A Mechanistic Model of Three Facets of Meaning

Deb Roy

October 17, 2007

Abstract

This chapter presents a physical-computational model of sensory-motor grounded language interpretation for simple speech acts. The model is based on an implemented conversational robot. It combines a cybernetic closed-loop control architecture with structured conceptual schemas. The interpretation of directive and descriptive speech acts consists of translating utterances into updates of memory systems in the controller. The same memory systems also mediate sensory-motor interactions and thus serve as a cross-modal bridge between language, perception, and action. The referential, functional, and connotative meanings of speech acts emerge from the effects of memory updates on the future dynamics of the controller as it physically interacts with its environment.

1 Introduction

This volume is the result of a meeting which was organized around a contrast between two approaches for analyzing and modeling semantics. In the first camp are cognitive scientists who model linguistic meaning as structural relations between symbols. The term "symbol" in this context is taken to mean, roughly, discrete information elements that may be given word-like labels that make sense to humans (e.g., DOG, IS-A). Examples of this approach includes semantic networks such as WordNet (Miller, 1995) and Cyc (Lenat, 1995) and statistical methods such as Latent Semantic Analysis (Landauer, Foltz, & Laham, 1998). In contrast to this "ungrounded" approach, the grounded camp treats language as part of a larger cognitive system in which semantics depends in part on non-symbolic structures and processes including those related to perception and motor planning. Examples include Bates' grounding of language in sensory-motor schemas (Bates, 1979) and Barsalou's proposal of a "perceptual symbol system" that grounds symbolic structures in sensory-motor simulation (Barsalou, 1999). The grounded camp pays more attention to how symbols arise from, and are connected to interactions with the physical and social environment of the symbol user.

The ungrounded camp has the advantage of mature computational modeling tools and formalisms. Although these have proved to be immensely useful, there are also clear limits to the explanatory power of any approach that deals strictly with relations among word-like symbols. Many aspects of linguistic phenomena that depend on conceptual and physical processes, many of which are centrally involved in child language acquisition, are beyond the scope of purely symbolic analysis. For example, to understand the meaning of the assertion that "there is a cup on the table" includes the ability to translate this utterance into expectations of how the physical environment will look, and the sorts of things that can be done to and with the environment if the utterance is true. Presumably, those working with ungrounded models would readily acknowledge the limits of their approach in dealing with such issues. Thus I am not sure whether there is really a debate to be had between the two camps, but rather a difference of opinion in choosing which parts of an immensely complicated overall problem to focus on.

My contribution will be to offer a new model that I believe makes some progress towards understanding interactions between linguistic, conceptual, and physical levels in strictly mechanistic terms. I chose the phrase "mechanistic model" in the title of this chapter to indicate that the model encompasses both physical processes of sensory and motor transduction and computational processes. My hope is that this model will illuminate some aspects of linguistic meaning related to intentionality, reference, connotations, and illocutionary force that are difficult if not impossible to address using conventional symbolic modeling methods alone, but emerge naturally by emphasizing the underlying processes of grounding.

The model bridges the symbolic realm of language processing with the control-theoretic and machine perception realms of robotics. I will present the model in stages, starting with a simple cybernetic control loop that provides the foundation for goal directed systems. Memory systems and processes of this core model are enriched with schema structures and refined planning mechanisms, yielding a pre-linguistic cognitive architecture that is able to support processing of language about an agent's here-and-now environment. Once linguistic interpretation processes are grafted onto this pre-linguistic substrate, the result is a model of linguistic meaning with unique explanatory power. One novel aspect of the model is an explanation of connotative meaning that is derived directly from the design of the agent control architecture.

The model may be useful for guiding the construction of embodied/situated natural language processing systems (e.g., conversational robots, video game characters, etc.), and to shed light on aspects of human language processing, especially children's language acquisition. My emphasis on cognitive architectures and processes in this paper complements an earlier paper (Roy, 2005) that focuses on schema structures underlying language use.

The remainder of the paper is structured as follows. I begin by suggesting

that a viable strategy for modeling language use is to focus on *simple* language use, such as that of young children. I define three facets of "meaning" that need to be explained, leading to functional specifications for a model of language use framed as semiotic processing. The next section describes the embodiment and behavior of Ripley, a conversational robot built in our lab that serves as a concrete launching point for a more general model that is developed in the next sections. The final sections discuss the meaning of "meaning" suggested by this model.

1.1 Simple Languages

Given the immense complexity of language use, it is critical to focus on a manageable subset of the phenomena if we are to gain traction. Rather than partition the problem along traditional pragmatic-semantic-syntactic-phonological boundaries, we can instead take a "vertical" slice through all of these levels by modeling simple but complete language use. Language use by young children is a paradigm case.

Children acquire language by hearing and using words embedded in the rich context of everyday physical and social interaction. Words have meaning for children not because they have memorized dictionary definitions but rather because they have learned to connect words to experiences in their environment. Language directed to, and produced by young children tends to be tightly bound to the immediate situation. Children talk about the people, objects, and events in the here-and-now. The foundations of linguistic meaning reside, at least in large part, in the cognitive representations and physical processes that enable this kind of situated language use.

Consider the language production capabilities of a normally developing toddler at the two-word phase (e.g., "more milk"). The child's lexicon is tiny; her grammar is trivial compared to that of an adult. She will mainly refer to the here-and-now – more complex reference to the past, future, and to distant and imagined worlds will develop later. She will ignore most social roles in choosing her words to express herself – sensitivity to social roles will also develop in time. In other words, toddlers use *simple* language, simple along dimensions of lexicon size, syntactic complexity, extent of reference, social/cultural sensitivities, and so forth.

For all its simplicity, a toddler's use of language nonetheless demonstrates many of the hallmarks of mature adult language: descriptive and directive speech acts consisting of compositions of symbols that relate to the child's environment and goals. How is it that young children produce and interpret simple speech acts that simultaneously refer (are about something) and serve social functions? How do the child's mental lexicon and grammar interact? How do symbolic (word-level) and subsymbolic (concept-level) processes interact? A detailed mechanistic analysis of speech acts used by children has yet to be offered in the cognitive sciences. Any such model must explain aspects of perception,

memory, motor control, planning, inference, and reasoning capacities that play a role in situated language use. In this paper, I will describe a model of embodied, situated speech act interpretation that is motivated by these questions. The goal of this model is to make progress towards a mechanistic explanation of meaning, so first we need a working definition of "meaning".

1.2 Three Facets of Meaning

To analyze and model the meaning of speech acts produced or interpreted by humans there are (at least) three main facets we need to consider. First, speech acts refer to objects, actions, properties and relations in the physical environment – words have **referential meaning**. Second, speech acts are produced intentionally to achieve effects on the interpreter – speech acts have **functional meaning**. Finally, words and larger linguistic structures have **connotative meaning** where connotation is defined as "an idea or feeling that a word invokes in a person in addition to its literal or primary meaning" (McKean, 2005).

Consider first referential meaning. Taking a semiotic perspective[1] (Figure 1), we can view the symbolic constituents of language (words, clauses, etc.) as a layer of interconnected structures that interacts with a conceptual layer via language interpretation and production processes. The conceptual layer in turn interacts with the physical environment of the language user via perception and motor action. These three layers are essentially an expanded view of Odgen and Richards' classic semiotic triangle (Odgen & Richards, 1923) in which the link from words to the environment are mediated by concepts (or "ideas"). In contrast to the static structural view intimated by the original semiotic triangle, Figure 1 identifies the dynamic processes that link each layer. The semiotic relation brings our attention to referential meaning: language can be used to refer to (and to depict) objects, events, and relations in the physical environment. To ensure that referential meaning is truly grasped by language users, we can require that language users have the ability to verify referential assertions against the physical environment for themselves[2].

Figure 2 shows how all three facets of meaning may be analyzed in the context of a conversational turn. Two people are at a table and there is a cup on the table. The speaker says, "This coffee is cold". What do these words mean to the listener? The referential content of the utterance regards the temperature of the liquid in a particular container. "This" and "is" bind the referential meaning of the utterance to the here-and-now object that is physically accessible to both speakers. The meaning of "coffee" and "cold" have shared

[1] I use the term *semiotic* based on my interpretation of C.S. Peirce which emphasizes the process of interpreting sensory patterns as indications of the future possibilities for action.

[2] This is a strong requirement, one that we certainly would want to relax when we consider reference to referents distant in space, time, and more obviously for fictive/hypothetical referents. But in the simplest case of talk about the here-and-now it is a reasonable requirement, one that rules out any approach that relies merely on perceptual associations without the ability to act.

