Incremental Natural Language Processing for HRI

Timothy Brick
A.I. & Robotics Lab
University of Notre Dame
Notre Dame, IN 46556, USA
tbrick@cse.nd.edu

Matthias Scheutz
A.I. & Robotics Lab
University of Notre Dame
Notre Dame, IN 46556, USA
mscheutz@cse.nd.edu

ABSTRACT

Robots that interact with humans face-to-face using natural language need to be responsive to the way humans use language in those situations. We propose a psychologically-inspired natural language processing system for robots which performs incremental semantic interpretation of spoken utterances, integrating tightly with the robot's perceptual and motor systems.

Categories and Subject Descriptors: I.2.7 [Computing Methodologies]: Artificial Intelligence– *Natural Language Processing/Robotics*

General Terms: Human Factors, Design

Keywords: Embodied NLP, incremental processing, HRI

1. INTRODUCTION

There are at least three characteristic features of human language processing in face-to-face communication: (1) humans process language incrementally when possible, (2) humans automatically make use of perceivable context in the production and resolution of referential expressions, and (3) humans react both verbally and non-verbally to language as they process it. For example, listeners will rapidly and incrementally interpret spoken utterances to establish reference as soon as a speaker's utterance provides sufficient information to distinguish the intended referent from perceivable alternatives, even when this information occurs before the end of the syntactic constituent [9, 2, 25, 29]. Listeners will also react to the utterance using linguistic and non-linguistic methods as including "back-channel" responses such as eye movements, head nods, gestures, vocalizations, and even interruptions to signal their intentions and their understanding of the linguistic discourse.

We believe that all three of these characteristic features of human language processing are critical to HRI. Humans will converse differently based on their expectations of the abilities of their co-conversant, and robots will need to be sensitive to human interaction styles if they are to communicate with people in natural ways. Given the human tendency to anthropomorphize artifacts, we surmise that people will base their expectations about a robot's capabilities (perceptual, linguistic, etc.) on their observations of its appearance and behavior.

A robot with two camera"eyes" on a movable "head" will, for example, be expected to see objects in the world and direct its gaze to them, and will be spoken to as though it were able to do so. Unless robots are capable of using perceivable context in processing language and producing back-channel feedback to a human speaker at appropriate times, these expectations will be violated, resulting in communication failures and unnatural-seeming interactions.

In order to be able to keep up with human expectations (about linguistic abilities, timing, etc.), a robot *as an embodied agent* must, at the very least, perform its language processing incrementally (to be able to respond while an utterance is in process) and integrate perceivable context (to appropriately determine reference). In particular, semantic interpretation must occur in parallel with other processing stages and linguistic processing must be intrinsically intertwined with both perceptions (e.g., to determine referents) and actions (e.g., to provide feedback).

In this paper, we present our first attempt at defining a robotic incremental semantic engine (called RISE), which we hope will eventually be capable of interacting face-to-face[1] with humans in natural ways. RISE incrementally processes syntactic and semantic information and integrates of perceptual and linguistic information, and can therefore generate feedback or other actions during the processing of utterances. In the rest of the paper, we first expand on the need for embodied incremental language processing in HRI, summarize the related work, and argue for the insufficiency of incremental NLP systems that do not intrinsically take an agent's body (with sensors and effectors) into account. We then describe the architecture of RISE and illustrate its operation in three demonstrations that point to the utility of incremental embodied NLP, concluding with a brief discussion and outlook for future developments.

2. BACKGROUND AND RELATED WORK

Humans in natural face-to-face conversations have stylistic conventions and pragmatic constraints on their linguistic interaction. For example, referential *overspecification*, the

[1]Throughout this paper, we will use the term "face-to-face" interactions to refer to communicative interactions between two embodied agents in close enough proximity to share a common perceptual context.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
HRI'07, March 8-11, 2007, Arlington, Virginia, USA.
Copyright 2007 ACM 978-1-59593-617-2/07/0003 ...$5.00.

use of more properties to describe an object in shared visual space than is required to single it out (such as referring to "The apple on the towel" when only one apple is visible), can confuse a listener [4] in face-to-face conversation, even though it is grammatically and semantically valid. In interaction situations other than face-to-face situations, particularly in computer-mediated communication such as instant messaging and even videoconferencing, humans show significant deviations from their natural interaction styles (e.g., many culture- and timing-related requirements are relaxed in these environments [15, 16, 26]). Consequently, it is likely that the expectations about natural language processing for desktop computers (akin to a videoconferencing environment) will be different than those for robots (akin to face-to-face interactions). While it is obvious from their appearance and behaviors that robots are capable of perceiving and actively exploring their environment, desktop computers are bound to a fixed location in their environment and show little or no evidence of being able to perceive the outside world. From a human perspective, computers are clearly not embodied agents in the sense robots are (or can be), i.e. autonomous entities that appear to perceive and interact with their environments. To accomodate, at least to some extent, human expectations, incremental natural language processing has been explored extensively in the past, at different levels of incrementality.

