— DRAFT —
Grounding Language in the World: Signs, Schemas, and Meaning

Deb Roy

Cognitive Machines Group
The Media Laboratory
Massachusetts Institute of Technology

Abstract

A theoretical framework for grounding language is introduced that provides a computational path from sensing and motor action to words and speech acts. The approach combines concepts from semiotics and schema theory to develop a holistic approach to linguistic meaning. Schemas serve as structured beliefs that are grounded in an agent's physical environment through a causal-predictive cycle. Speech acts are interpreted in terms of grounded schemas. The theory reflects lessons learned from implementations of several language processing robots and provides a framework for the practical design of situated, physically-embedded natural language processing systems.

Key words: grounding, representation, language, situated, embodied, functional, semiotic, schemas, predictive, causal, meaning

1 Language and Meaning

The relationship between words and the physical world, and consequently our ability to use words to refer to entities in the world, provides the foundations for linguistic communication. Current approaches to the design of language processing systems are missing this critical connection, which is achieved through a process I refer to as *grounding* – a term I will define in detail. A survey of contemporary textbooks on natural language processing (e.g., [27]) reveals a rich diversity of data structures and algorithms concerned solely with manipulation of human-interpretable symbols (in either text or acoustic form) without any serious effort to connect semantic representations to the physical world.

Is this a problem we should really care about? Web search engines and word processors seem to work perfectly fine – why worry about distant connections between language and the physical world? To see why, consider the problem of building natural language processing systems which can in principled ways interpret the speaker's meaning in the following everyday scenarios:

An elderly woman asks her aide, "Please push that chair over to me".

A man says to his waiter, "This coffee is cold!".

A child asks his father, "What is that place we visited yesterday?".

A cell phone call from a car is intercepted, "We'll be there in 10 minutes".

How might we build a robot that can respond appropriately in place of the aide or waiter? The words make sense only when related to the particulars of the physical situation. How might a web search engine be designed to handle the child's query? How should we design a speech understanding system that can make sense of the cell phone message in relation to knowledge about space and time? These are of course not questions that are typically considered part of natural language processing, but these *are* basic questions that every human language user handles with deceiving ease. The words in each of these examples refer to the physical world in very direct ways. The listener cannot do the right thing unless he / she (it?) knows something about the physical situation to which the words refer, and can assess the speaker's reasons for choosing the words as they have.

In recent years, several strands of work have emerged that begin to address the problem of connecting language to the world [58, 18, 59, 39, 4, 48, 12, 61, 23, 9]. Our own efforts have led to several implemented conversational robots and other situated language systems [50, 51, 52, 20, 53, 55]. For example, one of the robots [55] is able to translate spoken language into object manipulation actions guided by perception.

Motivated by our experiences of implementing these systems, I present a theoretical framework for language grounding that provides a computational path from embodied, situated, sensorimotor primitives to words and speech acts – from sensing and acting to symbols. Building upon a rich body of schema theory [29, 44, 2, 38, 40, 57, 16] and semiotics [43, 41, 17, 37], I present an original holistic framework for representing the meaning of words and speech acts that can be used to guide the construction of a broad range of grounded language systems.

A gist of the framework is as follows. Agents are able to translate between speech acts, perceptual acts, and motor acts. An agent that sees a fly or hears the descriptive speech act, "There is a fly here" is able to translate either observation into a common representational form. Upon hearing the directive speech act, "Swat that fly!", an agent forms a mental representation that

guides its sensorimotor planning mechanisms towards the intended goal. Signs are physical patterns in the world which can be interpreted by agents to stand for entities (objects, properties, relations, actions, situations, and, in the case of speech acts, goals). Speech acts, constructed from lexical units, are one class of signs that can be observed by agents. Sensor-grounded perception leads to two more classes of signs which indicate, roughly, the "what" and "where" information about some entity. To interpret signs, agents activate structured networks of beliefs [1] called schemas. Schemas are made of continuous and discrete elements that are linked through six types of projections. Two of these projection types, sensor and action projections, provide links between an agent's internal representations and the external environment. These links are shaped by the specific physical embodiment of the agent. The four remaining projection types are used for internal processes of attention, categorization, inference, and prediction.

Taxonomic distinctions made in the theory are motivated by recurring distinctions that have appeared in our implementations – distinctions which in turn were driven by practical engineering concerns. Although the theory is incomplete and evolving, I believe it will be of value to those interested in designing physically embedded natural language processing systems. The theory may also be of value from a cognitive modeling perspective although this is not the focus of this paper.

Connecting language to the world is of both theoretical and practical interest. In simple terms, the world is full of language about concrete stuff that machines cannot make sense of because they have no way to jointly represent words and stuff. We talk about places we are trying to find, about the action and characters of video games, about the weather, about the clothes we plan to buy and about the music we like. How can we build machines that can converse about such everyday matters? From a theoretical perspective, I believe that language rests upon deep non-linguistic roots. Any attempt to represent natural language semantics without proper consideration of these roots is fundamentally limited.

[1] Although this paper deals with topics generally referred to as knowledge representation in AI, my focus will be on beliefs. From an *agent's point of view*, all that there are are beliefs about the world marked with degrees of certainty. Admittedly, as a designer of robots and other agents, I share the intuition that a robot's belief x is *true* (and thus may be called knowledge) just in the cases for which the corresponding situation is the case – a correspondence that I as the designer can verify (Bickhard calls this "designer semantics" [7]). However, I prefer to develop an approach that avoids reliance on notions of correspondence, since the autonomous agents we design do not have the option of "stepping out of their skins" to verify correspondences.

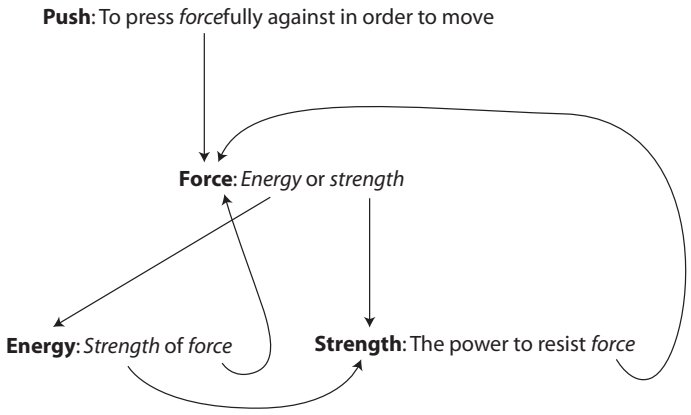

Fig. 1. A network of definitions extracted from Webster's Dictionary containing circular definitions. To make use of such symbolic networks, non-linguistic knowledge is essential to ground basic terms of linguistic definitions (symbolic descriptions).

Inherent to current natural language processing (NLP) systems is the practice of constructing representations of meaning that bottom out in symbolic descriptions of the world as conceived by human designers. As a result, computers are trapped in sensory deprivation tanks, cut off from direct contact with the physical world. Semantic networks, meaning postulates, and various representations encoded in first order predicate calculus all take objects and relations as representational primitives that are assigned symbolic names. Without additional means to unpack the meaning of symbols, the machine is caught in circular chains of dictionary-like definitions such shown in Figure 1 (Harnad [22] has raised this objection previously in the context of motivating symbol grounding). Efforts to encode knowledge using symbolic forms which resemble natural language and that can be written down by human "knowledge engineers" (for example, Cyc [32] or WordNet [35]) are variations of this theme and suffer from the same essential limitations. Dictionary definitions are meaningful to humans in spite of circularity because certain basic words (such as the words infants tend to learn first) hook into non-linguistic experiential knowledge and non-linguistic innate mental structures. How can we design machines that do the same? To address this question, let us shift our attention to a very different kind of machine intelligence: robot perception and control.

Consider the problem of designing a robot that avoids obstacles and navigates to light sources in a room. Robot designers have learned that it is a bad idea to simply tell robots where obstacles and lights are and expect the robot to work. This is because in practice, with high probability, human mediated descriptions will not quite match the state of the actual environment. No matter how accurately we draw a map, and how accurately we provide instructions for navigation, the robot is still likely to fail if it cannot sense the world for itself and adapt its actions accordingly. These are well known lessons in cybernetics and control theory. Closed-loop control systems robustly achieve goals

in the face of uncertain and changing environments. Predictive control strategies are far more effective than reactive ones. Insights into a mathematical basis of teleology derived from developments in control theory are every bit as relevant today as they were sixty years ago [49]. Cyclic interactions between robots and their environment, when well designed, enable a robot to learn, verify, and use world knowledge to pursue goals. I believe we should extend this design philosophy to the domain of language and intentional communication.

A comparison between robotics and NLP provides strong motivation for avoiding knowledge representations which rest on symbolic, human generated descriptions of the world. Language processing systems that rely on human mediated symbolic knowledge have no way to verify knowledge, nor any principled direct way to map language to physical entities in the world. An NLP system that is told what the world is like will fail in the same ways that we know that a robot will fail.

1.1 Language is Embedded in the Physical World

In everyday language, it is the rule rather than the exception that speech acts leverage non-linguistic context to convey meaning. Barwise and Perry [6] call this the *efficiency of language* – the same words can mean infinitely different things depending on the situation of their use. For practical reasons, we should care about the absence of computational theories for representing the meaning of situated speech acts. Whether we consider a spoken dialog system in a car, a conversational interface for an assistive robot, or analysis of messages flowing between mobile humans, the technical problems we face return to the same underlying questions regarding the relationship between language and the physical world.

Although we might design *ad hoc* solutions for specific restricted applications, I believe a principled solution to address in-the-world language processing requires a basic rethinking of how machines process language. The theory I develop is motivated by such concerns. This theoretical framework has emerged through practice. Over the past several years, we have implemented a series of systems which learn, generate, and understand simple subsets of language connected to machine perception and action. These engineering activities have been guided by the intuition that language needs to be connected to the real world much the way that infants learn language by connecting words to real, visceral experience. What has been lacking in our work, however, is a coherent way to describe and relate the various systems, and provide a theoretical framework for comparing systems and designing new ones. This paper is an attempt to address this latter concern. No attempt has been made to prove that the theory is complete or correct in any formal sense given the early

stages of the work.

Although this paper is focused on systems with tight physical embeddings, the underlying theoretical framework may be applied to communication tasks in which direct physical grounding is not possible or desirable. The prediction I make, however, is that an approach which is shaped primarily by concerns of physical grounding will lead to a richer and more robust general theory of semantics. This prediction is based on my belief that intentional communication evolved atop layers of sensorimotor control that were shaped by the nature of the physical world. By building systems that must connect internal representations to external physical entities, the evaluation of our designs is directly shaped by the same natural constraints which shaped our own evolution.

1.2 Three Aspects of Meaning

Consider the coffee scenario, illustrated in Figure 2. What do the speaker's words mean, and perhaps more to the point, *how* do they mean what they mean? There seems to be a basic duality in the nature of linguistic meaning. On one hand, the downward pointing arrow suggests that the speech act achieves its meaning by virtue of its "aboutness" relationship with the physical situation shared by communication partners. On the other hand, we can interpret speech acts within a larger theory of goal directed actions taken by rational agents as indicated by the upwards arrow.

Everyday common usage of "meaning" also includes an additional sense, roughly the emotional connotation of something ("My Father gave me that cup – it has great meaning for me"). I believe connotative meanings of this kind are more complex and emerge from more basic aspects of meaning, roughly as a summary statistic of an individual agent's goal-directed experiences. I will thus set aside connotations and focus the more basic aspects of meaning.