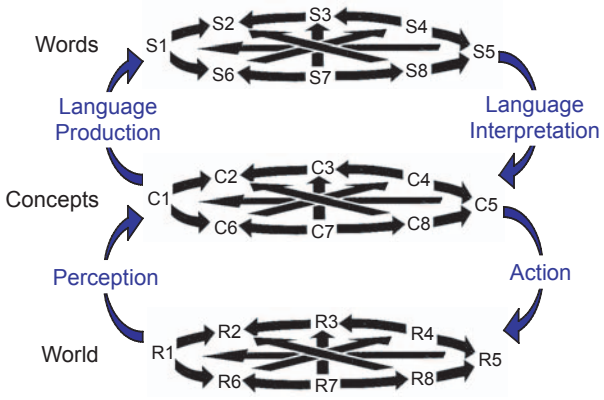

Figure 1: Semiotic processes.

meaning for the speaker and interpreter to the extent that they have associated similar personal experiences to the terms. The connotative meanings of "coffee" might differ markedly and depend on personal tastes. The speaker's purpose in producing this speech act (upper left thought bubble) is layered from higher level to lower level intended functions. If communication succeeds as intended, the cooperative interpreter should infer a set of layered beliefs and goals that are aligned with those of the speaker.

The degree and kind of alignment of speaker and interpreter meaning is different for each facet of meaning. Referential meaning in the depicted situation should overlap substantially. Shared sensory-motor access to a common physical environment forces alignment of referential meaning. Functional meaning should be complementary at matched levels (e.g., A's attempt to convince B that X should translate to B believing that X; A wanting Y should translate to B planning to get Y for A). Connotative meaning may in general diverge significantly as in this example, or alignment might depend on similarities in cultural background or personal experiences.

Within the framework of conventional computational linguistics typified by WordNet and Cyc, my choice of three facets of meaning seems to miss an obvious one: semantic relations. These are diagrammed as arrows within the top layer of Figure 1. Virtually all computational models of semantics, from early semantic networks (Quillian, 1968) to electronic thesauri such as WordNet (Miller, 1995) and commonsense databases such as Cyc (Lenat, 1995) all the way to purely statistical approaches such as latent semantic analysis (Landauer et al., 1998) share one fundamental attribute: they model relations between words. Modeling semantic relations can be traced back to Saussarian tradition of structural analysis and in modern semantic analysis in terms of lexical relations (Cruse, 1986). Without doubt semantic relations are essential for a mature language

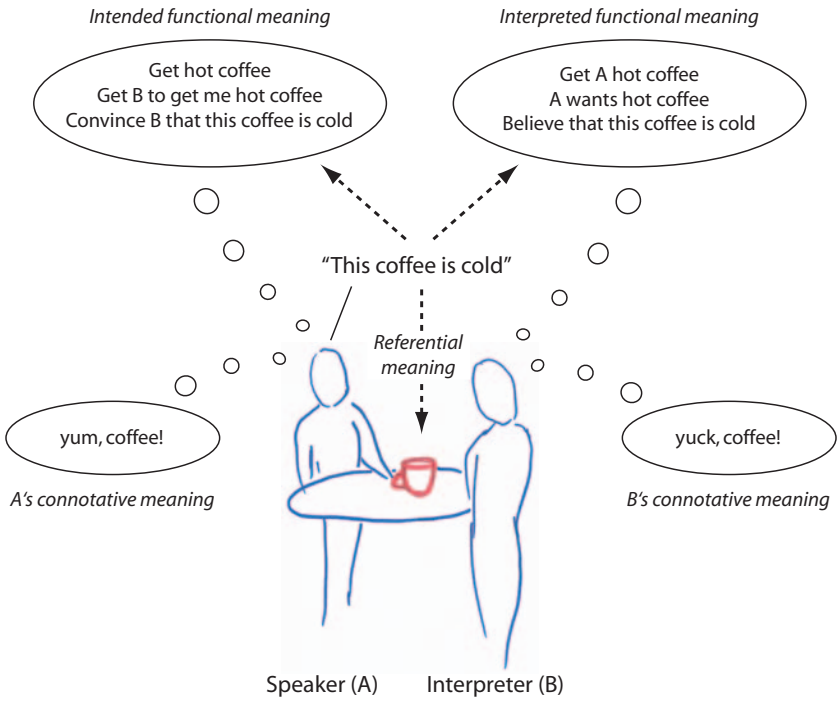

Figure 2: Three facets of linguistic meaning.

user. We would never learn from texts (e.g., text books, dictionaries, news articles, etc.) without them. These relations, however, must ground out in non-linguistic relations. The meaning of semantic relations such as *is-a*, *has*, or *opposite* must be fleshed out in terms of non-linguistic structures and processes. My assumption is that semantic relations are derived from the three core facets of meaning, thus my focus on these foundations. Ultimately, semantic relations (and for that matter, syntactic word classes) will gain some degree of autonomy from underlying conceptual structures[3]. The model I discuss here does not address these more advanced aspects of linguistic structure but prepares a way for them[4].

[3]For example, although we might assume that syntactic word classes are grounded in conceptual categories (nouns derived from objects, verbs from actions, etc.), a fully autonomous syntax moves beyond conceptual groundings (e.g., "the flight" or "cup the ball").

[4]The failure to realize the age-old AI dream of building machines that learn by reading books is perhaps due to an underestimation of the value of first working out the non-linguistic underpinnings of semantic interpretation.

1.3 Modeling Approaches: Modularity vs. Holism

A grand challenge for the cognitive sciences, then, is to develop a mechanistic model that simultaneously addresses all three facets of meaning. This is clearly a tall order that will not be met any time soon, but tangible progress can be made as I hope to show.

A potential objection to taking the grand challenge seriously is the overwhelming number of factors that must be considered together to analyze and model situated language use. When we don't understand how any one "module" works, how can we possibly consider such a challenge? This objection leads to a natural alternative strategy: divide and conquer. Kintsch conveys just this sentiment in his chapter in this volume:

> "...we would have images, concepts, somatic markers, S-R links, features, schemata, abstractions, words, etc., etc., all linked somehow organized in one grand system, at least in principle, for I have no idea how one would go about constructing such a model. Alternatively, one could *model each level of representation separately.*" (emphasis added)

This is in fact a widely held attitude by those interested in developing computationally precise models of language. For example, Chomsky expresses a similar position when considering the prospects for explaining language use (as opposed to mere syntactic competence) (Chomsky, 2000):

> "It would...be a mistake, in considering the nature of performance systems, to move at once to a vacuous "study of everything"...[instead we should] try to *isolate coherent systems* that are amenable to naturalistic inquiry and that yield some aspects of the full complexity." (emphasis added).

The idea of focusing on a single level or component is of course a reasonable strategy. The problem is that we might miss the forest for the trees. For example, although presumably the reason for studying any aspect of language is to understand how words are used to communicate, as research into syntax delves deeper and deeper into the subtleties of "rules" of word order, it is unclear how progress towards understanding these rules in any way furthers our understanding of the process of social communication. I believe that holistic models of language use can be developed if we start with simple kinds of language use and slowly expand the boundaries of analysis to more complex forms. This strategy was anticipated by Wittgenstein (Wittgenstein, 1958):

> "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*." (emphasis added).

1.4 Functional Specifications for a Semiotic Processor

From a designer's perspective, we can translate the task of here-and-now speech act interpretation into a set of functional specifications that must be met by a "semiotic processor". There are two basic kinds of speech acts, descriptives and directives, that I will be concerned with here. Descriptives are used by speakers to make assertions about the current state of affairs, providing the same function as perception. Directives are used to make requests/demands for actions upon the environment. In contrast to descriptives and perceptual input that inform the interpreter of how things *are*, directives suggest how things *should be*.

Imagine a robot that can see its environment through cameras and manipulate objects using its arms. A fly lands next to the robot and the robot happens to be looking in the fly's direction. The fly's presence gives rise to a pattern of activation in the camera. If successfully interpreted, the robot can come to correctly believe that there is a fly out there at some location relative to the robot, with certain properties of size, shape, color, and so forth. If instead of seeing the fly, the robot is told (by a trustworthy source) that "there is a fly next to you", the robot must translate these words into beliefs that are of the *same format* as those caused by the fly's image. If the robot is asked to "swat the fly", those words must be translated into appropriate motor actions selected to make appropriate changes to the environment, and that can be verified by sensory-motor interaction. The robot must translate directives into goals that are in a format compatible with its perceptually-derived beliefs.