Incremental processing at the syntactic level, for example, has been explored by Mori et al. [14], Rosé et al., [18] and others [7], using incremental chart parsing (or a similar incremental parser) to perform fast and robust syntactic parses. These systems then pass their completed syntactic parses on to a separate module for semantic understanding. While such systems can be quite robust to dysfluencies in natural language, they require multiple stages for understanding an utterance. As such, the semantic interpretation of an utterance cannot begin until the utterance itself is syntactically parsed in full (which must be at least the end of the utterance itself). Therefore, it is not possible for these systems to provide the feedback about its own state of understanding while an utterance is still being spoken, or respond to an utterance before it is complete.

Other systems such as Terry Winograd's SHRDLU [32], some of the Results of the ARPA Speech Understanding Research Program, such as HEARSAY [5] and HARPY [13], and several more recent systems [3, 1, 30] are incremental at both syntactic and semantic levels. Purver and Otsuke's system [17] additionally generates referential phrases using a reversible grammar. While these systems would be able to perform some of the "back-channel" responses necessary for natural interactions, they are desktop systems, not embodied systems, and therefore more similar to videoconferencing or instant messaging communication than to face-to-face. As such, they have neither the need nor the necessary effectors to make back-channel responses. They are also not sensitive to the pragmatics of human communication, such as the use of a shared visual environment.

The Stanford Computational Semantics Laboratory's work [11, 31] is also quite robust to partial, incomplete and interrupted sentences, maintains shared context, and allows for back-channel and asynchronous responses. It is not sensitive, however, to the pragmatic constraints imposed on embodied systems, and relies on explicit gesture and past discourse for resolution of reference, not on shared context.

Some recent true robotic architectures can resolve references to objects in visual scenes (e.g., [19, 20]). These systems are focussed primarily on specific aspects of the construction of shared discourse, such as for example, how word meanings emerge and can be used to refer to objects in a shared scene, e.g., [28], or how reference can be resolved with minimal computational effort in a behavior-based robotic system, e.g., [8]). They are not intended to address the pragmatic concerns faced by embodied systems.

3. DESIGN AND ARCHITECTURE

RISE is our first attempt to step beyond incremental language processing on desktop system and include perceptions and actions during linguistic processing, modeling the pragmatic constraints and interactional conventions used by humans in face-to-face discourse. It is intrinsically parallel and incremental both at the syntactic and semantic levels and works closely with the robot's perception and action systems. It can thus be sensitive to issues like referential overspecification (mentioned above) as well as provide important back-channel feedback and interruptions during an utterance, which can trigger early clarifications.

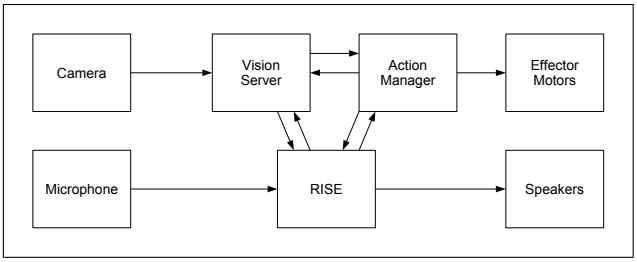

Figure 1: An overview of the architecture used for RISE.

3.1 Overall Robotic Architecture

The current implementation is part of our distributed integrated affect, reflection, and cognition (*DIARC*) architecture [23] (a high-level schematic showing the relevant relations between RISE and other parts is shown in Figure 1). The architecture runs on an ActivMedia Peoplebot (P2DXE) with two Unibrain fire-wire cameras mounted on Directed Perception pan-tilt unit, a Sick laser range finder, a Voice Tracker Array Microphone, and two speakers.

In the context of the *DIARC* architecture, RISE is used for various kinds of natural language interactions that are part of team tasks [24], from understanding and initiating simple commands like "turn right" or "move forward", to more complex natural language dialogues and action sequences involve multiple effectors, attentional subsystems, and other parts of the robotic architecture.

A typical example is the a dialogue-like interaction with humans about objects in the environment and their relation: <Human>:"Look fifteen degrees to the left, do you see a screen?" – <Robot turns cameras to the left>:"Yes."; <Human>: "From the perspective of the screen, is the box to the left of the folder?" – <Robot>:"Yes.". These kinds of interactions were recently demonstrated at the 2006 AAAI HRI robot competition with a simpler version of our incremental system and received an award for natural language processing and action execution.

Our action manager is script-based, taking script names and arguments as inputs, and running scripts step by step. The interpreter is capable of handling empty arguments by supplying defaults. For example, if told to execute the "startMove:" action script without a direction or extent, the interpreter will begin to move forward until another instruction (or some other concern, such as an intervening wall) causes it to stop. The action manager is also capable of requesting bindings of specific types from the discourse engine. For example, if the action manager knew it was required to move, but could not assume a direction for some reason, it could request a direction for movement from RISE.