Words are about (refer to) entities and situations in the world

The sensorimotor associations of taste and temperature conjured by "coffee" and "cold" rely on agents having similar embodied experiences caused by common underlying aspects of reality (the chemical composition of coffee and the dynamics of heat transfer as they interact with bodily actions and senses).

Language use is situated

The speech act in Figure 2 is an assertion about the state of a very specific part of the world: "this" coffee. The word "coffee" has meaning for the listener because, in part, it is directed towards a particular physical object as jointly conceived by speaker and listener. The words "this" and "is" connect the

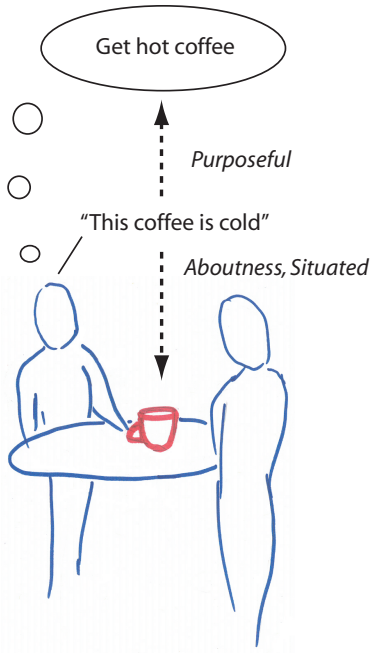

Fig. 2. Three aspects of the meaning of an everyday situated speech act.

speech act to a region of space-time, in this case a part of the agents' here-and-now.

Agents use language to pursue goals

Speech acts can be considered within a broader theory of purposeful action [21]. Beyond the literal meaning of "this coffee is cold" interpreted as an assertion about the state of the world, in certain contexts the speaker may also intend an implied meaning to the effect of "I want hot coffee". Note that even the literal reading of the sentence can be analyzed with respect to the speaker's intentions. For example, the speaker might have just been asked, "Do you know if that coffee is hot or cold?".

I believe that finding a computationally precise and tractable representation of linguistic meaning which covers all three aspects of meaning outlined above is a grand challenge for artificial intelligence and cognitive science. I believe the theory presented here takes steps towards addressing central aspects of this challenge (but of course much much more work remains to be done!). Before getting into the framework, however, it will be useful to define a term which has given rise to much confusion in the past: grounding.

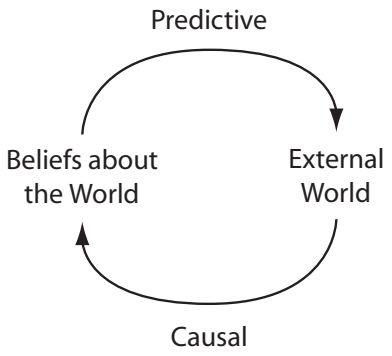

Fig. 3. Grounding is an interactive process of predictive control and causal feedback.

2 A Definition of Grounding

I define grounding as a causal-predictive cycle (Figure 3) by which an agent maintains beliefs about its world. Consider an agent that is situated next to a table that supports a cup. Let "belief" be broadly defined to be an explicit information structure that exists in the mind of the agent [2] (details of the internal structure of beliefs are presented in Sections 4-6). For the agent to hold the grounded belief that *that* particular cup is on the table, two conditions must hold: (1) that cup must have caused the belief via the natural physical laws of the universe (the flow of information via photons, physical contact, sensory transduction, etc.), and (2) the belief must support predictions of future outcomes regarding that cup conditioned on actions which the agent might take. On this definition, the grounding process requires both causal and predictive relations between referent and belief. This cyclic process corresponds to an interpretation-control loop that must be implemented by an agent that holds grounded beliefs.

By virtue of being embedded in a shared physical world, the beliefs of agents are compelled to alignment, providing the basis for coordinated action. Communication gets off the ground because multiple agents can simultaneously hold beliefs grounded in common external entities such as cups of coffee.

I take beliefs about the concrete, physical world of objects, properties, spatial relations, and events to be primary. Agents can of course entertain more abstract beliefs, but these are built upon a physically grounded foundation,

[2] For a simple enough system (e.g., a thermostat), it is perhaps unnecessary to attribute beliefs at all. However, the complex sorts of system I am most interested in (e.g., conversational robots), I prefer to "assume the intentional stance" [13]. Given that my goal is to design and implement such systems, I will take the step (perhaps a controversial step from a philosophical perspective, but a sound step from an engineering perspective) of using a theory of belief as a guide for designing representations and constraints of synthetic agents.

connected perhaps by processes of analogy and metaphor.

The use of metaphor to bridge concrete and abstract domains is supported by Lakoff and Johnson's observations of the ubiquity of physical metaphors appearing in everyday language [31]. For example, we understand and use the metaphor *ideas are food*. By leveraging grounded knowledge regarding food, we can conceptualize the highly abstract concept of an *idea* in sensorimotor (especially gustatory) terms: What he said *left a bad taste in my mouth*. Now there's a theory you can really *sink your teeth into*. That's *food for thought*. This is the *meaty* part of the paper. That argument *smells fishy*. Beyond observations of metaphors in the surface form of language, there is emerging experimental evidence that cross-domain analogical influences affect underlying non-linguistic conceptual structures in humans ([8]). Related to Lakoff and Johnson's observations, Talmy shows that the nature of physical force dynamics (the interaction of forces and objects) appears to structure our conceptions of abstract domains, at least as revealed in how we talk about abstract topics [63]. I expect similar physically grounded structures can be used by machines to conceive of, and communicate about abstract domains.

An agent's basic grounding cycle cannot require mediation by another agent. This requirement excludes many interesting classes of agents that exist in purely virtual worlds. This exclusion is purposeful since my goal is to develop a theory of physically grounded semantics. If A tells B that there is a cup on the table, B's belief about the cup is not directly grounded. If B sees a cup on the table but then permanently loses access to the situation (and can no longer verify the existence of the cup), then B's belief is not directly grounded. I am *not* suggesting that an agent must ground all beliefs – that would lead to a rather myopic agent that only knows about what it has directly experienced and can directly verify. In order to communicate with humans and build higher order beliefs from that communication, an agent must have a subset of its beliefs grounded in the real world without the mediation of other agents. From a practical point of view, the necessity for real world unmediated grounding is well known to roboticists. An autonomous robot simply cannot afford to have a human in the loop interpreting sensory data on its behalf. Furthermore, complex inner representations must be coupled efficiently, perhaps through layering, for operation under real-world uncertainty. For autonomous robots to use language, we have no choice but to deal with internal representations that facilitate conceiving of the world as objects with properties that participate in events caused by agents. The need for unmediated grounding can also be argued from a cognitive development perspective. Infants don't learn language in a vacuum – the meanings of first words are learned in relation to the infant's immediate environment. Language is bootstrapped by non-linguistic experience and non-linguistic innate structures, paving the way for comprehension of dictionary definitions and other sources of ungrounded beliefs. I return to this topic in Section 8.

It is worth heading off one possible criticism of the theory which may arise from a misinterpretation of my definition of grounding. Although we have investigated language learning in several systems (e.g., [50, 54, 52, 51], the focus of this paper is on representational issues and many of the implemented structures that this theory is based on have not yet been learned by any fully automated system. We have instead used a pragmatic approach in which some aspects of a representation (typically topological structure) are designed manually, and machine learning is used to determine settings of parameters only when standard statistical estimation algorithms are easily applicable. The potential criticism arises from the fact that human designers are creating representations for the machines – in essence, it might appear that we are describing the world for the machine – precisely what I said I wanted to avoid. However, there is in fact no contradiction when we consider the definition of grounding carefully. The definition places constraints on the process by which a particular set of beliefs come to be, are verified, and maintained. The definition does *not* make any demands on the source of the underlying design of representational elements. These might be evolved, designed, or discovered by the agent (in terms of some lower level evolved or designed representational primitives). In all of our robotic implementations, the synthetic agent does indeed construct and maintain its representation autonomously, and link language to those belief structures.

2.1 Causal Sensor-Grounding is Not Enough

Based on the definition of grounding I have stated, *causality alone is not a sufficient basis for grounding beliefs*. Grounding also requires prediction of the future with respect to the agent's own actions. The requirement for a predictive representation is a significant departure from purely causal theories. For example, in his 1990 paper on the symbol grounding problem, Harnad suggested a causal solution based on categorical perception of sensor-grounded signals [22]. In my own past work (e.g., [50], [52]) I have used "grounding" to describe language systems with similar bottom-up sensory-grounded word definitions. The problem with ignoring the predictive part of the grounding cycle has sometimes been called the "homunculus problem" (Figure 4). If perception is the act of projecting mental images in an "inner mental theater", who watches the theater [3] ? How do *they* represent what they see? A "pictures in the head" theory without an accompanying theory of interpretation passes the representational buck. The problem of interpretation is simply pushed one layer inwards, but leaves open the question of how those internal models have meaning for the beholder. If the inner process constructs a model of the model, we are led to an infinite regress of nested models which is of course

[3] Dennett calls this the Cartesian theater [14].

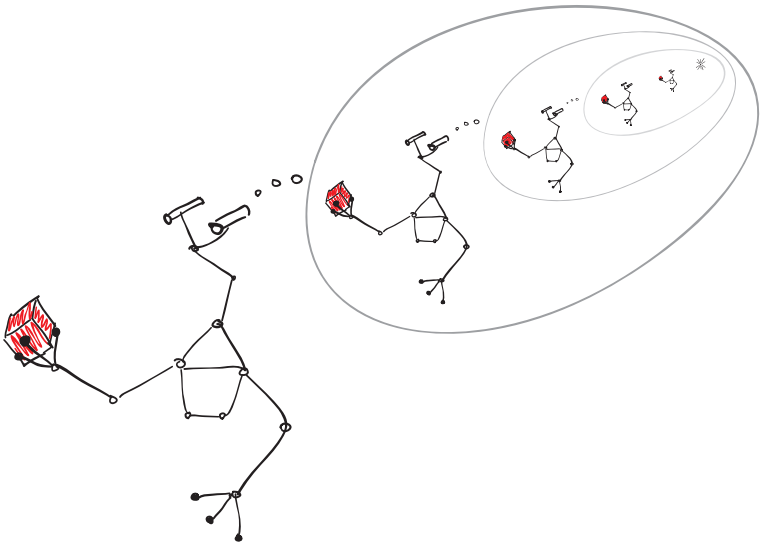

Fig. 4. If representation is treated purely as a problem of description (the "pictures in the head" approach), an agent must have some process within it which interprets the description, essentially pushing the representation problem inwards without actually addressing it. An infinite regress of layered descriptions results without some way to ground out the process.

unsatisfactory.

By requiring that the agent be able to translate beliefs into predictions (not necessarily about the immediate future) with respect to the agent's own actions (where not acting at all is considered a kind of action), we have a broad working definition of interpretation that avoids descriptive regress. Beliefs have meaning for the agent because they have the potential to predict future outcomes of the world, which the agent can verify for itself by comparing predictions to actual sensations. As a result of this framing, beliefs that have no possible impact on the agent's abilities to make predictions about the outcomes of its actions are deemed valueless [4].

3 Desiderata for a Theory of Language Grounding

If a theory of language grounding is to provide the basis for agents to use physically situated natural language, I suggest that it must satisfy three criteria:

(1) Unification of representational primitives: Objects, properties, events, and situations should be constructed from the same set of underlying primitives. This requirement is desirable if we are to have a way for beliefs

[4] This is consistent with Peirce's pragmatic approach to epistemology [42].

about concrete objects (e.g., cups) to be efficiently translated into expectations with respect to actions.