This set of cross-modal (perception-action-language) requirements of compatible beliefs and goals constitute the functional specifications of a here-and-now semiotic processor at an abstract level of description.

I will now turn to a specific robotic implementation to make some of these ideas concrete, leading to a generalized model of speech act interpretation.

2 Ripley, A Conversational Robot

Ripley is an interactive robot that integrates visual and haptic perception, manipulation skills, and spoken language understanding. The robot uses a situation model, a kind of working memory, to hold beliefs about its here-and-now physical environment (Roy, Hsiao, & Mavridis, 2004). The mental model supports interpretation of primitive speech acts (Mavridis & Roy, 2006). The robot's behaviors are governed by a motivation system that balances three top-level drives: curiosity about the environment, keeping its motors cool, and doing as told (which includes responding to directives and descriptives) (Hsiao & Roy, 2005). The following sections briefly describe Ripley's embodiment and behavior, which are then translated into a general model that Ripley partially instantiates.

2.1 Ripley's Embodiment

Ripley's body consists of a five-degree of freedom (DOF) arm (or torso) with a gripper at its end. Thought of as an arm, the five DOFs provide a shoulder (two DOFs), elbow (one DOF), and wrist (two DOFs). The gripper has one DOF allowing it to open and close. A pair of miniature video cameras are mounted on either side of the gripper. The torso is anchored to a table. The table top serves as Ripley's primary physical environment. The robot can examine objects on the tabletop via vision and touch, and move objects around using its gripper. Each DOF of the torso includes a position and a force sensor so that the robot can sense both its body pose and the forces that are acting upon it (e.g., due to gravity or if something resists the robot's movements). The gripper tips house pressure sensors that detect contact with objects.

2.2 Ripley's Behavior

The robot is designed for physical and verbal interaction with a human partner who sits across the table from the robot. Using a visual face detector, the robot is able to detect and track the location of the person, using this information to guide its motor movements (e.g., when handing the person an object) and to guide its interpretation of speech (e.g., to differentiate the meaning of "my left" versus "your left"). The following is a typical interaction between Ripley and a human partner:

Ripley: Looks at left side of table, detects a new object that the human has just placed on the table, updates its mental model to encode this new information.

R: Looks to the right, finds that the table is empty there, which matches its expectations based on its current mental model, so no further changes to the model are needed.

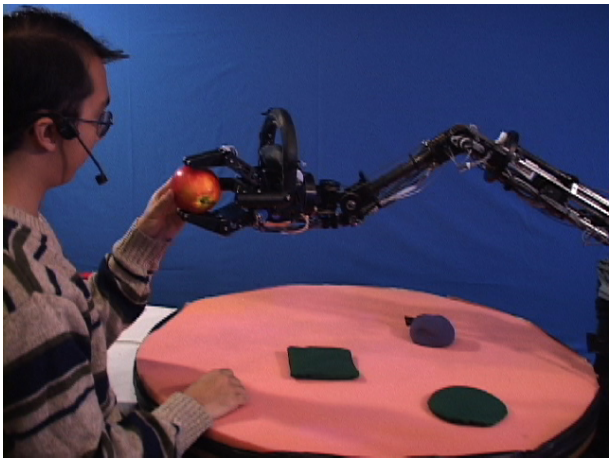

Figure 3: Ripley hands an apple to its human partner in response to a spoken directive.

R: Goes to sleep by positioning its gripper to a resting position on the table and cutting off electrical power to its motors in order to cool them.

Human: "Hand me the blue one"

R: Wakes up, leans down, picks up the blue cup, lifts it and holds it out towards the human.

H: Takes the cup from Ripley.

R: Goes back to sleep.

R: Wakes up by lifting its torso back to an upright position, then looks up towards the human to verify the human's location based on face detection. The person has shifted slightly, so an appropriate update is made to the mental model.

H: "There is a red cup on your left"

R: Updates its mental model to integrate a new belief, that there is a cup on the table but currently not in view, says "OK"

H: "Where is the red cup?"

R: "On the left"

H: "How big is the red cup?"

R: "I have no idea" (Ripley has only been told about the color of the cup, but has no information about its size)

H: "Look to the left"

R: Ripley looks to the left, detects the cup, consolidates new visually derived information (precise location, precise color, size) into mental model.

H: "How big is the red cup?"

R: "It is small"

H: "Pick up the red cup"

R: Attempts three times to grasp object but fails because it is too slippery, updates mental model to record this failure, says "sorry".

The following sections explain Ripley's behavior in terms of cognitive architecture, information structures and processes. This model yields an analysis of all three facets of meaning, from Ripley's point of view. The model focuses on interpretation of speech acts. Although Ripley does generate limited speech output, speech production is not yet addressed by the generalized model.

3 A Model of Situated Language Interpretation

In a previous paper (Roy, 2005) I developed the concept of *semiotic schemas* which "serve as structured beliefs that are grounded in an agent's physical environment through a causal-predictive cycle of action and perception." The focus of that paper was on conceptual structures underlying lexical semantics. I chose the term *semiotic* to highlight the cross-modal (language-action-perception) interpretive and control issues at stake in grasping physically situated language. The following sections build on these ideas but turn attention to the architecture and processes required for semiotic processing. I will treat Ripley as a concrete instance of a class of cognitive architectures. My goal is to develop a model with minimal complexity that yields the capacity for interpreting (i.e., acting successfully in response to) descriptive and directive speech acts about the physical here-and-now.

3.1 Goal Directed Behavior

As we shall see, to account for any of the three facets of meaning, the language interpreter must have its own autonomous interests/goals/purposes. A framework for goal-directed behavior is thus a natural starting point. In cybernetic terms, at the core of the simplest goal driven system is a comparator that drives the agent's actions (Rosenblueth, Wiener, & Bigelow, 1943). A comparator, or difference engine, is shown in Figure 4. Rectangles indicate memory systems and ovals indicate processes. The difference engine compares target and actual situations and generates a plan of action designed to eliminate differences. The

specification of the target situation must be in a format compatible with the contents of the situation model although the levels of specificity may differ (e.g., the target might specify a constraint on acceptable situations but not specify a particular target situation).

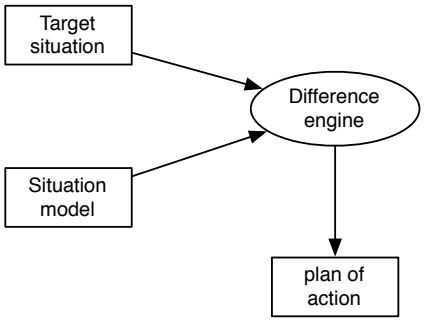

Figure 4: Core elements of a goal driven system.

Since we are interested in physically situated here-and-now semantics, the difference engine must be embodied and embedded in a physical environment. Figure 5 introduces interaction processes that couple the difference engine to the agent's environment. A simple (and classic) example of an embodied, situated difference engine is a thermostat coupled with heating and cooling equipment. Consider how parts of the thermostat system map onto Figure 5. The target situation is a desired temperature, perhaps dialed in by a human. The situation model is a reading of the current air temperature. The body of the thermostat includes a temperature sensor. Bodily interaction includes mapping the sensor reading into a format compatible with the encoding of the target temperature stored in the target situation memory system. The difference engine must compare inputs and decide whether to turn on a furnace, air conditioner, or neither. The output of the difference engine which encodes this three-way decision constitutes its plan of action. The bodily controller (which is part of the bodily interaction processes) executes actions by activating appropriate bodily actuators (furnace, air conditioner). Acting upon and sensing the environment form a cycle, enabling the system to adapt to changes in either the environment or the target temperature.