Our vision system uses its two cameras and pan-tilt unit to perform Scale-Invariant Feature Transform (SIFT) [12] point identification of objects in the visual field. It is capable of estimating the location of each element in space (using binocular vision to determine depth), and can calculate spatial relationships between these objects accordingly.

A short-term visual memory has been added to the visual system to maintain rough descriptions and locations of objects no longer currently in the robot's field of view. These representations can be verified by visually examining each object as necessary to confirm its continued existence and location.

Currently, phonetic tokenization and resolution is performed by the Sphinx-4 [27] language recognition system. Speech synthesis, when necessary, is performed by the Festival speech production system. These phases of language recognition are performed in parallel with the operation of RISE so that speech can be processed incrementally at the lexical level.

Each of the unique systems (RISE, the vision server, the action manager, etc) is a small Java server running as an ADE module. Modules communicate through a variant of the Java Remote Method Invocation (RMI). Modules call each other through synchronous RMI calls, but implement a callback system to allow simultaneous operation for time-consuming commands (i.e. SIFT calculation) (for more information about the ADE architecture, see [21]).

RISE itself is designed to be psychologically plausible and to resemble the process used by humans in natural language processing system. To this end, each recognized lexical item in an utterance (i.e., words) is simultaneously syntactically analyzed and semantically interpreted as soon as it becomes available to the system. Similarly, utterances are incrementally produced from intentions in a way that lends itself to the conventions of incremental recognition and interpretation used by humans in face to face interactions. This is important because it will allow more natural interaction in a face to face environment, avoiding problems like the over-specification of referents, which can cause confusion[4].

3.2 Integration with Perceptual Systems

RISE's online processing of semantic constraints also allows for better disambiguation of otherwise ambiguous statements, especially if perceptual context can be utilized. Humans with access to a shared context (e.g. a visual environment) are capable of easily interpreting otherwise ambiguous statements.

The referent resolution algorithm in our system follows the core assumptions about incremental reference resolution that are part of the computational robotic model of human incremental reference resolution proposed in [22].

Essentially, these state that the reference relations between terms and objects of reference are established incrementally. At any given time during reference establishment, the expansion of a complex term (e.g., a noun phrase) is a set PR of possible referents to which the term could refer. Reference is established when the set PR has precisely one member ($|PR| = 1$). After this point, a listener will continue to attach incoming words to that term only due to syntactic constraints (for example, in "the white block", "block" will be attached to "white" even if reference is established because syntax forbids "block" from being separated from "white").

RISE, therefore, deals with a referential phrase by creating a set PR of possible referents (in order to save memory and processing time, the set PR is not constructed until at least one significant constraint term exists in the utterance). In processing each word as it arrives, RISE reduces the size of PR based on word's associated syntactic constraints and the semantic constraints established after processing its meaning. Incoming terms are considered constraints on the referent until the set PR contains only one member and syntax constraints allow it to return processing "attention" to the enclosing clause.

Consider the two scenarios in Figure 2.

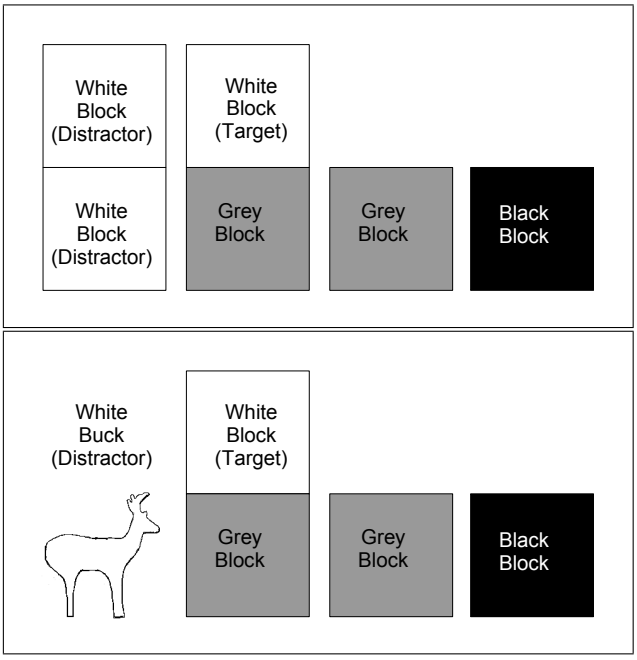

Figure 2: Two different scenarios in a blocks world setup (see text for details).

The phrase "Put the white block on the grey block on the black block" means something different in each scenario. In the upper scenario, "the white block" does not uniquely determine a single block, and so "on the grey block" is taken to be a clarification of the referent. "on the black block", then, is regarded as the destination location of the "put" action. In the lower scenario, it is clear which white block is meant, so "on the grey block" is taken to specify the destination of "put". When "on the black block" registers, it is difficult to understand the meaning of the sentence, since both blocks have already been uniquely specified.