(2) Cross-modal Translatability: Information derived from perception and language should be interpretable into a common representational form since we want to design agents that can talk about what the observe and do.

(3) Integrated space of actions: Motor acts (e.g., leaning over to resolve a visual ambiguity) and speech acts (e.g., asking a question to resolve a visual ambiguity – "is that a cup or a can?") should be expressed in a single integrated space of actions so that an agent may plan jointly with speech and motor acts to pursue goals.

The framework that I will now present is motivated by these requirements. In Section 7 I will assess to what extent each goal has been satisfied. In this first attempt to formalize aspects of what is still very clearly a developing theory, I have chosen to emphasize representational structures, leaving for the future a formalization of operations on these structures.

4 A Theory of Signs, Projections, and Schemas

The theoretical framework is a product of building systems and represents my attempt to explicate the theoretical elements and structures that underlie these complex engineered systems. Rather than separate the description of implementations into another section of the paper, I will highlight relevant implementations in the course of presenting the framework.

4.1 Signs

Signs are a basic building block of the theory. Physical patterns in the world may serve as signs for an interpreter. For example, a particular structured configuration of photons caused by the presence of a fly may be a sign of that fly, if appropriately interpreted by an agent. I take Peirce's definition as a starting point [43]:

> *A sign...is something which stands to somebody for something in some respect or capacity.*

I will interpret Peirce's definition in the following way. A sign is a physical pattern (first instance of "something" in Peirce's definition) which only exists as a sign relative to an interpreter ("somebody"). A sign signifies an object, some entity in the world (second instance of "something"). Signs may take

other signs as their objects, leading to nesting of signs. For example, a shadow might be a sign of a cloud. If the shadow leads to a cooler patch of ground, the temperature of the ground serves as a sign for both the shadow, and chains through to serve as a sign of the cloud. This does not necessarily mean that an interpreter can make the connection from a sign to its object, only that the physical causal link exists. Signs signify (stand for) only limited aspects of their objects ("some respect or capacity") and thus can serve to abstract and reduce information.

4.2 Three Classes of Signs: Natural, Indexical, Intentional

Signs may be classified as *natural*, *intentional*, and *indexical* [5]. This classification scheme is not mutually exclusive – a physical pattern may be interpreted as both a natural and an indexical sign. Natural signs are shaped by nomic physical laws (natural flow of photons, pull of gravity, etc.) whereas intentional signs are generated by volitional agents for some purpose. The configuration of photons signifying the fly is a natural sign. The speech act, "there's a fly!", is an intentional sign. The words exists as a physical pattern of vibrating air molecules, as much a part of the sensible world as photons, but their origins are fundamentally different. The word "fly" signifies the fly by convention and is uttered by a rational agent with some purpose in mind.

Indexical signs situate beliefs relative to a spatiotemporal frame of reference. The location of the fly within an agent's field of view may lead to an indexical sign of its spatial position relative to the viewer's frame of reference. The semantics of indexical signs arise from their use as parameters for control. As we shall see, an indexical that specifies the spatial location of an object may serve as a control parameter in a robot to control reaching and visual saccade behaviors directed towards a target. An alternative approach would be to treat the spatiotemporal location of an object as simply another property of the object like its color or weight. We have found, however, that in construction of robotic systems, separation of spatiotemporal information leads to cleaner conceptual designs.

I will now focus in some detail on natural signs, and how an agent can create beliefs about objects via natural signs. Indexicals will then be folded in, leading to spatiotemporally situated beliefs about objects. Finally, we will consider the comprehension and generation of intentional signs (grounded speech acts).

[5] This classification scheme is related to Peirce's three-way classification of iconic, indexical, and symbolic signs. However, I prefer Ruth Millikan's distinction between natural and intentional signs [37] for reasons I explain in the text.

4.3 Sensing Signs

Sensors transduce signs (physical patterns in the world) into internal signals (for robots, electrical signals) which the agent can further transform, interpret, store, and use to guide actions. The only way for a sign to enter an agent from the environment is through a sensor. The embodiment of an agent determines its sensors and thus directly effects the signs which an agent can pick up.

The agent is attuned to specific channels of sensory input and only detects signs that appear within those channels. Attunement may be innate and unalterable, or determined by the agent's state of attention. For example, an agent can be attuned to high contrast closed forms that are picked out from a visual environment, localized high intensity pains from a haptic environment, or speech signals from an acoustic environment while ignoring other signs from those same channels. Multiple channels can be derived from a single sensor (e.g., color and shape are different input channels, both of which might be derived from the same camera). On the other hand, multiple sensors can contribute to a single input channel [6].

Sensor-derived channels define a continuous space which I will call the channel's domain. Incoming signs project into their corresponding domain. When a sign is detected within a channel to which the agent is attuned, the projection of the sign is called an observation. To take a simple example, imagine a robot which represents the shape of a closed visual region based on the region's maximum height and width (its bounding box). The sign (an optical projection of some underlying object) is transduced by the robot's visual sensor into a pair of continuously varying signals which in a robotic implementation might be a pair of real numbers, h and w. The range of possible values of h and w, and thus the domain of incoming sign observations for this channel, range from 0 to H and W, the height and width of the robot's visual field (measured in pixels). An observation is a particular pair of (h, w) values resulting from a sign.

4.4 Beliefs about Signs (a-signs)

Define an analog belief to be a distribution over all possible observations within an input channel's domain. An analog belief can serve as both an element of memory which encodes a history of observations within a channel, and a prediction of what will be observed within a channel. To be useful in practice,

[6] The superior colliculus of cats contain neurons which only fire under the conditions of simultaneous auditory, visual, and somatosensory evidence [62]. This is an example in nature of multiple sensors leading to a single channel of input.

analog beliefs must be context-dependent. As we shall see, context is defined by the structure of schemas within which beliefs are embedded. To simplify terminology, I will refer to analog beliefs about sign observations within a channel as a-signs. A-signs represent beliefs about signs without any attempt to functionally categorize them. Returning to the earlier example, an a-sign for the shape input channel can be implemented as a probability density function defined over the two-dimensional $H \times W$ domain.

To recap, a-signs are distributions about signs in the environment since they are causally shaped by incoming signs and make predictions of future signs. Natural signs, in turn, are about aspects of their objects by definition – they are causally connected to their objects due to nomic physical conditions of the environment. *By chaining these relations, we see that a-signs are about objects.* A-signs form elements of schemas which enable an agent to both encode causal histories of signs and make context-dependent predictions about the observation of new signs, satisfying the causal-predictive grounding cycle defined in Section 2.

4.5 Sensor Projections

I now introduce a graphical notation of typed nodes and typed edges that I will use to represent schemas. Figure 5 shows the notation for a-signs as ovals. A-signs may have names (the uppercase label A in Figure 5) for notational convenience only. The names are not accessible to the agent. The meaning of an a-sign from the agent's perspective is derived strictly from its function in guiding the agent's interpretative, predictive, and control processes. Figure 5 also introduces a representation of the sensory transduction and observation process as a projection. I will define five more projections as we proceed.

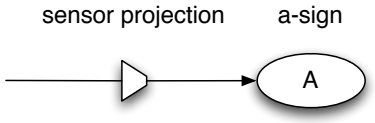

Fig. 5. Graphical notation for a sensor projection connected to an a-sign.

4.6 Schema Types and Tokens

Figure 5 is our first example of a schema, a structured network of beliefs connected by projections. We will encounter a series of schema diagrams of this kind as we progress. The purpose of these diagrams is to show how elements of the theory are combined to implement various functions such as active sensing, representation of actions and objects, and higher level situational,

goal, and linguistic structures. Agents maintain schema types in long term memory schema store. An agent interprets its environment by instantiating, modifying, and destroying schema tokens which are instances of structures such as Figure 5. For example, if an agent is attuned to an input channel represented by the sensor projection in Figure 5, then an observation in this channel may be interpreted by instantiating a token of the schema, resulting in an instantiation of an a-sign. The decision on whether to actually instantiate a schema depends on the control strategy employed by the agent. The structure of possible interpretations of an observation are determined by the contents of the agent's schema store. The contents of the store might be innate, designed, learned, or some combination thereof.

4.7 Transformer Projections

A second type of projection is called a transformer. A transformer performs a mapping from one analog domain to another. Transformers may be used to pick out features of interest from one a-sign to project a new a-sign, or might be used to combine multiple a-signs. An observation from a source domain may be transformed into an observation in a target domain by a transformer. For example, an observation of a shape represented by h and w may be transformed into a one-dimensional domain by taking the product of the terms. In this case, the transformer is simply an analog multiplier. An agent might want to make this transformation in order to ground words such as "large" which depend on surface area. A division transformer (i.e., one that computes the ratio w/h) could be used to ground words which depend on visual aspect ratios such as "long" (for an implementation along these lines, see [51]). The graphical notation for transformers is shown in Figure 6:

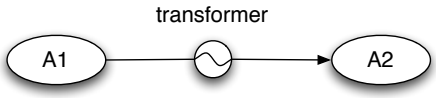

Fig. 6. Graphical notation for a transformer projection which maps a source a-sign, $A1$, to a target a-sign, $A2$.

4.8 Discrete Beliefs About Signs (d-signs), Categorization Projections

The second elementary form of belief is a d-sign, which is a belief about the output of a discrete categorization process which maps a continuous domain a-sign to a discrete domain [7]. Categorization is performed by categorizer projections.

[7] My definition of a-signs and d-signs is similar to Harnad's iconic and categorical signs [22].

The output domain of a categorizer is always a finite discrete set of outcomes. A d-sign is thus a discrete distribution (typically a discrete probability distribution in our implementations). In contrast to a-signs, d-signs rely on categorization – they may be thought of as beliefs about answers to verbal questions one might ask about analog observations (e.g., will the brightness of this patch of pixels will be greater than 0.5? Is this shape a square?). Figure 7 introduces notation for categorizer projections and d-signs.

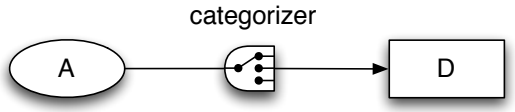

Fig. 7. Graphical notation for a categorizer projection which maps a source a-sign, A, to a target d-sign, D.

The questions implied by the categorization function in a rational agent should be designed to support the selection of actions that the agent must choose from in order to pursue its goals. For example, if a conversational robot needs to distinguish between square and non-square objects (because the robot's human communication partner makes this conceptual distinction and communicates on its basis), then somewhere in the robot, a square / non-square categorizer will be of great utility. In cases where all belief is concentrated on a single discrete outcome, the specific outcome can be given a lowercase label and shown explicitly in the graphical notation as illustrated in Figure 8. The interpretation of this diagram is that the agent believes (remembers, predicts) that the outcome of the categorizer will with high likelihood be the indicated outcome. In underlying implementations, residual belief in other outcomes might be maintained – the notation simply makes the information structure clear for purposes of conceptual design and analysis.

4.9 Empirical Signs in the World vs. Theoretical Signs in the Mind

Consider the distinction between a natural sign on one hand, and an analog observation on the other. The two are linked by analog transductions and transformations, yet signs are physical patterns "out there" in the world, while observations are firmly in the mind of the agent. Next consider the status of an observation that passes through a categorizer. The result is a decision,

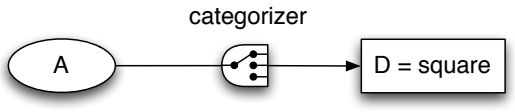

Fig. 8. Graphical notation for a categorizer projection which maps a source a-sign, A, to a target d-sign with concentrated belief in a single outcome. The label of this outcome (*square*) is a notational convenience and is unavailable to the agent.