3.2 Semiotic Schemas for Conceptual Representation and Control

To begin scaling up the capabilities of the basic architecture of Figure 5, we will need to endow our system with richer situation models. Rather than elaborate on specific implementation details, a description of Ripley's software system at

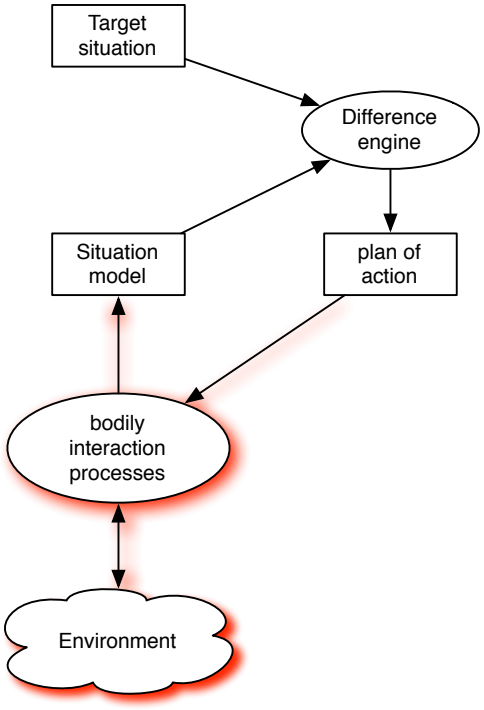

Figure 5: Embodied, situated goal driven system.

the level of semiotic schemas (Roy, 2005) provides a clearer functional explanation suited to our current purposes. Several key attributes of semiotic schemas are:

- Actions, objects, properties, and spatial relations are all represented using a common set of primitives. In other words, rather than treat objects and actions as distinct basic ontological types, both are constructed from the same elementary structures.
- Schemas are used both to encode representations and guide control. Beliefs, goals, and plans are created by instantiating schemas in the situation model, target situation, and plan memory systems respectively.
- Schema *types* encode concepts and are stored in long term memory. Schema tokens are instantiations of schema types. Tokens can be created in target, situation, or plan memory systems. Schemas are instantiated in response to (as an interpretation of) bodily and linguistic interactions with the environment as described below.

- Actions are chains of precondition-movement-result triplets. The result of each action provides the precondition of the next. An action chain may be represented compactly by suppressing all but the initial precondition and final result, providing a causal representation of what the chain achieves without specifying the means by which the change is achieved (Drescher, 1991).

- Objects schemas are collections of possible actions that are licensed by the existence of external objects. As a result, the composition of actions and objects is seamless. To lift a cup is to activate the appropriate action schema from the collection of schemas that represents the cup[5].

- The properties of an object are parameters of schemas which may be measured when those schemas are executed (e.g., the weight of an object is represented in terms of expected forces on joints when that object is lifted).

- Schemas may be used to represent both topological concepts (e.g., that a triangle has three sides corresponding to three connected actions that are spatially mutually constrained) and graded effects (e.g., that blue occupies a certain part of color space with certain parts of the space that are most typical).

An example of an object schema in Ripley is shown in Figure 6. A detailed explanation of the elements of this schema are provided in (Roy, 2005). The main idea is that for Ripley to believe that there is a cup at location L is to believe that two sensory-motor control loops will execute successfully. The upper control loop involves visually detecting a region at location L. The lower loop involves grasping, touching, and moving the object. Both the visual and the haptic loop can effect change in the value of L. Shape and color parameters may optionally be measured while the visual loop is executed. Although not shown in this figure, touch information may also inform the shape parameter (this was not implemented in Ripley since Ripley's touch sensors are too primitive to characterize object shapes).

An example of an action schema corresponding to the verb *lift* is shown in Figure 7. Executing this action causes the lifted object to change from location from L1 to L2. Note that the lift schema is embedded within the cup schema. This embedding relation enables composition of schemas corresponding to the compositional semantics of "lift the cup".

Equipped with schemas, Ripley can move beyond the stimulus-response world of thermostats and represent its physical environment in an object-oriented fashion. In Figure 8, bodily interpretation processes gain access to a long term

[5]In contrast, other models of verb grounding that use schemas (e.g., (Narayanan, 1999; Siskind, 2001) do not provide a direct path for modeling composition with objects that are acted upon. I view this as a serious shortcoming of these alternative approaches.

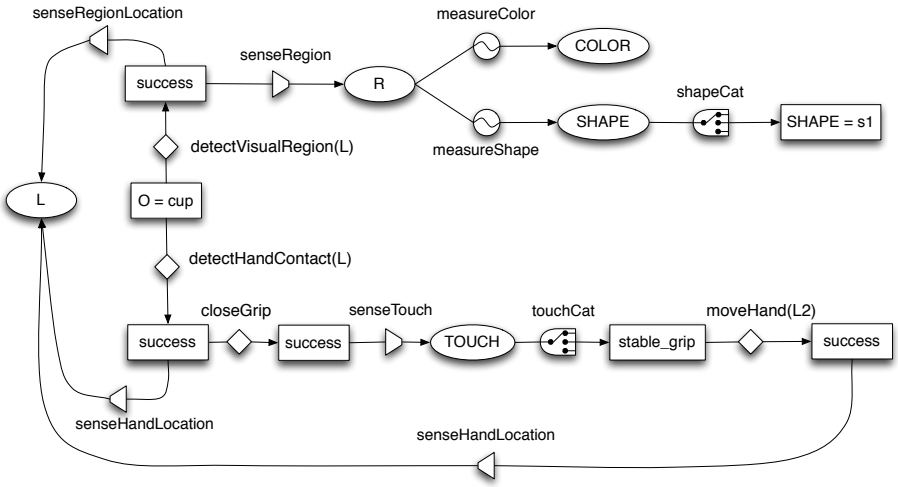

Figure 6: Cup schema.

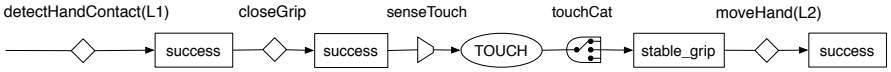

Figure 7: The lift schema is embedded in the cup schema.

memory store of schema types (concepts). Ripley's implemented schema store includes representations of simple handle-able objects, their color, size, weight, and compliance (i.e., soft, hard) properties, and a few actions that may be performed on objects to move them about the table top and to give them to the human partner.

Schema use is goal driven. The first of three top-level drives implemented in Ripley is its curiosity drive. The purpose of the curiosity drive is to maintain an up-to-date set of beliefs about the objects in Ripley's environment. Belief maintenance is challenging for Ripley due to its limited visual field of view. Ripley's cameras allow it to see only a portion of the table top at a time (recall that the cameras are mounted at the end of Ripley's gripper so its visual perspective shifts whenever the robot moves). Also, when Ripley looks up to detect and track humans, its field of view only covers part of the space where the human might stand and loses sight of the table altogether.

Ripley's environment is dynamic since the human partner may move around, and may introduce, remove, and move objects on the table. To keep track of changes, Ripley must continuously visually scan its environment. The curiosity drive is implemented as follows (for more details see Hsiao and Roy (2005)). The visible environment (defined by the full extent of motions of the robot) is

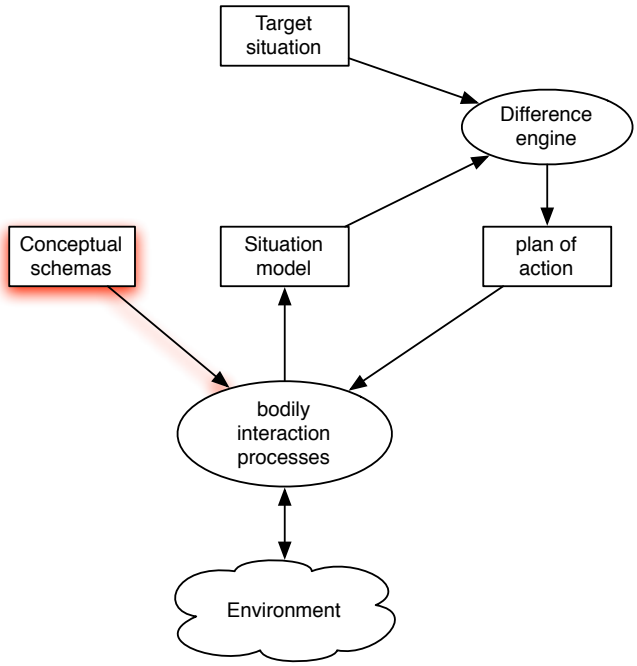

Figure 8: Conceptual schemas in long term memory constrain and accelerate interpretation of bodily interactions with the environment.

partitioned into a number of visual zones. A scalar weight that ranges from 0 to 1 is associated with each zone and tracks when that zone was last updated based on visual input. The longer it has been since Ripley looks at a zone, the higher its associated weight grows. After a sufficiently long period without a visual update, the weight for the zone will hit the ceiling value of 1 and not change further. The collection of weights constitutes Ripley's curiosity and is stored in the situation memory.

A corresponding set of target curiosity values that are stored in the target situation memory (in the current implementation all target curiosity values are set to 0). The difference engine strives to satisfy curiosity by executing a basic control loop in which each iteration consists of selecting the zone with the highest difference between target and actual curiosity levels, and instantiating a plan to eliminate this difference. The plan is constructed by selecting a target pose for the robot that corresponds to the targeted visual zone, computing a sequence of body poses that will guide the body to that pose, and then pausing at that pose for a sufficient period (about 0.5 seconds) for Ripley's visual processes to incorporate new input into the situation model.