It is this kind of referential overdetermination that reliably causes comprehension difficulties in face-to-face interactions among humans (though it may not in other situations). Most systems will pay no attention to the critical differences in how human construct referential expressions for these scenarios, because their syntactic parsers will not be guided by the incremental interpretation of meaning inherent in these referential phrases. That is, since both parses are possible and seem equally likely (or, at least, the parser will have little to no information about their likelihood), both will be considered. In fact, they need not be. In the upper scenario, the first "on" will never indicate the target location, and in the second, the sentence will not be produced by humans in face to face interaction (or if it is, confusion based on it will be expected and quickly handled). While a desktop system would not need to worry about these rules, an embodied system that will be engaged in face to face interactions will need to be. RISE will resolve these referents incrementally, not even considering the incorrect interpretations unless it becomes clear that it has misinterpreted the referential phrase.

3.3 Integration with Action Execution

In order to allow the necessary level of incremental processing, RISE utilizes combined syntactic/semantic knowledge structures to determine both syntactic properties and semantic interpretations of lexical items at once[2]. These structures are similar to synsets in WordNet ([6]) and verb senses in VerbNet([10]), but maintain both semantic and syntactic values, and are tailored to the semantic interpretations used by the action manager.

For the duration of an utterance, RISE maintains multiple possible interpretations (or *hypotheses*) of the parts of the utterance it has heard so far. As each new word is processed, its semantic and syntactic characteristics are retrieved from a table and added to the current interpretations. The semantic and syntactic constraints of the new word may remove or change some of the existing interpretations of the utterance. Similarly, only those senses of the new word which are appropriate to a given interpretation are applied to it. Each new word added to the discourse is thus able to propagate constraints both forward and backward in the discourse stream (by changing or removing the existing interpretations) and is itself constrained by the existing interpretations.

When an utterance is completed and only a single interpretation remains, the sentence is already understood, and no further processing is needed. In the event that more than one interpretation remains, a clarifying question can be generated, or the "most likely" interpretation can be used[3].

The overall algorithm is similar to an incremental chart parsing system (e.g., [14]). The primary difference is the integration of semantic understanding into the structures. For example, when given the word "move" at the beginning of a command, it is unclear whether the robot is being ordered to move itself or to move an object. In terms of the robot's actions, these are two totally distinct tasks. We therefore represent these tasks as two separate constraint systems, because the semantic and syntactic constraints placed on later words are entirely dissimilar. Because we additionally allow unstructured semantic and syntactic constraints to be propagated both forward and backward, our system more closely resembles a constraint-propagation view of language processing than a chart-parsing view. While our method results in a larger series of possible interpretations than a classic chart-parse, it also more closely reflects the distinct semantic meanings of words.

The combined semantic and syntactic engine does not fall prey to combinatorial explosion as might be predicted, because it does not blindly add all possible interpretations to the list. Instead, it utilizes both syntactic constraints and simple semantic constraints (such as those posed by a shared visual environment) while the utterance is spoken to maintain a minimum number of possible interpretations at any given time.

One major benefit of our propagation algorithm lie in the ability of the system to process an utterance while that utterance is still being spoken. This results in higher reactivity of the system, and the ability to act faster and in ways more similar to those expected by humans, including "back-channel" feedback and interruptions.

The meaningful partial interpretations of the utterance can be used to initiate actions while the utterance continues. Consider the sentence, "put the white block on the black block" for the lower scenario in Figure 2. By the end of the word "white", RISE can initiate an action to confirm the existence of both white objects by looking at them. This action both overcomes the limitations of the robot's vision (if it could not see both objects, and therefore had to confirm they were still there), and lets the speaker know that the robot has understood so far. By the end of "put the white block", RISE will already understand the first part of its task. That is, it will understand that it is initiating a "put" action, and that the white block is the object to be put. The robot could therefore begin to reach for the white block and pick it up, knowing that it will have to be moved to a new location. This is similar to behavior exhibited by humans immediately after hearing a referential phrase (and sometimes even during the phrase)[4]. While not a formal vocal response to the human's utterance, such movements can indicate clearly to the human that the robot has (whether correctly or incorrectly) determined the white block to be the object to be moved.

By the end of the word "black", RISE will already have uniquely specified the location to place the white block, and can begin to move the block accordingly. This is especially useful in cases where the end of a sentence might be garbled, lost, or simply omitted. Furthermore, the anticipation of what should come next might allow the system to generate timely or even overlapping responses to the speakers' utterance, as demonstrated in Section 4.1.

4. THREE DEMONSTRATIONS

We will now demonstrate the utility of RISE's incremental semantic engine and integration with the visual system.

[2]We will refer to semantic interpretations of an utterance in terms of the underlying robotic representation. In the case of a command, these interpretations take the form of arguments which can be passed directly to the action manager for execution.

[3]In the absence of actual statistical information about word sense frequencies, likely interpretations are currently determined by retrieval order based on hypothesized frequency and likely occurrence in the given context. Each interaction type can have its own vocabulary and its own weighting of word meanings.

4.1 Demonstration of Anticipation and Visual Integration

We begin with a demonstration of incremental natural language understanding and action execution to show how our system can generate actions based on partial interpretations of incomplete utterances and start action execution while the rest of the utterance is still being processed.