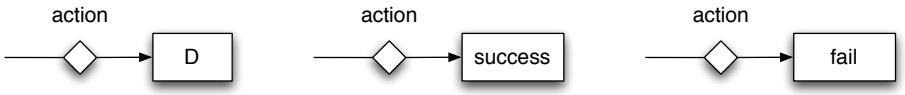

Fig. 9. Graphical notation for action projections.

made on utilitarian grounds, about the observation. For example, imagine a triangle detector which takes as input an analog visual region observation, and outputs a binary triangle / not-a-triangle decision. We can interpret this latter discrete outcome as a "theoretical sign" since it is an inference made by the agent regarding an analog "empirical sign". An agent constructs theories by constructing schemas that include d-signs.

4.10 Action Projections

The specific physical embodiment of an agent gives rise to a natural set of action primitives. For example, the robots we have constructed [50, 52, 45] have servo motors dedicated to each degree-of-freedom (DOF) of the robot. Using standard position-derivative control, each motor is associated with a lowest level action primitive, essentially "move to position x along a specified spatiotemporal path" subject to failure conditions due to unanticipated collisions or other external conditions which require reactive response. When an agent attempts to execute a primitive action, it either succeeds or fails.

Actions provide a new representational element, an action projection, which results in a discrete binary (success / fail) outcome identical in form to the output of categorizer projections. This can be seen in the graphical notation for action projections indicated by diamonds in Figure 9. Actions lead to d-signs, either indicated as distributions over binary outcomes (left most graph) or alternatively, specific beliefs about the success or failure of an action (for notational convenience, I write "success" rather than "D = success").

The use of a d-sign to represent the outcome of an action binds actions into the theory of signs at a most basic level. Each time the agent executes an action primitive, a d-sign observation (about the world it has acted upon) results. Action and sensing are thus intimately intertwined.

4.11 Active Perception / Perceptive Action

Success or failure provides very limited information about an action. In general an agent may want information about the *manner* in which an action succeeds or fails. An agent can achieve this through active sensing – sensing while an action is performed. An example of this arose from experiments with one of our

robots, Ripley [45, 55], which I now briefly introduce. Only details relevant to the development of the theory are mentioned here. More technical descriptions of the robot may be found previous papers.

Ripley, pictured in Figure 10, is a custom-constructed manipulator robot that was designed for grounded language experiments. All of its 7 DOFs are driven by back-drivable, compliant actuators instrumented with position and force sensors, providing the robot with unique proprioceptive sense. Two miniature video cameras are placed at the gripper which also serves as the robot's head (when the robot talks with people, it is hardcoded to look up and "make eye contact", to make spoken interaction more natural). Ripley's gripper fingers are instrumented with force-resistive sensors giving it a sense of touch.

The visual system of the robot includes several low-level image processing routines for segmenting foreground objects from the background based on color, finding closed form connected visual regions, and extracting basic shape and color features from regions. A higher level visual sub-system tracks regions over time and maintains correspondence between regions as the robot's perspective shifts. When a region is detected and tracked over time, we call it an *object* that is instantiated in Ripley's *mental model*. The mental model provides Ripley with object permanence. Ripley can look away from the table (such that all the objects on the table are out of sight), and when it looks back to the table, retain correspondences between objects from before. If a human intervenes and adds, removes, or moves physical objects, Ripley instantiates, destroys, and updates objects in its mental model. Each object in the mental model encodes basic visual attributes of the object (shape, color) and object locations encoded with respect to Ripley's body configuration (we will return to this last point in the discussion on indexical signs in Section 4.12). Ripley's visual system also includes a face tracker to locate the position of its human communication partner. It is able to use this information to modulate spatial language to distinguish, for example, "the cup on my right" from "the cup on your right" [55].

The robot's work space consists of a round table placed directly in front. The robot's motor control system allows it to move around above the table and view the contents of the table from a range of visual perspectives. A visually-servoed procedure lets the robot move its gripper to the centroid of visual regions. Several other motion routines enable the robot to retract to a home position, to lift objects from the table, and to drop them back onto the table.

Ripley understands a limited set of spoken requests. Output from a speech recognizer is processed by a spatial language interpreter [20] which maps requests onto goals with respect to objects in Ripley's mental model. A limited look-ahead planner chooses actions to satisfy goals such as looking at, touching, grasping, lifting, weighing, and moving objects.

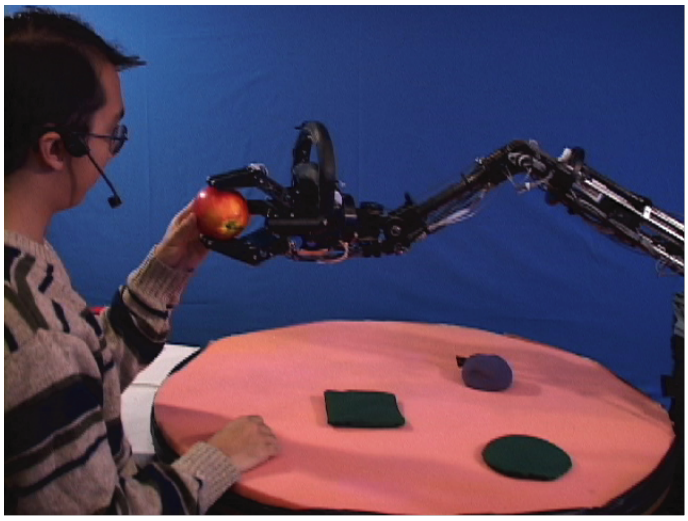

Fig. 10. Ripley is a 7 DOF manipulator robot terminating in a gripper, pictured here handing an apple to its human partner. The human speaks into a head-worn microphone to communicate with the robot. Two video cameras and touch sensors are mounted on the robot's gripper. Each actuated joint contains both a position and a force sensor, providing proprioceptive sensing.

Fig. 11. A schema for active sensing of compliance through grasping.

We are now ready to consider how Ripley might represent the meaning underlying words such as "soft" or "hard" used in their most literal, physical sense. An obvious approach, one that we implemented, is to sense the degree of resistance which is met in the course of gripping. The resistance reading indicates the compliance of the object, providing the basis for grounding words that describe these properties.

Figure 11 shows how to combine some of the elements introduced earlier into a schema to represent active perception required for touching to gauge compliance, providing the basis for grounding words such as "soft" and "hard". The schema may be interpreted as follows. The action primitive *closeGrip*, when executed, runs a motor controller connected to the grip motor. The gripper may or may not reach the targeted position (if the robot successfully grasps a large rigid object, the object will block the gripper from closing). The outcome of the action is represented by the d-sign D. A sensor projection, *senseGripResistance*, is connected to D and projects an a-sign with the designer-friendly (but invisible to agent!) annotation *COMPLIANCE*. The connection from D to the projection is interpreted to mean: run *senseGripResistance* while the source action connected to D is executed.

4.12 Indexical Signs and Schema Parameters

Indexical signs signify spatiotemporal locations – regions of space-time. These signs give rise to beliefs about locations, which in turn provide the grounding for language about space and time which are of course common elements of everyday speech acts. I will use the Ripley implementation once again as an example of how belief structures can be constructed about locations, and then generalize the idea to develop the theoretical framework.

To represent a belief about spatial location, consider how Ripley perceives indexical signs of objects such as cups. For Ripley to move its gripper to touch a cup, it must set 6 joint angles appropriately (the 7th joint is the gripper open / close angle). When Ripley touches an object, the six-dimensional joint configuration at the moment of contact provides an encoding of the object's location. Similarly, when Ripley looks around the table and detects that same object, again its 6 joint angles encode position when combined with the two-dimensional coordinates of the object's visual region within Ripley's visual field, leading to an eight-dimensional representation of space. To connect these two representations of spatial location, we implemented a coordinate translation algorithm using principles of forward kinematics and optical projection combined with knowledge of Ripley's physical embodiment. All object positions, regardless of which modality detected them, are transformed into a two-dimensional space corresponding roughly to positions on the surface of the robot's work surface. As currently implemented, the location of an object is represented deterministically. However, similar to Isla and Blumberg [24], we can extend the implementation to a support a probabilistic representation of spatial location by assigning a distribution over possible two-dimensional positions.

When an object is detected by Ripley through touch, the configuration of the robot's body provides a six-dimensional value which is an observation of the indexical sign originating from the physical object. We can consider body pose to be an input channel, and the proprioceptive sensor reading to be an observation of an indexical sign. The domain of the input channel spans Ripley's permissible body poses. A transformer projection maps indexical observations into a two-dimensional domain, which can be transformed again to guide grasping or visual targeting.

To generalize, just as in the case of natural signs, an agent may hold beliefs about indexical signs using the same forms of representation, a-signs and d-signs. Indexical a-signs are distributions over possible locations within a continuous spatiotemporal domain. Indexical d-signs are distributions over discrete spatiotemporal categories. D-signs can be used to represent relative temporal and spatial relationships such as Allen's temporal relations [1] or

topological spatial relations [47].

4.13 Parameters in Schemas

We can design a search routine, which I will call *detectHandContact*, that requires a parameter L, an a-sign defined over a location domain that the agent can map into arm positions. The routine $detectHandContact(L)$ is not an action primitive, but instead implements an iterative search procedure in which the peak value in L is used to select where to reach, and if no hand contact is detected, the region of L around that peak is set to 0, and the next highest peak in L is tried.

The same a-sign that guides the hand control routine can also be used to drive a visual routine, $detectVisualRegion(L)$ which performs a similar visual search thorugh the control of visual saccades. Isla and Blumberg [24] have implemented something along these lines in their virtual three-dimensional environment. As we shall see in Section 5, the use of a shared indexical a-sign as the control parameter for multimodal action routines provides a basis for deriving a sensorimotor grounded semantics of spatial location which can be extended to represent location in space and time.

4.14 Complex Actions and Abstraction

Building on the idea of parameterized actions, we can now construct structured schemas representing complex actions which will provide the basis for grounding concrete action verbs. The top schema in Figure 12 gives an example of a schema for lifting. Scanning the schema from the left, when interpreted as a control procedure, to lift means to search and find the object (using the a-sign $L1$ to guide the haptic search), close the gripper, query the gripper touch sensors, make sure a stable grip is found, and then to move the gripper to a new location specified by the peak value of another a-sign parameter, $L2$. The same schema can be denoted by an abstracted schema (bottom) which shows a single action projection that carries the designer-friendly label *lift* and its two indexical a-sign parameters, the source and destination locations. Note that other implementations of lifting which differ from the top schema, but which take the same input parameters and lead to the same change in situations can be represented by the single schema at bottom. The abstraction process suppresses "by means of" details and retains only the parametric form of the whole.

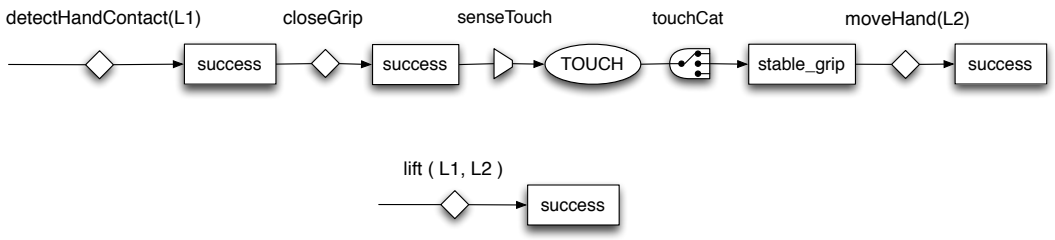

Fig. 12. Top: Schema for lift; $L1$ specifies a distribution over possible start locations and $L2$ specifies a distribution over the target completion locations. Bottom: Abstracted representation of the same schema.

