, 61–69.
57. Ekman, P. An argument for basic emotions. *Cogn. Emot.*, **1992**, *6*, 169–200.
58. Wimmer, H.; Perner, J. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* **1983**, *13*, 103–128.
59. Baron-Cohen, S.; Leslie, A.; Frith, U. Does the autistic child have a 'theory of mind'? *Cognition* **1985**, *21*, 37–46.
60. Bloom, P.; German, T.P. Two reasons to abandon the false belief task as a test of theory of mind. *Cognition* **2000**, *77*, 25–31.
61. Milliez, G.; Warnier, M. A framework for endowing an interactive robot with reasoning capabilities about perspective-taking and belief management. In Proceedings of the 2014 RO-MAN: The 23rd IEEE International Symposium on Robot and Human Interactive Communication, Edinburgh, Scotland, 25–29 August 2014.
62. Sindlar, M.; Dastani, M.; Meyer, J. BDI-based development of virtual characters with a theory of mind. *Intell. Virtual Agents* **2009**, *5773*, 34–41.
63. Hyman, J.; Steward, H. *Agency and Action*; Cambridge University Press: Cambridge, UK, 2004.
64. Gardenfors, P.; Warglien, M. Using conceptual spaces to model actions and events. *J. Semant.* **2012**, *29*, 487–519.
65. Corkill, D. Hierarchical planning in a distributed environment. *IJCAI* **1979**, *79*, 168–175.
66. Whiten, A.; Flynn, E.; Brown, K.; Lee, T. Imitation of hierarchical action structure by young children. *Dev. Sci.* **2006**, *9*, 574–582.
67. McDermott, D.; Ghallab, M.; Howe, A.; Knoblock, C. *PDDL—The Planning Domain Definition Language*; Technical Report CVC TR98-003/DCS TR-1165; Yale Center for Computational Vision and Control: New Haven, CT, USA, 1998.
68. Alili, S.; Warnier, M.; Ali, M.; Alami, R. Planning and plan-execution for human-robot cooperative task achievement. In Proceedings of the 19th International Conference on Automated Planning and Scheduling, Thessaloniki, Greece, 19–23 September 2009.
69. Lenz, A.; Lallee, S.; Skachek, S.; Pipe, A.G.; Melhuish, C.; Dominey, P.F. When shared plans go wrong: From atomic- to composite actions and back. In Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vilamoura, Portugal, 7–12 October 2012; pp. 4321–4326.
70. Lallée, S.; Lemaignan, S. Towards a platform-independent cooperative human-robot interaction system: I. perception. In Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Taipei, Taiwan, 18–22 October 2010; pp. 4444–4451.
71. Lallée, S.; Pattacini, U.; Lallee, S.; Boucher, J.D.; Lemaignan, S.; Lenz, A.; Melhuish, C.; Natale, L.; Skachek, S.; Hamann, K.; *et al.* Towards a platform-independent cooperative human-robot interaction system: Ii. perception, execution and imitation of goal directed actions. In Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), San Francisco, CA, USA, 25–30 September 2011, pp. 2895–2902.
72. Lallée, S.; Wierenga, S.; Pattacini, U.; Verschure, P. EFAA—A companion emerges from integrating a layered cognitive architecture. In Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction, Bielefeld, Germany, 3–6 March 2014; p. 2008.

73. Google Speech Recognition Engine. Available online: http://www.google.com/intl/fr/chrome/demos/speech.html (accessed on 18 February 2015).
74. WYSIWYD Speech Recognizer. Available online: https://github.com/robotology/speech (accessed on 18 February 2015).
75. What You Say Is What You Did (WYSIWYD) Project. Available online: http://wysiwyd.upf.edu/ (accessed on 18 February 2015).
76. Lallée, S.; Hamann, K.; Steinwender, J.; Warneken, F.; Martienz, U.; Barron-Gonzales, H.; Pattacini, U.; Gori, I.; Petit, M.; Metta, G.; *et al.* Cooperative human robot interaction systems: IV. Communication of shared plans with Naïve humans using gaze and speech. In Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Tokyo, Japan, 3–7 November 2013.
77. Pointeau, G.; Petit, M.; Dominey, P. Successive developmental levels of autobiographical memory for learning through social interaction. *IEEE Trans. Auton. Ment. Dev.* **2014**, *6*, 200–212.
78. Pointeau, G.; Petit, M.; Dominey, P. Embodied simulation based on autobiographical memory. *Biomim. Biohybrid Syst.* **2013**, *8064*, 240–250.
79. Pointeau, G.; Petit, M.; Dominey, P.F. Robot learning rules of games by extraction of intrinsic properties. In Proceedings of the Sixth International Conference on Advances in Computer-Human Interactions, Nice, France, 24 February–1 March 2013; pp. 109–116.
80. Pointeau, G.; Petit, M.; Gibert, G.; Dominey, P. Emergence of the use of pronouns and names in triadic human-robot spoken interaction. In Proceedings of the 2014 Joint IEEE International Conferences on Development and Learning and Epigenetic Robotics (ICDL-Epirob), Genoa, Italy, 13–16 October 2014.
81. Petit, M.; Lallée, S.; Boucher, J.; Pointeau, G.; Cheminade, P.; Ognibene, D.; Chinellato, E.; Pattacini, U.; Gori, I.; Martinez-hernandez, U.; *et al.* The coordinating role of language in real-time multi-modal learning of cooperative tasks. *Trans. Auton. Ment. Dev.* **2013**, *5*, 3–17.
82. Hinaut, X.; Petit, M.; Pointeau, G.; Dominey, P.F. Exploring the acquisition and production of grammatical constructions through human-robot interaction with echo state networks. *Front. Neurorobot.* **2014**, *8*, doi: 10.3389/fnbot.2014.00016.