Because our constraints are processed incrementally (as described above) and continually limit the possible actions that the robot can take, it is sometimes possible to anticipate what a command or utterance means even before the command or utterance is complete. Once sufficient constraints have been added that the script to be executed is known, the script can be executed, resulting in the anticipatory behavior. If this anticipation is violated, it can be detected immediately. The robot can either interrupt to clarify the statement, or (for example, if the end result is nonsensical) ignore the nonsensical parts as errors of speech recognition (or production errors on the part of the speaker).

We begin with the utterance "move the white block onto the black block" and the shared visual environment shown in the lower pane of Figure 2, above. To complicate matters slightly, our human speaker will mispronounce the last word, replacing the last instance of "block" with the word "buck".

This dysfluency will generate a nonsensical statement, but RISE will be able to resolve it reasonably. Similarly, the human will be able to notice if the robot is attending to the wrong objects, because the robot will clearly look at each object involved to verify that it is still where it was left.

The output of the processor[4], below, is annotated with effects of action manager scripts, etc. so that it is easy to see the benefits of incremental processing. Different syntactic/semantic interpretations are listed on separate lines connected by vertical bars ("|"). Parentheses ("()")surround optional but highly likely syntactic constructions, and brackets ("<>") indicate syntactic locations for incoming arguments. Noun phrases are enclosed in square brackets ("[]") preceded by the letters "NP:", and are surrounded by asterisks ("*") when "attention" focuses on them. Below each syntactic interpretation is the associated semantic meaning. Question marks ("?") precede the names variables which may yet be matched. Arrows ("−−>") show calls to outside systems or denote actions of the robot. An ellipsis in brackets ("[...]") indicates a point where a large portion of output (say, the processing of one or more words) was removed for space reasons.

```
... Beginning new utterance
Processing word: move
Matched Verb: move
Current utterance:
|Move <prep> <location>
| (startMove:?location)
|
|Move (<prep> <article:def>) <direction>(<extent>)
| (startMove<direction>:?extent)
|
|Move <object> <prep> <location>
| (moveObject:?prep:?location)
   Possible Actions:  startMove, moveObject

... Processed: move
Processing word: the
Matched article: the added to NP:[the <?>]
Current utterance:
|Move *NP:[the (<adjective>)<name:object>]*
                                        <prep><location>
   Confirmed Action: moveObject:?object:?prep:?location

... Processed: move the
Processing word: white
Matched Adjective: +white added to NP:[the white <?>]
--> Visual System Query: listObjects:+white
--> Response: {block3 buck1}
Possible Referents: {block3 buck1}
--> Visual System Query: confirmObjects:{block3 buck1}
--> Robot looks at block3, then buck1
Current Utterance:
| move *NP:[the white (<adjective>) <name:object>]*
                                        <prep><location>
   Confirmed Action: moveObject:?object:?prep:?location

... Processed: move the white
Processing word: block
Matched noun: name "block" added to NP: [the white block]
--> Visual System Query: listObjects: block +white
--> Response: {block3}
Possible Referents: {block3} Reference established.
--> Visual System Query: confirmObjects:{block3}
--> Robot looks at block3
Current Utterance:
| move *NP:[the white block]* <prep><location>
   Confirmed Action: moveObject:block3:?prep:?location

--> To ActionManager: moveObject:block3:?prep:?location
--> Robot begins to reach for block3.

[...]

... Processed: move the white block onto the
Processing word: black
Matched Adjective: +black added to NP:[the black <?>]
--> Visual System Query: listObjects: +black
--> Response: {block1}
Possible Referents: {block1} Reference established.
--> Visual System Query: confirmObjects:{block1}
--> Robot looks at block1
Current Utterance:
| move NP:[the white block] onto(atop) *NP:[the black
                                  (<adjective>)<name:object>]*
   Confirmed Action: moveObject:block3:atop:block1
--> To ActionManager: moveObject:block3:atop:block1
--> Robot begins to move block3 toward block1

... Processed: move the white block onto the black
Processing word: buck
Matched noun: name "buck" added to NP: [the black buck]
NP: Reference already established
--> Visual System Query: getType(block1)
--> Response: block
NP: Error: block is not a buck!
--> Visual System Query: listObjects: buck +black
--> Response: null
--> Action Manager call: stop
--> Robot stops moving
NP: Maintaining referent {block1} & using "block"
--> Call to speech production
--> Robot says: "you mean the black block"
```

Notice that immediately after the word "white" it is impossible to tell which item will be moved. After "block", however, the referent is uniquely specified, and action can begin. Back-channels occur after each specification. Though block3 has already been confirmed after the word "white",

[4]These output have been edited and reformatted for legibility and space reasons, but we believe we have provided enough for an understanding of the mechanism of the system.

an additional check is added both to be certain the location is absolutely accurate while the robot attempts to pick up the object, and to perform the additional back-channel response. Reaching for the object can begin here, even before the utterance is complete.