5 Schematization of Objects

> *Consider what effects, which might conceivably have practical bearings, we conceive the object of our conception to have. Then, our conception of these effects is the whole of our conception of the object. (Charles Sanders Peirce, 1878).*

We are now able to construct schemas for tangible physical objects using a combination of natural a-signs, natural d-signs, indexical a-signs, sensor projections, categorizer projections, and action projections. We have already seen some examples of schemas for properties (Figure 11) and complex actions (Figure 12). Object schemas subsume action and property schemas. This is in contrast to many previous computational interpretations of schema theory (e.g., [57, 56]) which take objects as representational primitives distinct from the actions that act upon them. I believe that for an agent to efficiently generate affordances [8] of novel situations for dynamically changing goals on the fly, the only practical option is to represent objects, actions, and goals with a common set of lower level primitives.

My approach to the construction of objects from sensorimotor grounded primitives is consistent with Drescher's approach [16]. Drescher's schema mechanism represents an object as a set of expected interactions with the environment. Drescher, however, chose not to allow parameterization and other structuring elements to enter his framework, which led to difficulties in scaling the representation to higher order concepts of the kind I seek to address. Smith's conception of the "intentional dance" [60] has also directly influenced my approach to object perception and conception as a dynamic, constructive process.

[8] *Affordances* is used here as defined by J.J. Gibson to be a function of both the external real situation and the goals and abilities of the agent [19].

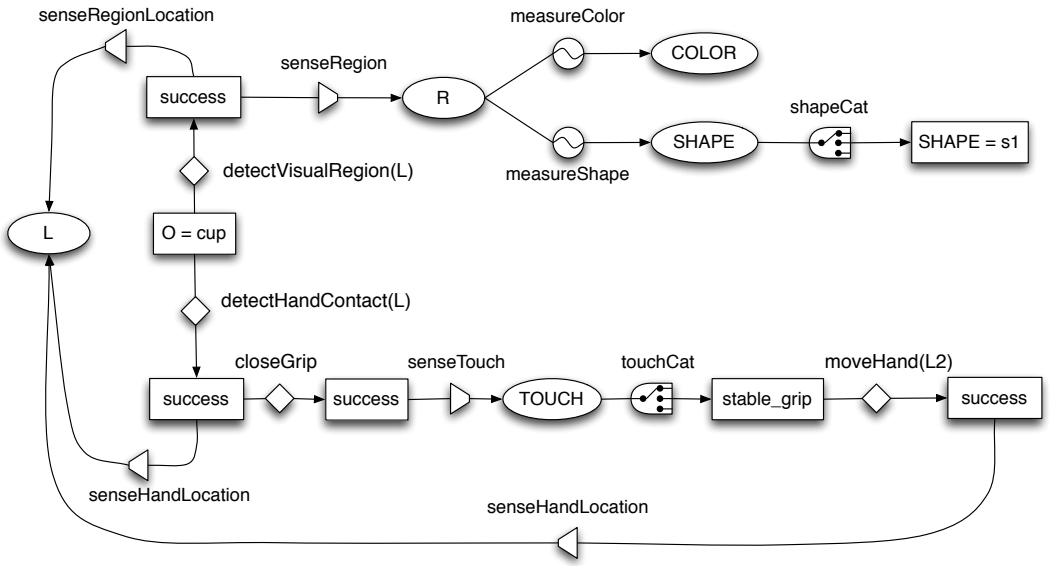

Fig. 13. Schema for a tangible (touchable, graspable, moveable, visible) object such as a cup.

Figure 13 illustrates a schema for a tangible physical object such as a cup [9]. A functionally equivalent structure has been implemented in Ripley as the object permanence part of the robot's mental model which coordinates visual perception, motor control for grasping, and referent binding for speech based understanding of directives [55].

Let us walk through the main paths of this schema to see how it works. The handle of this schema is the d-sign labeled $O = cup$. Symbolic names (e.g., "cup") will be attached to handles of schema types. The domain of O is a discrete set of possible objects known to the agent [10]. The label $O = cup$ indicates that this schema encodes beliefs that are held in the case for which belief within the domain of O is concentrated on the outcome *cup*. As with all labels, these are provided for us to design and analyze schemas. From the agent's perspective, O is simply a d-sign which gets its semantics from its relations to other elements of the schema.

Two action projections connect to the schema handle O. Following first the top action projection, $detectVisualRegion(L)$ projects a binary accomplishment d-sign. Two sensor projections emanate from this d-sign. The first, *senseRegionLocation* feeds back the actual location at which a visual region is found to update L. An agent can execute this path, for instance, to actively track the location of an object. The *senseRegion* sensor is attuned to the

[9] Arbib, Iberall, and Lyons have also suggested detailed schemas for multimodal integration of vision and grasping of objects such as cups [3].
[10] Of course the agent may be able to learn new categories of objects and thus increase the span of the domain over time.

output of $detectVisualRegion$ and projects R, an a-sign with a domain over possible region geometries. Two transformers project (extract) analog color and shape information about R onto separate a-signs. A categorizer projects the shape a-sign onto a specific shape category outcome, $s1$ which corresponds to the shape of cups (if the distribution of belief in O was concentrated on a different object type, say balls, then the distribution over the $SHAPE$ d-sign would shift as well). To specify a cup of a particular color, the distribution of belief would simply be shifted accordingly in the $COLOR$ a-sign.

The lower pathway of the schema may look familiar – it is an embedding of the lift schema that we have already seen (Figure 12). Two feedback loops are used to update L based on haptic sensation using the $senseHandLocation$ sensory projection. The indexical L can serve as a coordinator between modalities. In Ripley, for example, we have implemented a coarse-grained vision-servoed grasping routine which relies on the fact that a single spatial indexical coherently binds the expected success locations for vision and touch.

The object schema is an organizing structure which encodes various causal dependencies between different actions that the agent can take and expectations of sensory feedback given that a cup actually exists at L. To believe that a cup is at L, the agent would be committed to the expectations encoded in this schema. If the agent executed some of the action projections of the schema and encountered a failure d-sign, this would provide cause for the agent to decrease its belief that $O = cup$. Conversely, if the agent is unaware of the presence of a cup, it may inadvertently discover evidence which leads it to instantiate this schema and thus develop a new belief that there is a cup at L.

The object schema serves as a control structure for guiding action. Embedded in the network are instructions for multimodal active perception and manipulation directed towards the object. Given a goal with respect to the object (e.g., finding out what its color is, or moving it to a new location), the schema provides predictions of sequences of actions which will obtain the desired results.

A central aspect of the concept of a cup, that its function is to carry stuff, is not yet captured in this schema. To represent this, containment of objects relative to other objects must be represented (Section 5.3).

5.1 Construction of Objects: Individuation and Tracking

To perceive an object, the agent must instantiate a schema that stands for that object. A particular internal information structure within the agent serves as an "absorber" for signs from the environment which the agent attributes to an

individual object. It is by virtue of maintaining a particular mental absorber over time that the agent conceptualizes individuals over time. These internal structures stand for entities in the world and provide the agent with a basis for grounding names and categorical labels that refer to the entities.

Partial evidence may cause an agent to instantiate a complex schema token that makes various predictions about possible interactions with the object. The schema is grounded in the actual object because (1) physical signs caused by the object are transduced by the agent and interpreted into schemas, and (2) these schemas in turn generate a cluster of expectations of future interactions with the object as observed through future signs.

5.2 Ambiguity in Interpretation

A sign may give rise to multiple possible interpretations. For instance, any tangible object may be placed within an agent's path leading to physical contact. The resulting unanticipated d-sign (from, say, $detectHandContact$) might have been caused by any physical object, not just a cup. Prior context-dependent beliefs encoded as a distribution over O play an important role in such cases. If the agent has an a priori basis for limiting priors to a reduced set of objects, then ambiguity is reduced at the outset. If an agent knows it is indoors, the priors on things usually found outdoors can be reduced.

Regardless of how low the entropy of an agent's priors, sensory aliasing is a fact of life. A circular visual region impinging on an agent's retina might signal the presence of a ball, a can viewed from above, a flat disc, and so forth. In response, the agent might activate multiple schemas, one for each significantly likely interpretation. If the potential number of schemas is too large, a pragmatic approach for the agent might be to instantiate a likely subset, which can be revised on the basis of future observations.

If the agent needs to disambiguate the type of object that caused the ambiguous sign, its course of action lies within the schemas. The instantiated alternative schemas are represented in terms of expected outcomes of actions with respect to the object, and so the agent can choose to execute actions which predict maximally different outcomes for different object classes. For the disc-ball-can problem, simply leaning over to obtain a view from a side perspective will suffice.

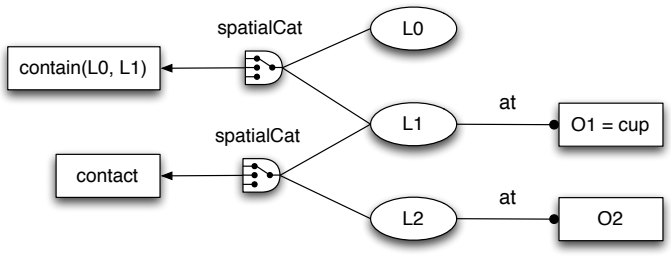

Fig. 14. The situation corresponding to, "There is a cup here. Something is touching the cup."

5.3 Situation Schemas

A situation is represented by linking schemas via their indexical elements. Figure 14 shows the schema corresponding to "There is a cup here. Something is touching the cup". Only the handle d-signs of the objects $O1$ and $O2$ are shown, along with their associated indexical a-signs $L1$ and $L2$. I use the notational shortcut of the *at* link to summarize object schemas by their handles and associated indexical a-sign. No expected outcome of $O2$ is specified, indicating a high entropy belief state with respect to $O2$'s type. A pair of categorizers projects beliefs about spatial relationships between $L1$ and $L2$, and between $L1$ and $L0$. The projected d-sign labeled *contact* serves as a situational constraint and encodes the belief that a contact relationship exists between $L1$ and $L2$.

$L0$ is a default spatial indexical a-sign corresponding to "here". $L0$'s domain spans the default spatial operating range of the agent which depends on the agent's embodiment. A second spatial relation d-sign encodes the belief that the cup is contained within $L0$. For Ripley, $L0$ is the surface of a table in front of Ripley which is the only area that Ripley is able to reach.

To represent "the ball is in the cup", the situational constraint between $L1$ and $L2$ is changed to $contain(L1, L2)$, a topological spatial relation. To reason about embedded indexical relationships during goal pursuit, relational constraints must be taken into account. For example, if the agent wishes to find the ball but can only see the cup, belief in a *containment* or *contact* relationship between the ball's indexical a-sign and the cup's indexical a-sign support the inference that the ball will be found in the proximity of the cup.

5.4 Negation, Disjunction, and Explicit Representations

Certain forms of negation are handled naturally in the proposed framework, others are more problematic. In Ripley's world, some objects can be seen but not touched because they are flat (e.g., photographs). The distinction

between tangible visible objects and intangible yet visible objects is handled by replacing the *success* d-sign projected by $detectHandContact(L)$ in Figure 13 with $fail$, and by removing all outgoing edges from that d-sign. In effect, the schema encodes the belief that the two kinds of objects are identical except that for photographs, the haptic pathway is expected to fail. The intangible object's indexical a-sign L is refreshable only through visual verification.