3.3 Bodily Interaction Processes using Object Schemas

Object schemas guide Ripley's interactions with the environment via a set of object schema management processes:

Create Instantiates a new object schema token in situation memory when a visual region is detected and no existing token corresponds to (explains away) the region. Correspondence is determined by comparing location, color, and size properties of the region to those predicted by each object already in situation memory.

Destroy Deletes an object schema token from situation memory if sensory-motor evidence does not support existence of an object. This may happen, for example, if an object is removed from the table while Ripley is not looking or if the schema was instantiated due to perceptual errors.

Update Modifies parameters of an existing object schema (e.g., size, color, position) based on current perceptual input. Moving objects are tracked by continuous updates of the position parameter based on visual sensing. Although not yet implemented, changes in parameters over time may be attributed to causes within the object (self-driven or agentive objects) or due to external causes (e.g., naive physics).

Temporal-merge Implements object permanence by tracking objects over perceptual discontinuities. If Ripley looks at an object on the table at time T1, looks away at time T2, and then looks again at the same object at T3, using only the processes listed so far would lead to creation of an object schema (call it Token 1) at time T1, then destruction of Token 1 at T2, and then creation of a new Token 2 at T3. The temporal-merge process detects such sequences and merges Token 2 (and all subsequent tokens) into Token 1. This process provides the basis for individuating objects, and grounding proper names.

Similar processes are used to maintain beliefs about the human's position using a special purpose face detector rather than generic region detectors. The collection of object schemas in situation memory provide referential content for both linguistic communication and motor actions. Various details regarding real-time operation of these bodily control and interpretation processes may be found in (Roy et al., 2004).

Several object schema management processes that have not been implemented in Ripley but a more complete maintenance system for a situation model should include are:

Temporal-split The complement of temporal-merge, this process splits tokens that have been mistakenly merged in the past back into separate tokens.

Spatial-merge type 1 Merges object schema tokens in situation memory that were erroneously instantiated by a single distal object. An object that is visually split by another object in front due to occlusion of it might cause such errors.

Spatial-split type 1 Splits an object schema token into two to correct for a previous interpretation error. Two similarly colored objects that are next to one another might initially be mistaken for a single object, but when one object is moved, this process would account for the unexpected split in visual regions by creating a new object token.

Spatial-merge type 2 Merges object schema tokens in situation memory that were instantiated in response to two objects that have since become connected. This process should be executed, for example, if the robot observes one object being attached to another.

Spatial-split type 2 Splits an object schema token into two to account for an actual split in what was originally a single object. For example, if an object is cut into two, the token corresponding to the original object is split into two tokens. This process is needed to ground conceptual actions (and ground corresponding verbs) for actions such as cutting or tearing.

Although the processes have been described with an emphasis on visual perception, touch plays a parallel role for all processes (e.g., creating an object schema instance upon touching it).

Schema types guide and limit bodily interpretation process. Since object schemas are collections of possible actions, the situation memory encodes possible actions that the robot expects would succeed if they were to be executed. Which of these possible actions is actually taken depends on the active drive(s) of the agent. When an action is performed on the basis of situated beliefs, failure of the action may lead to belief revision (e.g., removal of a object schema if the robot fails to touch it, update of properties such as weight or slipperiness if the robot fails to lift it).

Schemas in situation memory have intentionality (aboutness) with respect to their referents through causal histories and future-directed expectations. Physical objects cause patterns to impinge on sensors of the agent which are interpreted by the agent by creating object schema tokens (causal relationship). Instantiated schemas are the agent's beliefs about the here-and-now and guide future actions upon the environment (predictive relationship). Interpretation errors may be discovered when predicted outcomes of bodily interactions diverge from actual experiences, triggering belief revision.

With only the curiosity drive in place, Ripley's motor behavior is straightforward: it scans visual zones in a predictable sequence, always choosing the zone with highest curiosity value. We now consider a second top-level drive that competes with curiosity and enriches Ripley's autonomous behavior.

3.4 Multiple Top-Level Drives and Autonomy of Behavior

For practical reasons, Ripley has been designed to keep its motor temperatures within a target operating range. In early versions of Ripley's controller, the robot would sometimes be accidentally left operating over extended periods, causing motor burnout from overexertion. Since introducing the self-maintenance drive, Ripley has never lost another motor. This drive is implemented in a similar fashion to the curiosity drive. The situation model maintains the current estimated temperature of three motor zones. Each zone is the average temperature of a group of motors. The target situation memory is programmed to hold target temperatures for each zone. The difference engine compares actual and target temperatures and if any zone level of "tiredness" is too large, it generates a plan for the robot to bring its body down to a resting position on the table, and power to all motors is shut off for a cool down period.

The drive for self-maintenance conflicts with the drive for curiosity. If both drives are active at the same time the robot will not effectively satisfy either goal since it cannot rest at the same time that it moves to update its situation model. Ripley's architecture adds a layer of control beyond the components of a thermostat as shown in Figure 9. The drive memory system maintains long term motivations of the robot (e.g., target temperatures of motor zones, target curiosity values for visual zones) along with priority weights for each drive. The target selector compares the contents of situation memory and drive memory to select which drive "takes control" of the body. Once a new drive is selected, the target selector copies the associated drive variables into the target situation memory, and then the difference engine goes into action.

The interaction of self-maintenance and curiosity drives in Ripley often leads to surprisingly rich behavior. From a designer's point of view, it is difficult to predict when the robot will tire of visual scanning and decide to rest. Once it rests sufficiently to cool off its motors, it is also difficult to predict which visual zone it will first update in its situation memory.

The most unpredictable element of Ripley's environment is the human. For the human to affect Ripley's behavior using words, Ripley must be equipped with language interpretation skills. To get there, let us briefly reconsider interpretation of sensory-motor interaction as a kind of semiotic process, or sign interpretation. This view leads directly to a path for language interpretation.

3.5 Bodily Interpretation: Signs and Interpretive Frames

Recall the situation memory processes (Section 3.3) that make sense of an object when it first comes into Ripley's view. The visual presence of the object causes a two-dimensional pattern to impinge on the robot's camera sensor. This pattern, when properly interpreted, causes the creation of an object schema token in situation memory. The same is true of touch patterns. Patterns may signify the presence of an object in the here-and-now. The robot interprets patterns as

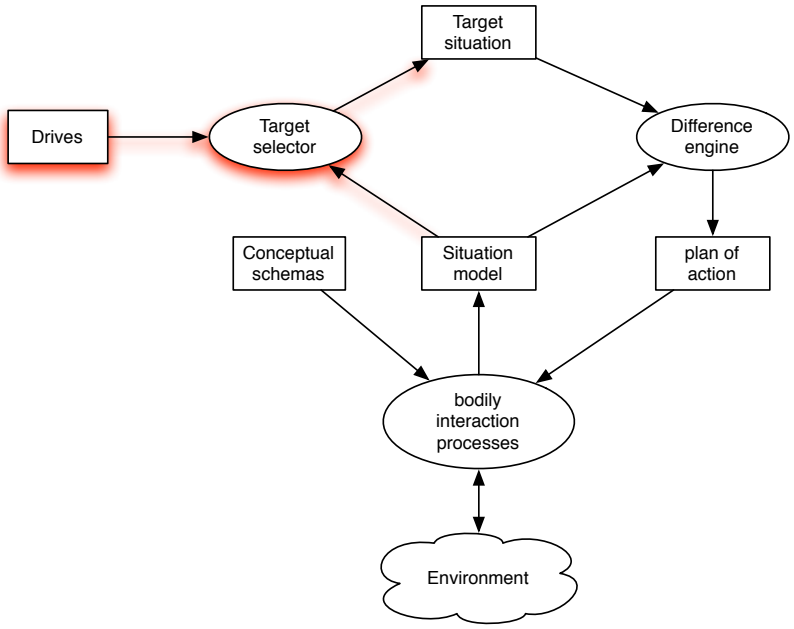

Figure 9: Autonomy of goal selection based on multiple top-level drives.

signs of objects that caused them.

The proper treatment of patterns as signs requires that the interpreter have contextual knowledge about where, when, and how the pattern was received. For example, if Ripley feels no sensation of pressure at its finger tips, how should this touch pattern be interpreted? In a contextual vacuum, this pattern is meaningless. The robot must integrate contextual knowledge of its body pose (which determines the position of its gripper) and whether the gripper is in motion, and which way it is moving. By combining these bits of knowledge, a touch sensation can be interpreted. Suppose Ripley holds the belief that there is an object within its grasp (i.e., an object schema is in situation memory that predicts grasp-ability), and that Ripley reaches to the location of this object and closes its gripper. The sensation at its finger tips now has clear meaning: either a tangible object is in fact present or it is not. I will call the collection of contextual information necessary to interpret a pattern its *frame of interpretation*. Streams of patterns that emerge as a result of bodily interactions are interpreted using the current frame of interpretation. Over time, the meaning of signs are integrated into situation memory.