After the word "black", by contrast, reference is already established. In this case, the referent can be anticipated even though the noun phrase is not complete. When the additional word "buck" conflicts with this referent, RISE attempts to justify the two. When it is clear that "buck" cannot refer to the expected referent, it checks to be sure there is no other referent that matches the criteria. Because no black bucks exist, it stops its current motion, and asks for clarification.

If the mismatch had occurred in the middle of the utterance instead of at the end, RISE could easily have interrupted the utterance with its clarification question, or waited until the end to ask.

Notice also that after the word "move", several possible semantic and syntactic interpretations exist for the utterance. The robot could be asked to move to a destination ("move to the door"), to move in a direction ("move to the left"), or to move an object. As the sentence progresses, it can determine that it needs only the last interpretation, and the others are removed from consideration. It is worth noting that should the parse fail entirely (if the sentence was misheard as "move the left"), the sentence could still be statically parsed after the fact. In this case, the speed benefits of incremental processing would be lost (although the robot could still act confused or ask for clarification in the meantime).

4.2 Demonstration of Disambiguation Using Shared Context

We demonstrate here RISE's ability to understand ambiguous phrases using a shared visual context.

We show here the evaluation of the phrase "push the box with the bumper to the wall". This statement could have any of several interpretations. For example, if one box exists and it lacks a bumper (but the robot has one), then the sentence would mean "use your bumper to push the box until it reaches the wall". If two boxes exist and have bumpers, but only box B's bumper is facing the wall, then the sentence means "push box B". If two boxes exist and are both away from the wall, and only box B has a bumper, then the sentence becomes "push box B until it reaches the wall".

We will consider the last of these configuration in order to demonstrate attachment to both the noun phrase and the verb. Below follows reduced output.

```
 ... Processed: push the
 Processing word: box
 Matched noun: name "box" added to NP: [the box]
--> Vision Query: listObjects: box
--> Response: {box7 box8}
 Possible Referents: {box7 box8}
--> Vision Query: confirmObjects:{box7 box8}
--> Robot looks at box7, then box8
 Current Utterance:
 | push *NP:[the box <prep><NP>]* (<prep><location>)
   Confirmed Action: moveObject:?object:?location

 [...]

 ... Processed: push the box with the
 Processing word: bumper
 Matched noun: name "bumper" added to NP: [the bumper]
--> Vision Query: listObjects: bumper
--> Response: {bumper9 myBumper}
 Possible Referents: {bumper9 myBumper}
--> Vision Query: confirmObjects:{bumper9 myBumper}
--> Robot looks at bumper9 (myBumper is known to exist)
--> Vision Query: Relationships:
       {box7 box8}, {bumper9 myBumper}
       Confirm: with
 Object box7 rightOf object bumper9
 Object box8 attachedTo object bumper9 (MATCH!)
 Object box7 inFrontOf object myBumper
 Object box8 inFrontOf object myBumper
 Using: with(attachedTo)

--> Vision Query: Confirm: box7 attachedTo bumper9
--> Robot looks at box7 and bumper9
--> Vision Response: Confirmed

  Comparison Result is: 1 local, 1 parental
     New PR for NP[the box...] : {box7}
     New PR for NP[the bumper] : {bumper9}
     Reference established.

 Current Utterance:
 | push NP:[the box with NP[the bumper]] (<prep><location>)
   Confirmed Action: moveObject:box7:?location
--> Action Manager call: moveObject:box7:?location
--> Robot begins moving towards box7

 ... Processed: push the box with the bumper to
 Processing word: to
 Matched preposition: to
 Current Utterance:
 | push NP:[the box with NP:[the bumper]] to <location>
   Confirmed Action: moveObject:box7:?location
```

Notice that RISE initially has two possible referents for the word "box", and two additional referents for the word "bumper". However, only one box and one bumper can be considered "the box with the bumper". RISE is therefore able to determine reference from the possible referents and the relationships between them.

Notice also that once reference is established, the noun phrases lose attention. That is, further incoming words are considered to be associated with the verb and not with the referents.

4.3 Demonstration of Incremental Language Production

We finally demonstrate RISE's ability to produce grammatical sentences in a manner that is both incremental in nature and conducive to incremental processing by the listener. Again we focus on the generation of unique reference, this time using the blocks arrangement shown in the upper pane of Figure 2.

Unlike speech understanding, where RISE operates incrementally at the word level, speech production occurs incrementally at the phrase level, since it is important to be sure that a referent is uniquely specified by the entire noun phrase.

In order to ensure that referents are neither under- nor overspecified, RISE runs its own listening engine in simulation mode. As each word is uttered, it is interpreted by the simulated listener to be processed. When a referent needs to be specified, the speaker sends the noun descriptor to the listener and checks to see if it is uniquely specified. If it is not, it chooses the most salient adjective to describe the object. RISE will add adjectives in this way until reference

is established, it reaches the limits of its working memory, or additional adjectives no longer reduce the size of the set of referents PR.