Difficult cases for handling negation arise from possible world semantics. For example, we might want to tell an agent that "there are no cups here". This sort of negative description is unnatural to represent in the approach I have outlined since the agent explicitly instantiates structures to stand for what it believes to be the case. The representation might be augmented with other forms of belief, perhaps explicit lists of constraints based on negations and disjunctions which are compared against explicit models to look for conflicts, but these directions are beyond the scope of this paper.

Although the difficulty with existential negation and disjunction might seem to be a serious weakness, there is strong evidence that humans suffer from very similar weaknesses. For example Johnson-Laird has amassed evidence [26] that humans make numerous systematic errors in dealing with existential logic that are neatly predicted by a theory of mental models according to which humans generate specific representations of situations and reason with these explicit models even in cases where they know the models are overly specific. Similar constraints on mental representations of machines may lead to better "meeting of the minds" since systems that conceive of their environment in similar ways can talk about them in similar ways. From a computational perspective, I believe my approach is closely related to Levesque's idea of "vivid representations", which have difficulty dealing with certain classes of existential negation and disjunction for similar reasons [33]. Levesque has argued that the choice of vivid representations is defendable when practical concerns of computational tractability are taken into account.

5.5 Event Schemas

Events are partial changes in situations. In Figure 15, an indexical anchor for time binds groups of spatial a-signs from two situations at two different points in time (indicated by the two large rectangular frames). Temporal a-signs ($T1$ and $T2$) encode regions along a one-dimensional local timeline exactly analogous to the two-dimensional spatial domain for spatial indexicals in Ripley. A temporal categorizer *temporalCat* projects the d-sign *after(T2, T1)*, specifying that the situation on the right follows in time.

In the initial situation at $T1$, a ball is believed to be contained in a cup,

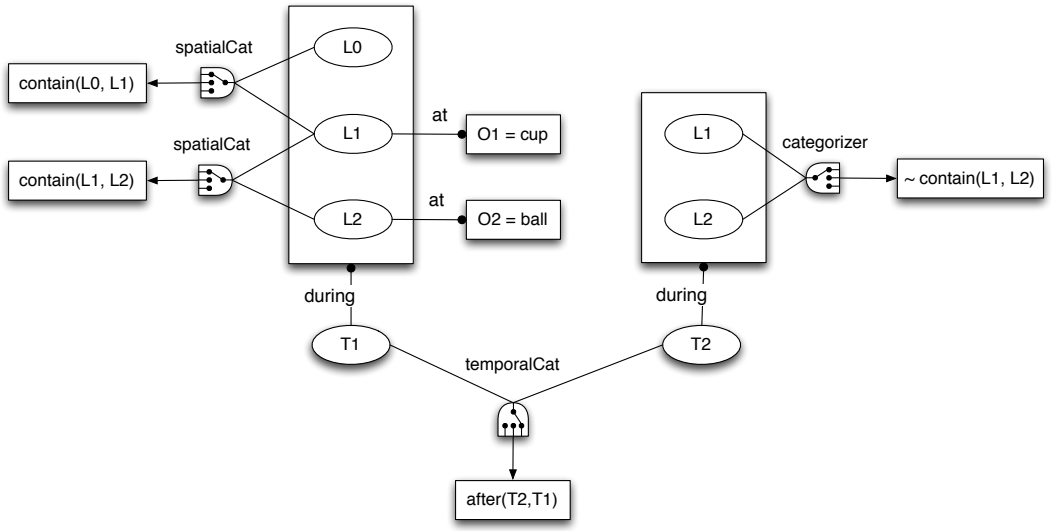

Fig. 15. The situation corresponding to, "The ball was in the cup. Now it is not."

which is contained in the default spatial domain. At some later time $T2$, the containment relation between the ball and cup becomes invalid – the agent places zero belief in the outcome $contain(L1, L2)$. Only changes from $T1$ are indicated in the situation for $T2$ – all other aspects of the original situation are assumed unchanged. Like actions, events may be represented at higher levels of abstraction to suppress "by means of" details, retaining only higher level representations about changes of state. At one level the particular trajectory of the motion of a cup might be specified, at a higher level only the before-after change in position and orientation.

6 Intentional Signs

The representational foundations are finally in place to address the motivation behind this entire theoretical construction: grounding language. Recall that there are three classes of signs. We have covered natural and indexical signs. The final class of signs, intentional signs, are used by agents for goal-driven communication.

Speech acts are the canonical intentional sign. Viewed from a Gricean perspective [21], speech acts are chosen by rational agents in pursuit of goals. I say "the coffee is cold" to convince you of that fact, and by Gricean implicature, to issue a directive to you to get me hot coffee. Intentional signs are emitted by an agent into the environment. Like all signs, intentional signs are physical patterns that stand for something to someone. In the case of intentional signs, as opposed to natural and indexical signs, the link from sign to signified is established by conventions agreed upon by a community of intentional sign

users. Gestures such as pointing may also constitute intentional signs but are not addressed here.

Speech acts are assembled from lexical units (words and other elements of the lexicon) using a grammar. Since my focus will be on primitive descriptive and directive speech acts, sophisticated grammars are not needed, only basic rules that map serial order to and from thematic role assignments. For this reason I will not say much more about parsing and grammar construction here [11] but instead simply assume the requisite primitive grammar is available to the agent.

6.1 Lexical Units (a.k.a. Words)

Figure 16 shows the internal representational structure of a word. A fifth type of projection is introduced in this figure, an intentional projection. This projection is an indicator to the agent that the sign projected by it, in this case the d-sign labeled *"cup"*, is a conventional projection, i.e., one that can only be interpreted in the context of communicative acts. Intentional projections block interpretative processes used for natural signs since hearing "cup" is not the same as seeing a cup. Hearing the surface form of the word that denotes cups will be interpreted differently depending on the speech act within which it is embedded in (consider "there is a cup here" versus "where is my cup?").

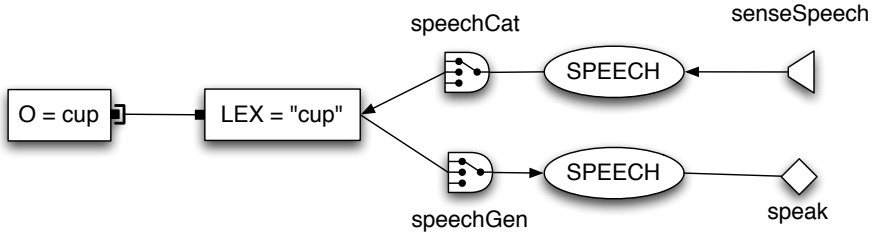

Fig. 16. The structure of a grounded word.

The domain of *LEX* in Figure 16 is the discrete set of all lexical units known to the agent. Like all instances of d-signs, an instance (token) of *LEX* is a theoretical sign (Section 4.9), one that only exists in the mind of agents. The agent using the schema in Figure 16 is able to convert discrete lexical units into surface forms in order to emit them into the environment through *speak* actions, hopefully in earshot of other agents attuned to speech through *senseSpeech* or functionally equivalent sensor projections. The *speechCat* categorizer has been implemented in our systems using standard statistical methods of continuous speech recognition using hidden Markov models. To invert the process, a sixth

[11] Elsewhere, we have explored the relationship between grammar acquisition and visual context [50, 51] and the interaction of visual context on parsing of text [20] and speech [53].

and final type of projection is introduced. *speechGen* is a generator projection which renders a representative of a discrete class. In this case, *speechGen* can be implemented using speech synthesis techniques. In previous work on visual-context guided word acquisition [50], we implemented word structures that are consistent with the schema in Figure 16 in which *speechGen* and *speechCat* shared acoustic models of word surface forms.

Word schemas can be connected to various schemas within an agent's store of schema types. I have made suggestions along the way for ways that schemas provide a grounding for lexical units which serve as labels for the underlying schema. We have seen examples of schemas for properties including visual property names ("red", "round"), affordance terms ("soft", "heavy"), spatial and temporal relation labels ("touching", "before"), verbs ("lift", "move"), and nouns ("cup", "thing"). In addition, the very notion of an individual arises from the act of instantiating and maintaining particular schemas, providing the basis for proper nouns and binding of indexical terms ("that cup", or more persistent proper names).

6.2 Using Speech Acts

As a basic classification schema for speech acts, Millikan [37] has suggested just two classes, descriptives and directives. Descriptives are assertions about the state of the world and are thus akin to natural signs (assuming the speaker can be trusted). Directives are fundamentally different – they are requests for action (including questions, which are requests for information). A speech act may be both descriptive and directive. In the situation depicted in Figure 2, "This coffee is cold" is a descriptive (it describes the temperature of a particular volume of liquid) and perhaps also a directive (it may imply a request for hot coffee). In systems we have constructed to date, only the more literal interpretation of speech acts have been addressed, thus I will limit the following discussion to this simplest case. I first discuss how the framework handles directives and then descriptives.

Directives are understood by agents by translating words into goals. The agent's planning mechanisms must then select actions to pursue those goals in context-appropriate ways. This approach suggests a control-theoretic view of language understanding. If we view a goal-directed agent as a homeostasis seeking organism, directive speech acts are translated by the agent into partial shifts in goal states which effectively perturb the organisms out of homeostasis. This perturbation causes the agent to act in order to regain homeostasis.

In our schema notation, a goal is represented using a dashed outline for the appropriate a-signs or d-signs which the agent must satisfy in order to sat-

isfy the goal. A transition ending in a spreading set of three rays (an iconic reminder that goals are reached for) connects the a-sign as it is currently believed to be to the desired target value. In Figure 17, the agent has set the goal of changing the cup's location such that the containment relation holds. This corresponds the directive, "Put the cup on the plate".

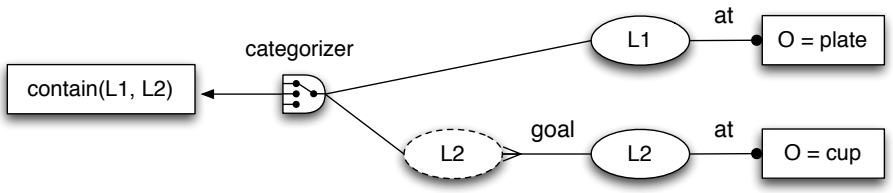

Fig. 17. "Put the cup on the plate."

Ripley understands limited forms of directives such as, "touch the bean bag on the left", or, "pick up the blue one". To perform the mapping from speech acts to goals, output of a small vocabulary speech recognizer is processed by a parser [20] which is integrated with Ripley's control system and mental model architecture [55]. In some cases, a directive will lead Ripley to collect additional information about its environment before pursuing the goal set by the directive. For example, if the robot is directed to "hand me the heavy one", but the weight of the objects in view are unknown, Ripley's planning system uses the implicit control structure of the schema underlying "heavy" [12] to lift and weigh each candidate object to determine which best fits the bill. Details of Ripley's planning algorithms are forthcoming. There are of course many other kinds of directives, but in essence, I believe treating the comprehension of directives as a problem of translation into goal schemas is a productive path forward (see also [28]).

A higher order verbal behavior, one that we have not yet explored, is the generation of directives. To accomplish this in a principled way, the agent must be able to plan with the use of instruments, and treat communication partners as instruments which can be controlled by setting their goals through speech acts.