The foundation is now laid for grounding language use. Words serve as signs, and syntax provides frames of interpretation for words. Sensory-motor signs embedded in their interpretive frames provide the equivalent of descriptive

speech acts – the environment makes "assertions" about how things are. Like linguistic assertions, beliefs derived from the environment may be faulty and thus always susceptible to revision. The meanings of speech acts are created in the process of translating word strings into updates on situation memory (in the case of descriptive speech acts) and drive memory (directive speech acts).

3.6 Grounding the Lexicon: Words as Signs

Open class words are defined by three components:

Spoken Form The sensory-motor schema for generating and recognizing the spoken form of the word. In Ripley these are implemented using standard speech recognition methods based on hidden Markov models combined with model-based speech synthesis.

Conceptual Association Associated schema type in conceptual schema memory.

Proto-syntactic class The word class on which syntactic rules will operate. In contrast to a fully autonomous syntax in which word classes are independent of conceptual categories, the *proto-syntactic* model instantiated by Ripley assumes that word classes are fully determined by their conceptual groundings.

Open class words that Ripley uses include names for objects, actions, object properties, and spatial relations. Examples of schema types for objects and actions were mentioned in Section 3.2. Property schemas specify expected values on measurable parameters of action schemas. Since object schemas embed action schemas, properties can be specified for either actions or objects. Spatial schemas specify expected results of movements between objects.

3.7 Grounding Proto-Syntax: Frames of Interpretation for Words

Just as signs can only be interpreted within the context of a frame, words too require an interpretive frame which is provided by syntactic structure. In Ripley, a small set of context free grammar rules are used to encode allowable word sequences. CFG rules are augmented with interpretation rules that specify the semantic roles and argument structure assigned to words based on syntactic structure[6]. Parsing according to these augmented CFG rules generates a semantic analysis of an utterance in terms of its propositional structure and speech act class. Closed class words (e.g., "there", "is") used in the CFG rules have meaning purely in their role in determining speech act classes and propositional

[6]Ripley's parser is adapted from (Gorniak & Roy, 2004).

structure. Of course other cues beyond syntax (e.g., prosody) may also be integrated to classify speech acts. The propositional structure specifies the role relations between the open class words of the utterance. The structure guides interpretive processes that translate word sequences into memory updates. The speech act class determines whether updates are made to situation or drive memory corresponding to descriptive and directive speech acts respectively.

Figure 10 introduces three final components of the model. The lexicon and rules of syntax are stored in dedicated memory systems, and the speech interpreter applies these memory structures to parse incoming spoken utterances and interpret them. As the figure shows, the effects of interpretation are to modify either situation memory in the case of descriptive speech acts, or drive memory in the case of directive speech acts.

3.8 Interpretation of Descriptive Speech Acts

A descriptive speech act is an assertion by the speaker about how the world is. Assuming a trustworthy relationship between speaker and interpreter, the literal meaning of a descriptive is constructed by making appropriate changes to the situation memory in order to translate the speaker's assertion into beliefs about the here-and-now environment.

When the syntactic parser determines that the utterance is a descriptive speech act, interpretation consists of translating the propositional structure of the utterance into execution of appropriate schema management processes applied to situation memory. Recall that these processes were defined in Section 3.3 and are also used for interpreting bodily interactions with the environment. The open class lexical items in the spoken utterance serve as indices into the schema types.

Let us work through an example to clarify the process of descriptive speech act interpretation. Suppose Ripley's human partner says, "there is a bean bag on your left" while Ripley is looking up at the person, and there are currently no object schemas instantiated in Ripley's situation memory. Interpretation of this descriptive speech act results in the creation of an object schema in situation memory. Here are the stages that lead to this result:

Speech recognition Speech is detected and translated into a sequence of lexical entries. The acoustic form of words are retrieved from the lexicon and constrain the recognition process.

Parsing The output of the speech recognizer is parsed using the augmented CFG rules. The parse will result in classification of this utterance as a descriptive speech act, which routes the resulting memory update to the situation memory. The parse also generates a propositional structure with the form *there-exists(bean-bag(location=left(landmark=ego), color=red))*.

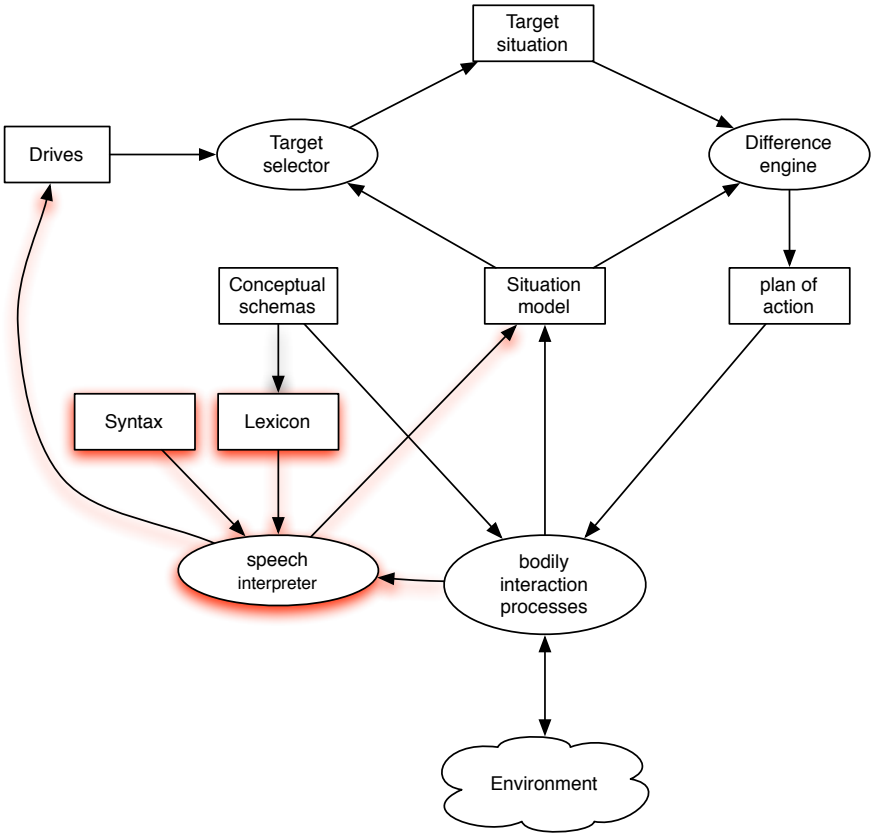

Figure 10: Interpretation of descriptive speech acts about the here-and-now environment parallel interpretation of bodily interactions with environment. The semantics of descriptives are fully constituted by their effects on the contents of the situation model.

Create schema token The top level of the propositional structure is interpreted by creating a new bean bag object schema in situation memory. The lexical item with form "bean bag" provides a link into the conceptual memory store for the corresponding object schema type. By default, the parameters of the schema (location, color, size, etc.) are set to be unknown. Parameters are encoded using probability density functions (pdfs). To specify a parameter as unknown, the corresponding pdf is set to a uniform distribution over all possible values (maximum entropy).

Update schema location parameter The location parameter of the object schema token is updated by replacing its location pdf with a new distribution that concentrates the probability density to the left region of the table top. The location distribution corresponding to "left" is stored in the concept memory as a property schema type. The meaning of "I" and "your" in the current implementation of Ripley is limited to the impact of these words in interpreting spatial semantics (Roy et al., 2004).

Update schema color parameter The color parameter of the object schema token is updated by replacing its color pdf with a new distribution that concentrates the probability density in the region corresponding to the English color term "red". This idealized color distribution is stored in the concept memory as a property schema type.

This example demonstrates the semiotic underpinnings of the model. To process language that depicts an object triggers the same processes that would be used to interpret bodily interactions with an object. As a result, the semiotic requirement of a consistent cross-modal format is met. Words are translated into the same format as sensory patterns since in both interpretive processes, the results are encoded in a common form in situation memory. This enables Ripley to come to believe in the existence of an object via language, and later verify, modify, or even discredit this belief via embodied sensory-motor interaction.