If the referent is still not uniquely specified RISE will attempt to add clauses to additionally constrain reference. Because these clauses will not change previous words, adjectives and nouns may be spoken in parallel with these computations. Additional clauses that refer to other referents require the instantiation of a new noun phrase, for which reference must then be established. We use a quick heuristic to determine which adjectives and relationships best specify an object, but the precise method of determination is not important.

RISE receives the command as an action script, as shown below. This representation is quickly converted into a syntactic-semantic structure, and words are fitted to the constraints of that structure.

We assume in this example that the robot is attempting to instruct a human teammate to move a block. The human shares the robot's visually perceivable context, and so overspecification will likely cause confusion. We will examine the instruction "moveObject:block3:atop:block1", where block3 is the white block on the grey block, and block1 is the black block.

```
 Received Action: moveObject:block3:atop:block1
 Script:moveObject:?object:?relationship:?location
 Associated syntactic/semantic frame follows:
    <placeVerb><object><prep><location>
 Executing listener in simulation mode: Ready.

 Processing <placeVerb>
 moveObject placeVerb: using "put"
 Speaking "put"
 Current utterance:
    put NP:[<object>] <prep:relationship>
                            NP:[<location>]

 Processing NP for <object>
 ?object = block3
 Unique unnamed object: Using "the"
 block3 is type "block"
 Attempting "block"
--> Listener: 6 possibilities
 Salient adjective for block3: white
 Adding "white"
--> Listener: 3 possibilities
 Salient adjective for block3: square
--> Listener: 3 possibilities
 Rejecting "square"
 No better adjectives: switching to clauses.
 Speaking: "the white block"
 Current utterance:
    put *NP:[the white block <prep> NP:[<relObj1>]]*
            <prep:relationship> NP:[<location>]

 block3 and block2 have relationship atop.
 Preposition "relationship:atop": using "on"
 Speaking "on"

 Processing NP for <relObj1>
 ?relObj1 = block2
 Unique unnamed object: Using "the"
 block2 is type "block"
 Attempting "block"
--> Listener: 6 possibilities
 Salient adjective for block3: grey
 Adding "grey"
--> Listener: Reference established.
 Established block2.
 Speaking: "the grey block"
--> Listener: Reference established.
 Established block3.
 Speaking: ""
 Current utterance:
    put the NP:[white block on NP:[the grey block]]
         <prep> NP:<location>
```

Notice in each case that specifiers are added to a noun phrase only until reference is uniquely determined.

5. DISCUSSION AND CONCLUSIONS

We have argued that it is important to HRI for robots to meet the expectations that humans hold for the interaction, and that these expectations are different for an embodied system in a face-to-face conversation than in a typed or spoken conversation with a desktop system. We further argued that in order to meet the expectations humans hold for face-to-face communication with an embodied system, a robot's semantic engine must at least perform incremental semantic processing and integrate tightly with both perceptual and action systems.

We presented RISE, our first attempt to define and implement an incremental language processing system designed to respect these expectations in order to improve human-robot interactions. We have further demonstrated some of the system's current capabilities such as dealing with partial and erroneous input, using information from the shared context to discern the meaning of ambiguous phrases, and using simulations of listeners with incremental language processing strategies to produce sentences that are easy for humans to process and understand.

Next steps will include experiments with human subjects, in which we formally evaluate the performance of our system against others, as well as extensions to our system to integrate prosodic analysis for accurate detection and repair of corrections in the verbal stream.

We believe that natural language processing systems that are sensitive to the ways that humans interact and that emulate the styles and conventions of human language processing will become critical components in HRI. By paying attention to the additional constraints that hold in face-to-face conversations between humans and other embodied systems, we believe interactions with these systems will become easier and more natural for the human participants.

6. ACKNOWLEDGMENTS

The authors would like to thank David Anderson and Aaron Dingler for their help with programming and integrating parts of the robot architecture on which our system runs. We would further like to thank Dr. Paul Schermerhorn and Jim Kramer for their work on our architecture, and Dr. Virgil Andronache for his help with the speech production architecture and his early work on the NLP in ADE.

7. REFERENCES

[1] J. F. Allen, B. W. Miller, E. K. Ringger, and T. Sikorski. A robust system for natural spoken dialogue. In A. Joshi and M. Palmer, editors, *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 62–70, San Francisco, 1996. Morgan Kaufmann Publishers.

[2] C. G. Chambers, M. K. Tanenhaus, K. M. Eberhard, H. Filip, and G. N. Carlson. Circumscribing referential domains in real-time language comprehension. *Journal of Memory and Language*, 47(1):30–49, 2002.

[3] D. DeVault and M. Stone. Domain inference in incremental interpretation. In *Proc. ICoS*, 2003.

[4] K. M. Eberhard, M. J. Spivey-Knowlton, J. C. Sedivy, and M. K. Tanenhaus. Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, 24:409–436, 1995.

[5] L. D. Erman, F. Hayes-Roth, V. R. Lesser, and D. R. Reddy. The hearsay-ii speech-understanding system: Integrating knowledge to resolve uncertainty. *ACM Computing Surveys*, 12(2):213–253, 1980.