Understanding descriptive speech acts is treated in a similar vein as interpreting natural signs since both provide information about the state of the world. The most interesting challenge in understanding descriptive acts is the problem of under-specification in linguistic descriptions. "The cup is on the table" tells us nothing about the color, size, orientation, or precise location of the cup. Looking at a cup on the table seems to provide all of this information at first sight, although change blindness experiments demonstrate that even

[12] The words "heavy" and "light" are grounded in active perception schemas similar to those for "soft" and "hard" shown in Figure 11. Accumulation of joint forces during lifting project the weight of objects.

short term memory encoding is highly goal-dependent (I might recall meeting someone yesterday and the topic of our conversation, but not the color of her shirt). The framework allows for various forms of descriptive underspecification. For example, to express uncertainty of spatial location, belief can be spread with high entropy across the domain of an indexical a-sign.

Generation of descriptive speech acts, like the generation of directives, requires some notion of "theory of other minds" in order to be able to anticipate effective word choices for communication descriptions. The Describer system [51] uses an anticipation strategy to weed out descriptions of objects which the system predicts will be found ambiguous by listeners.

There are numerous ideas which we could explore at this point ranging from context-dependency of word meanings (categorizers may receive bias shift signals from other categorizers, for example, to differentiate heavy feathers from light elephants) to the definition of connotative meanings for an agent (as long term summary statistics of objects, properties, and actions in their likelihood to assist or block goals – heavy objects probably block more goals of a low powered manipulator whose goal is to move things around, so the robot would develop a negative connotation towards the concept underlying "heavy"). Given our lack of specific implementations to support such ideas, however, I will not attempt to elaborate further.

7 Taking Stock

A summary of the elements of the theory provides a complete view of the framework developed thus far:

(1) Three classes of signs, natural, indexical, and intentional, carry different kinds of information to agents.
(2) Agents hold beliefs about the analog form of signs (a-signs), and beliefs about discrete decisions about a-signs (d-signs).
(3) Six projections (sensors, actions, transformers, categorizers, intentional projections, and generators) link beliefs to form schemas.
(4) Schemas may use parameters to control actions.
(5) Objects are represented by networks of interdependent schemas that encode properties and affordances. Object schemas subsume property and action schemas.
(6) Using schemas, an agent is able to interpret, verify, and guide actions towards objects, object properties, spatiotemporal relations, situations, and events.
(7) Lexical units are pairs of a-signs (encoding surface word forms), d-signs (encoding lexical unit identity) connected to defining schemas through

intentional projections.

(8) Speech acts are intentional signs constructed from lexical units.

(9) Two kinds of intentional signs, descriptive and directive, are used to communicate.

(10) Directive speech acts are interpreted into goal schemas that an agent may choose to pursue.

(11) Descriptive speech acts are interpreted into existential beliefs represented through schemas which are compatible with (and thus may be verified and modified by) sensing and action.

In my introductory remarks I suggested three aspects of meaning in language use: that words are about the world, situated, and purposeful. I defined grounding to be a process of predictive-causal interaction with the physical environment. Finally, I proposed three requirements for any theory of language grounding. Let us briefly review how the theory addresses these points.

All three aspects of meaning are addressed in the framework: (1) Words are about entities and situations in the world. Words project to schemas which are constructed out of beliefs about signs, and signs are about the world through nomic laws. The choice of a word's surface form is arbitrary and conventional, but the underlying mapping of its d-sign is shaped by causal-predictive interactions with the environment. (2) Language use is situated. Indexical a-signs and d-signs that are constructed in the process of using language. (3) Agents use language to pursue goals. Since all schemas may serve as guides for controlling action, and words are defined through schemas, the very representational fabric of word meanings may always be viewed from a functional perspective.

Schemas are networks of beliefs. Beliefs are both memories of what has transpired, and also predictions of what will transpire (contingent on action). This dual use of belief structures satisfies the predictive-causal definition of the grounding process provided in Section 2.

Finally, we may assess the framework with respect to the three requirements proposed in Section 3:

(1) Unification of representational primitives: Objects, properties, and events and higher level structures are all constructed from a unified set of a-signs, d-signs, and six types of projections.

(2) Cross-modal Translatability: Natural signs, indexical signs, and intentional speech acts are interpreted into schemas. Directive speech acts are interpreted as goal schemas. Descriptive speech acts (which are often vague when compared to perceptually derived descriptions) are interpreted into compatible schematized belief structures. In other words, speech acts (intentional signs) are translated into the same representa-

tional primitives as natural and indexical signs.

(3) Integrated space of actions: Although not explored in this paper, the framework lends itself to decision theoretic planning in which the costs and expected payoffs of speech acts and motor acts may be fluidly interleaved during goal pursuit.

8 Social Belief networks

In Section 2 I gave a relatively stringent definition of grounding that requires the believer to have direct causal-predictive interaction with the physical subjects of its beliefs. The theoretical framework I have developed does just this – it provides structured representations of various concepts underlying words and speech acts that are grounded strictly in sensorimotor primitives. But of course most of what we know does not come from first hand experience – we learn by reading, being told, asking questions, and in other ways learning through intentional signs. I argued that to make use of symbolically described information, an agent needs an independent path to verify, acquire, and modify beliefs without intermediaries. Once it has that (and that is what the focus of this paper has been), we can start to see social networks that collectively ground knowledge that not all members of community can ground. I depict such networks of belief amongst agents in Figure 18. Everything we have discussed thus far may be denoted by the graph on the left. It shows a single agent that holds the belief $B(x)$ about the world. The world (denoted as the rectangle with a electrical ground sign at bottom) indeed contains x and causally gives rise to $B(x)$ as indicated by the upwards arrow. The downward arrow from the agent back to the world denotes that the agent has the ability to directly verify $B(x)$ by interacting directly with the physical environment (the agent can, for example, look to verify x using appropriate schemas).

The right panel of Figure 18 shows a community of four agents. Only Agent 1 has full and direct grounded beliefs in x. Agent 2 came to know about x through intentional signs transmitted from Agent 1. Agent 2's only way to verify x is to ask Agent 3. Agent 3 also learned of x from Agent 1, but is able to verify by asking either Agent 3 or Agent 4. This kind of graph is reminiscent of Putman's linguistic division of labour [46] in which an expert about x (Agent 1) grounds beliefs about x on behalf of others in the belief network. The claim I began with is that there exists some basic set of concepts about the world we all share which each agent must ground directly for itself, and that language uses these shared concepts to bootstrap mediated networks such as the right side of Figure 18. The ubiquitous use of physical metaphor in practically all domains of discourse across all world languages [31] is a strong indication that we do in fact rely on physical grounding to get language off the ground.

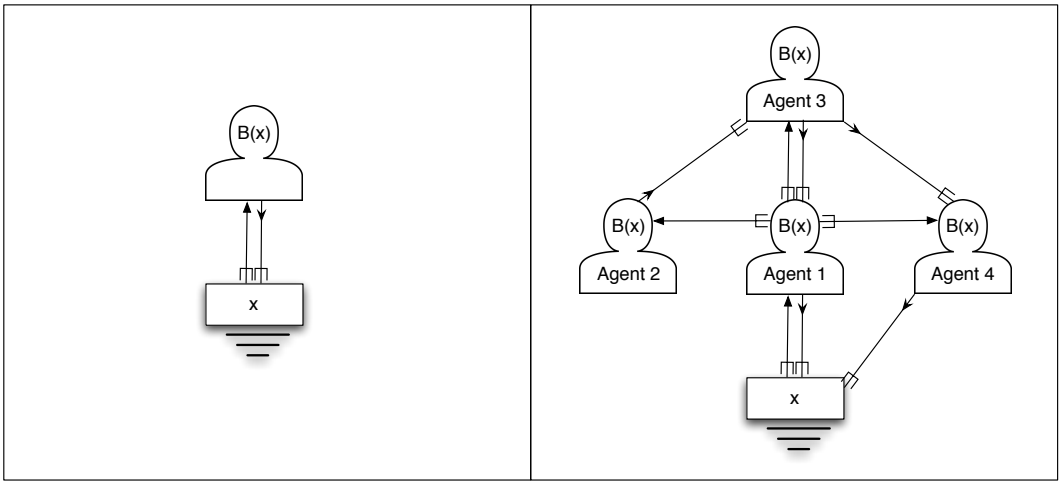

Fig. 18. Social belief networks.

9 Related Ideas

The theory I have presented brings together insights from semiotics (the study of signs) dating back to Peirce with schema theory dating back to Kant to form an original computational framework for grounding language. There is a great deal of prior work on which the theory rests and to which it relates. Rather than attempt a comprehensive survey, I highlight selected work that seems most closely connected and that I have not already mentioned elsewhere in the paper.

Minsky's seminal paper describing frames [38] is very similar in spirit to my approach. Frames are data structures that represent stereotyped situations, and are instantiated to interpret experienced situations much as I have suggested the role of schemas here. Minsky suggests frames as a structure for interpretation, verification, and control as I have for schemas. Minsky's paper covered a far wider range of domains, and thus naturally provided less specific details on any one domain. In contrast, the theory I have outlined is focused specifically on questions of language grounding and reflects specific structures that arose from a concerted effort to build language processing systems.

Schank and Abelson [57] developed an influential theory of scripts which are organizing knowledge structures used to interpret the meaning of sentences. Scripts are highly structured representations of stereotyped situations such as the typical steps involved in eating a meal at a restaurant. Scripts are constructed from a closed set of 11 action primitives but an open set of state elements. For example, to represent the stereotyped activities in a restaurant script, representational state primitives include *hungry*, *menu*, and *where-to-sit*. In contrast, I have suggested a theory which avoids open sets of symbolic primitives in favor of a closed set of embodiment-dependent primitives.

Several strands of work by cognitive scientists and linguists bear directly on the topics I have discussed. Miller and Johnson-Laird compiled perhaps the most comprehensive survey to date of relationships between language and perception [36]. Barsalou's perceptual symbol system proposal [5] stresses the importance of binding symbols to sensorimotor representations, but as Dennett and Viger [15] point out, Barsalou's proposal is more of a specification of a representation than an actual proposal (Dennett and Viger go on to say, "If ever a theory cried out for a computational model, it is here."). Jackendoff [25] presents a compelling view on many aspects of language that have influenced my approach, particularly his ideas on "pushing 'the world' into the mind", i.e., treating semantics from a subjective perspective.

Some noteworthy approaches in the robotics community are closely related to the use of schemas I have proposed. Kuipers' Semantic Spatial Hierarchy suggests a rich multilayered representation for the representation of space and navigation [30]. This representation provides a basis for causal-predictive grounding in spatial domains which I believe might be of great value for grounding spatial language. Grupen's work on modeling affordances [10] intermingles object and action representations and also deserves further study from a language grounding perspective.

Bailey [4] and Narayanan [39] propose the use of modified forms of Petri nets (a formalism used to model concurrent, asynchronous control flow in networks) to model schema-like structures underlying natural language verbs. Bailey models manipulation actions very similar to those we have experimented with (Bailey's implementations were only simulated, however). Narayanan uses modified Petri nets as a basis for understanding abstract economic news stories by analogy to underlying physical action metaphors (e.g., "the economy hit rock bottom"). Siskind [59] proposes an approach to modeling perceptually grounded representations underlying manipulation verbs by combining force dynamics primitives with Allen's temporal relations. The representation of events proposed by Bailey, Narayanan, and Siskind are all able to model more complex event structures than the approach I have presented here based on sequences of situation schema. However, my approach provides a holistic account for actions *and* other ontological categories such as objects, properties, and spatial relations, whereas these other approaches focus only on event structure. An interesting direction would be to investigate ways to incorporate the more expressive power of Petri nets or Siskind's representation to augment the schema structure while retaining the holistic nature of the framework I have presented.