There are differences in the level of ambiguity provided by sensory patterns versus words which are handled naturally in the model through the schema parameter update process. In some cases language is more ambiguous, in other cases less. Suppose Ripley is told that "there is a cup on the table". This assertion provides no constraints on the location of the cup beyond the fact that it is resting somewhere on the surface of the table. It also says nothing about the cup's size, color, orientation and so forth. In contrast, one clear view of the cup would provide far more precise information about all of these aspects of the object. Although the interpreter might not choose to actually extract and encode all information available in the visual pattern, the information is visually available as opposed to the less informative verbal evidence. This is a case where language is more ambiguous than perception. In contrast, suppose Ripley knows about (has conceptual schemas of) three-dimensional balls and cylinders[7]. If

[7]Ripley does not encode shapes as part of its concept schemas, but other robots we have

Ripley were to view an object from above and see a circular form, this pattern would be ambiguous since it might signify either a ball or a cylinder (which would also appear round when upright). On the other hand, the utterance "there is a ball on the table" would not harbour the same ambiguity. More generally, of course, language is a powerful medium for conveying information that is not perceptible at all (i.e., historical facts, theoretical assertions, etc.).

3.9 Interpretation of Directive Speech Acts

A directive speech act transmits a request for action (or in the case of a negative directive, a constraint on actions). The directive might specify an action (e.g., "hand me the cup"), a desired change in situation ("clear the table"), a request for information ("is that box heavy?"), and so forth. I will focus here on the directives that specify actions on objects by working through another example, "hand me the blue cup", which triggers the following interpretive processes:

Speech recognition Speech is detected and translated into a sequence of lexical entries using the same processes as descriptives.

Parsing The parser will classify the utterance as a directive, which routes the resulting memory update to the drive memory. The parser also generates a propositional structure with the form *there-exists(cup(color=blue), target-location=partner)*.

Create target structure The propositional structure is converted into a situation target linked to a top-level *do-as-told* drive. The situation target specifies a desired action (lift object and offer to human partner) and target object which specified by selecting an active object schema in situation memory which best matches the contents of the propositional structure. If no object schema in situation memory matches the requested object, this stage will fail.

Once the requested action is placed into drive memory, it will remain there until the target selector determines that the do-as-told drive should gain control. When this happens, the verbally triggered target situation is copied into target memory, which in turn triggers planning processes that carry out the requested motor action.

3.10 Summary

The model begins with the classic cybernetic structure of a closed loop controller. This skeletal structure provides scaffolding for introducing richer memory structures and processes for control and interpretation. A pre-linguistic architecture is developed that integrates drive arbitration with conceptual schemas

built do (Roy, 2003).

that guide the interpretation of sensory-motor interactions with the environment. Interpretation of bodily interactions can be viewed as a semiotic process of sign interpretation, where the results of interpretation lead to modifications of situation memory. We can then graft machinery for linguistic interpretation onto this underlying architecture. Interpretation of descriptive speech acts parallels the process of making sense of sensory-motor patterns by effecting situation memory. The meaning of directive speech acts emerges from the modification of drive memory structures.

This model addresses only a sliver of the linguistic phenomena exhibited by a young child. The list of what is missing is long and includes language production, understanding any kind of non-literal meanings, anaphoric resolution, reference beyond the here-and-now, and so forth. At a prelinguistic level, the conceptual representations I have sketched are also very incomplete. Object schemas as presented here only encode "do-to" affordances, i.e., the kinds of actions the agent can expect to take on an object such as grasping it or pushing it. Eventually we must also model "do-with" affordances (i.e., one can carry things in a cup) and "reacts-to" affordances (i.e., what a ball does when pushed). Type hierarchies for actions, objects, properties, and relations are also surely important, as is goal and belief attribution to communication partners. I believe all of these issues may eventually be addressed by building on the approach developed here, an approach that might be called embodied interactionist semantics[8].

This model makes some progress in explaining language use in holistic yet mechanistic terms. In contrast to typical computational models of language that focus on one aspect (e.g., syntax, semantic networks, word-to-percept associations, etc.), this model strives to encompass a broad range of cognitive structures and processes in a coherent framework to yield explanatory power.

4 Meaning, From Ripley's Point of View

In Section 1.2, I suggested three facets of meaning that are in need of explanation by a unified mechanistic model. In the model presented here, meaning arises as a consequence of the effects of words on an agent's memory systems, and the ensuing effects on the dynamics of the agent's control processes that are affected by these memory systems. Let us consider how the model relates to each facet of meaning.

4.1 Referential Meaning

The propositional content of descriptives and directives are bound to the environment via situation memory. Schema instances in situation memory are

[8]In philosophical terms, the approach is closely connected to Pierce's pragmatism (Peirce, 1878) and in contemporary philosophy to Millikan's functional semantics (Millikan, 2004) and Bickhard's interactivism (Bickhard, 1993)

cashed out in how they make predictions of, and are shaped by referents in the environment. Individual words provide constraints on the establishment of reference. For example, the meaning of "blue" constrains parameters of visual schemas embedded in whatever object is described to be blue. The model suggests that reference is the result of dynamic interactive processes that cause schemas to come into being and that keep them attuned to specific objects.

4.2 Functional Meaning

Functional semantics of speech acts emerge from their embeddings in the overall behavior system of the agent. The functional class of a speech act is determined by word order and the location of specific closed-class words. The meaning of the classification is determined by the memory system that is effected. Descriptive speech acts are translated into updates on situation memory, effectively committing the agent to new beliefs that have potential "cash value" in terms of future bodily interactions with the environment. Directive speech acts are translated into updates on drive memory, effectively biasing the agent to act in new ways in order to satisfy the new goals expressed by the speech acts.

4.3 Connotative Meaning

Connotation refers to the implied, subjective meaning of a word or phrase as opposed to its conventional, literal meaning. In the model, connotative meaning of a word is the accumulated effect of the word's conceptual structure with respect to the motivational drives of the language user. For example, consider the connotation of "heavy" for Ripley, given that the robot is driven by curiosity, the need to stay cool, and to do as its told. In terms of doing what it is told to do, the meaning of "heavy," like any property name, is useful since it lets Ripley pick out the object that the human wants. In this respect, the connotation of heavy is positive - we might say, *positive with respect to the verbal response drive.* In contrast, any object that is heavy will be negative with respect to Ripley's other top-level drives, since manipulating heavy objects accelerates motor heating, and heavy objects often slip. Although Ripley's curiosity drive is currently based on visual evidence, it would be possible to extend curiosity to include the weight of objects in which case Ripley would be compelled to lift all novel objects to characterize their weights. In such a setting, the negative relation of heaviness to motor heat would also emerge during the activity of satisfying curiosity, which would now involve lifting all novel objects. Overall, we might say that the meaning of "heavy" is neither totally negative nor positive, but rather is bittersweet. This mixed attitude is due to the interaction of Ripley's multidimensional drives, the particular conceptual structure underlying "heavy", and the physical laws of force and heat that govern Ripley's bodily interactions with its environment. The model yields an analysis of connotative meanings as the agent's embodied experiences "projected" through the lens of the particular

concept. Even with Ripley's simple drive system we obtain a three-dimensional space for connotative meanings. More generally, the richness of connotative meanings grows with the richness of the drive system of the language user. As a child grows and acquires culturally rooted interests and drives, the effective dimensionality and subtleties of connotative meaning become correspondingly more intricate.

4.4 Conclusion

To summarize, I believe the model presented here provides a new way of looking at how each facet of meaning emerges from common causes. Rather than propose three models of meaning, a single model gives rise to all three. The partial implementation of the model in the form of an embodied, interactive, autonomous robot provides a tangible example of a machine that "grasps" linguistic meanings in a way that goes beyond any symbol manipulation model of meaning.

Discussion

Reviewers of an earlier draft of this paper raised numerous questions and concerns, some of which I have tried to address in revisions to the paper. However, I felt that a subset of the questions were better answered in a question and answer format as follows in this section.

"Create" is a process of object schema token instantiation. If I understood, this means that the corresponding type schema must be stored in memory. My question: Is it the case that only objects whose type schema was precompiled by the programmers can be perceived, manipulated, and understood? In other words, if you provide Ripley with a banana and it does not have any previous banana type schema, could Ripley respond to descriptions and commands about it? My question is also related to Ripley's learning capabilities (or limitations).

Only pre-compiled schemas may be instantiated in the current model. Schemas have parameters such as shape and color which can accommodate new forms of objects. But the topological structure of schemas as defined by the network of possible actions is assumed to be fixed and preprogrammed. An important topic for future work is to explore how a machine could learn new schemas (i.e., concept learning), perhaps through a process of discovering new clusters of interaction patterns.