[6] C. Fellbaum, editor. *Wordnet: An Electronic Lexical Database*. MIT Press, Cambridge, 1998.

[7] M. V. Ferro and B. A. Dion. Efficient incremental parsing for context-free languages. In *Proceedings of the 1994 International Conference on Computer Languages*, 1994.

[8] I. Horswill. Tagged behavior-based architectures: Integrating cognition with embodied activity. *IEEE Intelligent Systems*, September/October 2001:30–38, 2001.

[9] Y. Kamide, G. T. M. Altmann, and S. L. Haywood. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, 49(1):133–156, 2003.

[10] K. Kipper, H. T. Dang, and M. Palmer. Class-based construction of a verb lexicon. In *Proceedings of AAAI 2000*, Austin, TX, 2000. AAAI Press.

[11] O. Lemon, A. Bracy, A. Gruenstein, and S. Peters. Information states in a multi-modal dialogue system for human-robot conversation. In *Proceedings of Bi-Dialog, 5th Workshop on Formal Semantics and Pragmatics of Dialogue*, pages 57 – 67, 2001.

[12] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[13] B. Lowerre and R. Reddy. The harpy speech understanding system. In W. A. Lea, editor, *Trends in Speech Recognition*, pages 340–360. Pretice Hall, 1980.

[14] D. Mori, S. Matsubara, and Y. Inagaki. Incremental parsing for interactive natural language interface. In *IEEE International Conference on Systems, Man, and Cybernetics*, 2001.

[15] A. Newlands, A. H. Anderson, and J. Mullen. Adapting communicative strategies to computer-mediated communication: An analysis of task performance and dialogue structure. *Applied Cognitive Psychology*, 17(3):325–348, 2003.

[16] B. O'Conaill, S. Whittaker, and S. Wilbur. Conversations over videoconferences: An evaluation of the spoken aspects of video-mediated communication. *Human-Computer Interaction*, 8(4):389–428, 1993.

[17] M. Purver and M. Otsuka. Incremental generation by incremental parsing: Tactical generation in dynamic syntax. In *Proceedings of the 6th annual CLUK Research Colloquium*, 2003.

[18] C. P. Rose, A. Roque, D. Bhembe, and K. VanLehn. An efficient incremental architecture for robust interpretation. In *Proceedings of Human Language Technology Conference*, 2002.

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

[20] D. Roy, P. Gorniak, N. Mukherjee, and J. Juster. A trainable spoken language understanding system. In *Proceedings of the International Conference of Spoken Language Processing*, 2002.

[21] M. Scheutz. ADE - steps towards a distributed development and runtime environment for complex robotic agent architectures. *Applied Artificial Intelligence*, 20(4-5), 2006.

[22] M. Scheutz, K. Eberhard, and V. Andronache. A real-time robotic model of human reference resolution using visual constraints. *Connection Science Journal*, 16(3):145–167, 2004.

[23] M. Scheutz, P. Schermerhorn, J. Kramer, and D. Anderson. First steps toward natural human-like hri. *Autonomous Robots*, page forthcoming, 2007.

[24] M. Scheutz, P. Schermerhorn, J. Kramer, and C. Middendorff. The utility of affect expression in natural language interactions in joint human-robot tasks. In *Proceedings of the 1st ACM International Conference on Human-Robot Interaction*, pages 226–233, 2006.

[25] J. C. Sedivy. Invoking discourse-based contrast sets and resolving syntactic ambiguities. *Journal of Memory and Language*, 46(2):341–370, 2002.

[26] L. D. Setlock, S. R. Fussel, and C. Neuwirth. Taking it out of context: collaborating within and across cultures in face-to-face settings and via instant messaging. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, 2004.

[27] Cmu sphinx speech recognition engines. http://fife.speech.cs.cmu.edu/sphinx/.

[28] L. Steels and F. Kaplan. Bootstrapping grounded word semantics. In E. J. Briscoe, editor, *Linguistic evolution through language acquisition: formal and computational models*, pages 53–74. Cambridge University Press, Cambridge, UK, 2002.

[29] M. K. Tanenhaus, M. J. Spivey-Knowlton, K. M. Eberhard, and J. C. Sedivy. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634, 1995.

[30] S. Varges and M. Purver. Robust language analysis and generation for spoken dialogue systems. In *Proceedings of the ECAI workshop on Development and Evaluation of Robust Spoken Dialogue Systems for Real Applications*, Riva del Garda, Italy, Aug. 2006.

[31] F. Weng, L. Cavedon, B. Raghunathan, D. Mirkovic, H. Cheng, H. Schmidt, H. Bratt, R. Mishra, S. Peters, L. Zhao, S. Upson, E. Shriberg, and C. Bergmann. A conversational dialogue system for cognitively overloaded users. In *Proceedings of the 8th International Conference on Spoken Language Processing*, 2004.

[32] T. Winograd. *Understanding Natural Language*. Academic Press, 1972.