Littman, Sutton and Singh [34] have proposed the idea of predictive representations of state through which states of a dynamical system are represented as "action conditional predictions of future observations". The exact relationship between those ideas and the ones I have presented will require detailed study,

but it seems to be very similar in spirit if not formulation. Also closely related is Cohen's work with robots that learn "projections as concepts" [11] which have been linked to linguistic labels leading to a limited form of language grounding [12]. Steels and Vogt take an evolutionary perspective on language, grounding the function of language for synthetic agents in terms of adaptive fitness [61].

Perhaps the most well known early work in this area is the seminal SHRDLU system constructed by Winograd [64]. This work demonstrated the power of tight integration of language processing within a planning framework. Thirteen years later, Winograd and Flores [65] wrote a penetrating critique of symbolic AI including the methods employed by SHRDLU. The basic gist of this critique is to point out the interpretive "slight of hand" that tends to underlie symbolic AI systems such as SHRDLU. I believe the core issues underlying Flores and Winograd's critique are the same as those which have motivated the framework I have presented.

10 Meaning Machines

There are many important questions that this framework raises that I have not begun to address. Where do schemas made of a-signs, d-signs, and projections come from? How and to what extent can their structure and parameters be learned through experience? How does an agent perform efficient inference and planning with them? How are abstract semantic domains handled? How are higher level event, action, and goal structures organized to support more sophisticated forms of inference and social interaction? These are of course challenging and deep questions that point to the immense number of future directions suggested by this work.

I have taken a holistic perspective on language and meaning rather than drilling deeper into relatively narrow sub-problems. Some readers might object that such approaches are destined to work on only toy problems of at most academic interest, but I believe this is not the case. We are currently exploring a variety of practical applications within this framework. Rather than point to real-world applications, however, a deeper reply to this objection is perhaps best left to Wittgenstein [66]:

> *If we want to study the problems of truth and falsehood, of the agreement and disagreement of propositions with reality, of the nature of assertion, assumption, and question, we shall with great advantage look at primitive forms of language in which these forms of thinking appear without the confusing background of highly complicated processes of thought. When we look at such simple forms of language the mental mist which seems to enshroud*

our ordinary use of language disappears. We see activities, reactions, which are clear-cut and transparent. On the other hand we recognize in these simple processes forms of language not separated by a break from our more complicated ones. We see that we can build up the complicated forms from the primitive ones by gradually adding new forms.

Acknowledgements

I have benefited immensely from interactions with members of the Cognitive Machines Group including Peter Gorniak, Kai-Yuh Hsiao, Nikolaos Mavridis, Andre Ribeiro, Michael Fleischman, Stefanie Tellex, and from discussions with Aaron Sloman.

References

[1] James Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26:832–843, 1983.

[2] Michael A. Arbib. Schema theory. In Michael A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks, 2nd Edition*, pages 993–998. MIT Press, 2003.

[3] Michael A. Arbib, Thea Iberall, and Damian Lyons. Schemas that integrate vision and touch for hand control. In *Vision, Brain, and Cooperative Computation*, pages 489–510. MIT Press, 1987.

[4] David Bailey. *When push comes to shove: A computational model of the role of motor control in the acquisition of action verbs*. PhD thesis, Computer science division, EECS Department, University of California at Berkeley, 1997.

[5] Lawrence Barsalou. Perceptual symbol systems. *Behavioural and Brain Sciences*, 22:577–609, 1999.

[6] Jon Barwise and John Perry. *Situations and Attitudes*. MIT-Bradford, 1983.

[7] Mark H. Bickhard. Representational content in humans and machines. *Journal of Experimental and Theoretical Artificial Intelligence*, 5:285–333, 1993.

[8] Lera Boroditsky. Metaphoric structuring: Understanding time through spatial metaphors. *Cognition*, 75(1):1–28, 2000.

[9] Angelo Cangelosi and Stevan Harnad. The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in perceptual categories. *Evolution of Communication*, 2000/2001.

[10] J. Coelho and J. Piaterand R. Grupen. Developing haptic and visual perceptual categories for reaching and grasping with a humanoid robot.

In *Proceedings of the First IEEE-RAS International Conference on Humanoid Robots*. IEEE-RAS, 2000.

[11] Paul Cohen. Projections as concepts. In *Proceedings of the 2nd European Conference on Cognitive Science*, pages 56–60, 1997.

[12] Paul Cohen, Tim Oates, Carole Beal, and Niall Adams. Contentful mental states for robot baby. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, 2002.

[13] Daniel Dennett. True believers. In *The Intentional Stance*. MIT Press, 1987.

[14] Daniel C. Dennett. *Consciousness Explained*. Little, Brown and Company, 1991.

[15] Daniel C. Dennett and Christopher D. Viger. Sort-of symbols? *Behavioral and Brain Sciences*, 22(4):613, 1999.

[16] Gary Drescher. *Made-up minds*. MIT Press, Cambridge, MA, 1991.

[17] Fred I. Dretske. *Knowledge and the Flow of Information*. MIT Press, 1981.

[18] J. Feldman, G. Lakoff, D. Bailey, S. Narayanan, T. Regier, and A. Stolcke. Lzero: The first five years. *Artificial Intelligence Review*, 10:103–129, 1996.

[19] James J. Gibson. *The Ecological Approach to Visual Perception*. Erlbaum, 1979.

[20] Peter Gorniak and Deb Roy. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21:429–470, 2004.

[21] Paul Grice. Logic and conversation. In p. Cole and J. Morgan, editors, *Syntax and Semantics: Vol 3, Speech Act*, pages 43–58. Academic Press, New York, 1975.

[22] Stevan Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.

[23] Gerd Herzog and Peter Wazinski. VIsual TRAnslator: Linking Perceptions and Natural Language Descriptions. *Artificial Intelligence Review*, 8:175–187, 1994.

[24] Damian Isla and Bruce Blumberg. Object persistence for synthetic creatures. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems*, 2002.

[25] Ray Jackendoff. *Foundations of Language*. Oxford Univeristy Press, Oxford, 2002.

[26] P.N. Johnson-Laird. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge University Press, 1983.

[27] Daniel Jurafsky and James Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall, 2000.

[28] Joshua Juster and Deb Roy. Elvis: Situated speech and gesture understanding for a robotic chandelier. In *Proceedings of the International Conference on Multimodal Interfaces*, 2004.

[29] Immanuel Kant. *Critique of Pure Reason*. Section A137. Cambridge

University Press, 1781 / 1998.
[30] Benjamin Kuipers. The spatial semantic hierarchy. *Artificial Intelligence*, 119:191–233, 2000.
[31] George Lakoff and Mark Johnson. *Metaphors We Live By*. University of Chicago Press, Chicago, 1980.
[32] Douglas Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.
[33] Hector J. Levesque. Making believers out of computers. *Artificial Intelligence*, 30(1):81–108, 1986.
[34] Michael Littman, Rich Sutton, and Satinder Singh. Predictive representations of state. In *Advances in Neural Information Processing Systems*. MIT Press, 2002.
[35] George Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
[36] George Miller and Philip Johnson-Laird. *Language and Perception*. Harvard University Press, 1976.
[37] Ruth Garrett Millikan. *Varieties of Meaning*. MIT Press, 2004.
[38] Marvin Minsky. A framework for representing knowledge. In P. Winston, editor, *The Psychology of Computer Vision*, pages 211–277. McGraw-Hill, New York, 1975.
[39] Srinivas Narayanan. *KARMA: Knowledge-based active representations for metaphor and aspect*. PhD thesis, University of California Berkeley, 1997.
[40] Ulric Neisser. *Cognition and Reality*. Freeman, 1976.
[41] C.K. Odgen and I.A. Richards. *The Meaning of Meaning*. Harcourt, 1923 / 1989.
[42] Charles S. Peirce. How to make our ideas clear. *Popular Science Monthly*, 12:286–302, 1878.
[43] Charles Sanders Peirce. Logic as semiotic: The theory of signs. In *Philosophical writings of Peirce*. Dover, 1897 / 1940.
[44] Jean Piaget. *The Construction of Reality in the Child*. Ballentine, 1954.
[45] Kai-yuh Hsiao, Nikolaos Mavridis, and Deb Roy. Coupling perception and simulation: Steps towards conversational robotics. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, Las Vegas, 2003.
[46] Hilary Putnam. The meaning of 'meaning'. In *Philosophical papers: Volume 2. Mind, language and reality*, pages 215–271. Cambridge University Press, 1975.
[47] D.A. Randell, Z. Cui, and A.G. Cohn. A spatial logic based on regions and connection. In *Proceedings 3rd International Conference on Knowledge Representation and Reasoning*, pages 165–176. Morgan Kaufmann, 1992.
[48] Terry Regier. *The human semantic potential*. MIT Press, Cambridge, MA, 1996.
[49] Arturo Rosenblueth, Norbert Wiener, and Julian Bigelow. Behavior, purpose and teleology. *Philosophy of Science*, 10:18–24, 1943.

[50] Deb Roy. *Learning Words from Sights and Sounds: A Computational Model*. PhD thesis, Massachusetts Institute of Technology, 1999.

[51] Deb Roy. Learning visually-grounded words and syntax for a scene description task. *Computer Speech and Language*, 16(3), 2002.

[52] Deb Roy, Peter Gorniak, Niloy Mukherjee, and Josh Juster. A trainable spoken language understanding system for visual object selection. In *International Conference of Spoken Language Processing*, Denver, 2002.

[53] Deb Roy and Niloy Mukherjee. Visual context driven semantic priming of speech recognition and understanding. *Computer Speech and Language*, In press.

[54] Deb Roy and Alex Pentland. Learning words from sights and sounds: A computational model. *Cognitive Science*, 26(1):113–146, 2002.

[55] Deb Roy, Kai-yuh Hsiao, and Nikolaos Mavridis. Mental imagery for a conversational robot. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 34(3):1374–1383, 2004.

[56] David Rumelhart and Andrew Ortony. The representation of knowledge in memory. In R.C. Anderson, R.J. Spiro, and W.E. Montague, editors, *Schooling and the Acquisition of Knowledge*, pages 99–136. Erlbaum, Hillsdale, N.J., 1977.

[57] Roger Schank and Robert Abelson. *Scripts, Plans, Goals and Understanding*. Erlbaum, 1977.

[58] Jeffrey Siskind. *Naive Physics, Event Perception, Lexical Semantics, and Language Acquisition*. PhD thesis, Massachusetts Institute of Technology, 1992.

[59] Jeffrey Siskind. Grounding the Lexical Semantics of Verbs in Visual Perception using Force Dynamics and Event Logic. *Journal of Artificial Intelligence Research*, 15:31–90, 2001.

[60] Brian Cantwell Smith. *On the Origin of Objects*. MIT Press, 1996.

[61] Luc Steels and Paul Vogt. Grounding adaptive language games in robotic agents. In C. Husbands and I. Harvey, editors, *Proceedings of the 4th European Conference on Artificial Life*. MIT Press, Cambridge, MA, 1997.

[62] Barry E. Stein and M. Alex Meredith. *The Merging of the Senses*. MIT Press, 1993.

[63] Leonard Talmy. Force dynamics in language and cognition. *Cognitive Science*, 12:49–100, 1988.

[64] Terry Winograd. *A Process model of Language Understanding*, pages 152–186. Freeman, 1973.

[65] Terry Winograd and Fernando Flores. *Understanding Computers and Cognition*. Addison Wesley, 1986.

[66] Ludwig Wittgenstein. *The Blue and Brown Books*. Basil Blackwell, Oxford, 1958.