"Destroy" seems psychologically a radical procedure, because object permanence takes place even if an object is hidden. The robot should distinguish between a hidden object that could reappear later on and a really destroyed object that will never come back. Of course you know that, but then what is the motivation for including destroy? Maybe Destroy is convenient to clean up the working memory

in a purely on-line processor such as Ripley, but what about long term memory? For instance, lets suppose that you want to implement an improved version of Ripley with episodic long term memory. We should ask Ripley displaced reference questions like: where was object X before (now)? To answer this, the destroy routine would be quite inconvenient, because the system should be able to track of past events and organize them into episodic memory traces.

In psychological terms, Ripley has working memory (the situation model, plan of action system, target situation memory, drive encodings) and long term semantic memory (for storing conceptual schemas) but no long term episodic memory. A promising future direction is to extend the model to include episodic memory by selectively encoding chains of activity of the various working memory systems. Furthermore, we might explore learning processes that compress multiple similar episodic traces to form new conceptual schemas.

Section 3.1 begins with the statement "to account for any of the three facets of meaning, the language interpreter must have its own autonomous interest/goals/purposes." This is a really interesting claim, but it is given little justification in section 3.1, and it is not until section 4.3 that it is cashed out. I would really like to see Roy justify the claim more completely in 3.1, or even in 4.3.

Schemas are structured bundles of possible actions. In other words, schemas are collections of *potential plan fragments*. As such, all instantiations of schemas (beliefs) and their meanings are inextricably imbued with the interests / goals / purposes of the beholder.

The question of whether a system is embodied is another layer beyond the matter of a model being grounded. Systems can be grounded in perception and action, but not be embodied, whereas it would not make sense for an embodied model to be ungrounded. In my view, what makes a model embodied is that its representations and processes are constrained by the properties of the body. So I wonder where Ripley stands on the question of whether it is embodied.

The terms "grounded" and "embodied" have unfortunately been used in many overlapping ways by authors and are thus bound to be confusing, and any specific definition contentious. That being said, I find it difficult to make sense of the assertion that "Systems can be grounded in perception and action, but not be embodied" if we are talking about perception of, and action in the *physical world* since that would be impossible without a physical body, and I think most definitions of embodiment involve bodies in one way or another. Ripley is a physically instantiated machine that senses and acts upon its physical environment so I would call it embodied, but I don't think there is much significance to being classified as such. The point of the model I have presented is not merely that it was implemented in an embodied system (robots have been around for

a long time). Rather, the important idea here is to develop a continuous path from sensory-motor interations to symbolic communication.

In Ripley, schemas are organized around objects (e.g., the lift schema is embedded within the cup schema). These are essentially affordances. I worry that such a position is too bottom up. A more goal-driven system would presumably be needed also.

I agree that schemas are closely related to the idea of affordances. Affordances as originally concieved by Gibson (1979) refer to the confluence of top-down and bottom-up factors, so there is no reason to "worry that such a position is too bottom up" anymore than one should worry that such a position is too top-down. Ripley's architecture is designed to blend top-down and bottom-up influences. On one hand the drive memory, target selection, and planning mechanisms drive top-down activity whereas interpretation of bodily interactions with the environment via belief revision adapt to bottom-up influences of the environment.

Ripley's continuous scanning of the environment may be a bit different than humans. Human attention is primarily drawn by changes in the environment.

Ripley's architecture is trivial in richness and capability when compared to humans and thus there are vast differences between its scanning processes and analogous functions in humans. That said, it would be inaccurate to say that human attention is "primarily drawn by changes in the environment". Attention depends critially on active goals and interests, not just changes in the environment (as shown, for example, by experiments in change blindness).

Ripley lives in a simplified toy world, much like Winograd's Blocksworld in the early 1970s. There is a small finite number of goals, a limited number of agents and objects, a small spatial layout, assumed common ground in person-system interaction, and so on. One reason that it is necessary to work with toy worlds is because of combinatorial explosion problems and modules that otherwise would be impossible to implement (such as goal recognition in the wild, and assumed-rather-than-computed common ground). It would be good to justify Ripley's being in the toy world for these and any other reasons.

As anyone who has tried to build an autonomous robot knows, operating on everyday objects in the physical world – even in the safe haven of Ripley's tabletop world – is *radically* different from operating on symbolically described blocks in a virtual world of the kind that was popular in early AI systems. This question appears to be based on a failure to grasp this difference. Since error and ambiguity do not arise in symbolic descriptions, there is no reason to make the fundamental distinction between belief and knowledge when dealing with symbolic microworlds. In contrast, all that Ripley can ever strive for are beliefs that translate into expectations with respect to self-action. The world may of course not satisfy those expectations, leading to belief revision. The model of language

interpretation and meaning I have presented are grounded in this conception of belief. Issues of error, ambiguity, and categorization are at the core of the model. None of these issues even arise in classic AI microworlds such as that of Winograd's SHRDLU (Winograd, 1973). A better way to think of Ripley's simplified environment is that Ripley exists in a carefully controlled *subworld* in the sense defined by Dreyfus and Dreyfus (1988). Like the controlled subworld of an infant, Ripley's world opens up into our world as opposed to microworlds that are physically disconnected from the rest of reality. Justifications for keeping Ripley's subworld simple can be found in Sections 1.1 and 1.3.

Acknowledgments

Ripley was built in collaboration with Kai-yuh Hsiao, Nick Mavridis, and Peter Gorniak. I would like to acknowledge valuable feedback on earlier drafts of this paper from Kai-yuh Hsiao, Stefanie Tellex, Michael Fleischman, Rony Kubat, Art Glenberg, Art Graesser, and Manuel de Vega.

References

Barsalou, L. (1999). Perceptual symbol systems. *Behavioural and Brain Sciences*, *22*, 577-609.

Bates, E. (1979). *The emergence of symbols.* Academic Press.

Bickhard, M. H. (1993). Representational content in humans and machines. *Journal of Experimental and Theoretical Artificial Intelligence*, *5*, 285-333.

Chomsky, N. (2000). *New horizons in the study of language and mind.* Cambridge University Press.

Cruse, D. (1986). *Lexical semantics.* Cambridge University Press.

Drescher, G. (1991). *Made-up minds.* Cambridge, MA: MIT Press.

Dreyfus, H., & Dreyfus, S. (1988). Making a mind versus modeling the brain: Intelligence back at a branchpoint. *Daedalus*, *117*(1), 15-43.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Erlbaum.

Gorniak, P., & Roy, D. (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, *21*, 429-470.

Hsiao, K., & Roy, D. (2005). A habit system for an interactive robot. In *AAAI fall symposium 2005: From reactive to anticipatory cognitive embodied systems.*

Landauer, T. K., Foltz, P., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, *255*, 259-284.

Lenat, D. (1995). Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, *38(11)*, 33-38.

Mavridis, N., & Roy, D. (2006). Grounded situation models for robots: Where words and percepts meet. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).*

McKean, E. (Ed.). (2005). *The new Oxford American dictionary.* Oxford Univeristy Press.

Miller, G. (1995). Wordnet: A lexical database for english. *Communications of the ACM*, *38(11)*, 39-41.

Millikan, R. G. (2004). *Varieties of meaning.* MIT Press.

Narayanan, S. (1999). Moving right along: A computational model of metaphoric reasoning about events. In *Proceedings of the national conference on artificial intelligence AAAI-99.* Orlando, FL.

Odgen, C., & Richards, I. (1923). *The meaning of meaning.* Harcourt.

Peirce, C. S. (1878). How to make our ideas clear. *Popular Science Monthly*, *12*, 286-302.

Quillian, R. (1968). Semantic memory. In M. Minsky (Ed.), *Semantic information processing.* MIT Press.

Rosenblueth, A., Wiener, N., & Bigelow, J. (1943). Behavior, purpose and teleology. *Philosophy of Science*, *10*, 18-24.

Roy, D. (2003). Grounded spoken language acquisition: Experiments in word learning. *IEEE Transactions on Multimedia*, *5*(2), 197- 209.

Roy, D. (2005). Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, *167*(1-2), 170-205.

Roy, D., Hsiao, K., & Mavridis, N. (2004). Mental imagery for a conversational robot. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, *34*(3), 1374-1383.

Siskind, J. (2001). Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal of Artificial Intelligence Research*, *15*, 31-90.

Winograd, T. (1973). A process model of language understanding. In *Computer models of thought and language* (p. 152-186). Freeman.

Wittgenstein, L. (1958). *The blue and brown books.* Oxford: Basil Blackwell.