SITUATED SPEECH UNDERSTANDING

by Peter Gorniak

Thesis Proposal for the degree of Doctor of Philosophy
at the
Massachusetts Institute of Technology

May 2004

Professor Deb K. Roy
Assistant Professor of Media Arts and Sciences
Massachusetts Institute of Technology

Professor Daniel C. Dennett
Professor of Philosophy
Tufts University

Dr. Allen Gorin
Director, Knowledge Discovery Research Laboratory
National Security Agency
U.S. Department of Defense

Professor Leslie P. Kaelbling
Professor of Computer Science and Electrical Engineering
Massachusetts Institute of Technology

SITUATED SPEECH UNDERSTANDING

Peter Gorniak
April 2004

Abstract

Speech and language do not occur in a vacuum - much of the meaning of an utterance comes from its context, including when and where it was uttered, what the person saying it was doing at the time, who was there to listen to it, and why the speaker decided to speak in the first place. In my research, I aim to unify a number of disciplines such as speech recognition, discourse modelling, plan recognition and natural language understanding into a single coherent framework that deeply understands situated speech. To do so, such a framework has to interact with the world it shares with other language users and model and understand this world in ways similar to theirs to establish a shared context. In this proposal, I describe the elements of such a framework that shares a virtual game playing world with several human players. The framework models the progress of the game in terms of the physical situation, the discourse situation and the players' plans and goals and grounds speech uttered by the players in terms of these rich situational models.

Note: The use of "we" in this document refers to myself and members of the Cognitive Machines group, unless otherwise noted.

1 Introduction

The idea of designing machines that understand human language is older than the oldest electrical computing machines [9]. Speech and language are powerful tools that let us affect the world in sophisticated ways, and it is only natural that we think of a language understanding machine as having the ideal human-computer interface. However, despite decades of intense research on the computational language understanding problem, we do not have machines today that even approximate the natural and flexible ways in which human beings use and understand language.

Encouraging progress has been made in several sub-problem areas. Speech recognition has flourished to the degree that commercial products using such technology have seen some success for constrained recognition settings like business-oriented dictation. Speech recognition has also seen great success in call routing, where learned associations lead to robust behaviour in a constrained task [19]. We have sophisticated understanding of many aspects of natural language phenomena such as grammar and morphology, and many efficient algorithms to parse syntax and other surface level aspects of human language [2]. There are also systems that attempt to tie words and syntactic structures to symbolic representations of a domain of discourse, such as airline reservations or product support information.

Finally, we have seen the emergence of statistical algorithms that exploit large semi-structured bodies of text such as the World Wide Web to link query words to other words in the text so as to retrieve results that are meaningful to the human user [6].

Why have none of these methods, each one useful in its own confined realm of applications, produced a language understanding system that shows some of the flexibility, robustness and ease inherent to human language? I will argue that there are crucial aspects to human language understanding that each of these approaches fails to acknowledge. These aspects can be summarized in the following key statement: *For a machine to understand language, it must assign meaning to words for its own purposes and from its own perspective by autonomously interacting with the world and other language users.* To elucidate the importance of this way of viewing meaning from a machine's point of view, I describe some of its implication in more detail in the following:

Situatedness Language and especially speech are produced in a certain situation. Often, the physical surroundings of the speaker play a large role as possible referents. This is especially clear in the use of deictics such as "this" or "the red one", but pervades all of language use.

Context Meaning depends not only on the physical situation, but also on other aspects of context such as the current state of discourse, shared knowledge and history, as well as shared goals of speakers and listeners. An utterance obtains much of its meaning from this context, leading to seemingly sparse utterances conveying much information, also termed the *efficiency of language* [4].

Embodiment Hand-in-hand with situatedness goes the notion of the language user's embodiment. Without some notion of a body (say, for example, a spatial location occupied by the speaker), situatedness is impossible to achieve. Embodiment contextualizes language in further ways, exemplified most clearly by the use self-relative meaning such as "the one on my left", "this is too heavy for me" or simply "I". Along with the notion of having a body comes that of having a specific body, that not only affords one a point of view, but dictates one's other interactions with the world. The type of these interactions in turn lead one to encode the world differently than one might have with a different body.

Cross-Realm Context While situatedness and embodiment in the physical world are important aspects of language understanding, there are many other domains of discourse that can provide a situating context and a differing instantiation of the speaker's body. To name only a few, we regularly speak of items we or others possess ("Can I have your torch?"), our social embedding ("he is my friend") and our goals ("I'm trying to open that door"). Each domain may require different ways of reasoning and speaking, but

can be linked to others by analogy and the use of indexicals (the meaning of "this one" or "I" bind to any of the domains, and often to several at once). A language understanding system must be able to reason and understand in the context of many domains and decide which domains are applicable for understanding any given utterance.

Dynamic Goal-Based Interaction While embedding language understanding in the context of a body in a concrete situation within one or more domains of discourse lets us begin to tackle some issues of reference and language understanding that cannot be addressed in single-domain unsituated systems without bodies, it still may be tempting to view language understanding as a one-directional encoding process. In this view, the situation is statically encoded by the understanding system, be it human or machine, and words are tied to this encoding to achieve meaning. But in fact, it is purposeful interaction with the world, not passive perception of the world, that leads to the useful encoding, efficient information transfer and rich meaning that language affords human beings.

The goal of the research proposed here is to design and build the first speech understanding system for a virtual environment that exhibits all of the aspects of deep understanding listed above. While none of these aspects will be as rich as they are in human beings, I believe that a system that incorporates even simplified solutions to these demands will demonstrate a robust flexibility in understanding that will encourage further research in more sophisticated situational embeddings. There are two types of machine understanding possible in this context. In the first type, a machine models the perception, situatedness, embodiment and interaction of other language users and understands language that these speakers use to talk to each other. It is easier to collect representative data for this type, as human beings speak to each other naturally in many situations without the need to first implement an understanding machine. In the second type, the machine itself is the embodied and situated conversation partner, with its own perceptive capabilities and goals. In my work, I pursue the path of first tackling the former problem, that of modelling human beings' understanding and situation, but limit the situation at hand and the perception possible on part of the language users to be such that it is possible to gain very similar understanding of the situation and very similar perceptive capabilities when the understanding problem is transferred to a machine that is a full-fledged language user in its own right. I have applied this approach successfully so far, and this proposal follows the same approach.

To record speakers in a rich yet measurable environment that includes a notion of embodiment, spatial location, possessions, roles and goals I turn to graphical multi-user role playing games. Players in these games use language to communicate within the game, to coordinate their action and to achieve their goals. The game setting also allows me to shape the scenario players find themselves in, to assign roles and impose problems, making otherwise hard-to-measure features

more easily estimable. Finally, the game setting is not only an effective research platform, but also has immediate relevance to the large and growing video game entertainment market, which currently does not include any games that use speech understanding. The type of situated and robust speech understanding that is the goal of this research would open the doors to a whole new type of game and gaming interface.

This research necessarily draws from a variety of fields including speech recognition, natural language processing, qualitative and causal reasoning, reasoning under uncertainty, linguistics, philosophy of language and mental representation, cognitive science, speech and discourse understanding and plan recognition research. As a thoroughly interdisciplinary project that puts the problem of understanding language, one of the core means of human expression, this work fits ideally into the Media Laboratory's research agenda, and especially within the Cognitive Machines group's interests.

2 My Proposed Research

In the following sections I detail the elements and results of my research so far. I also explain in each section how I will proceed to expand and change the prior research towards my newly set goals. Figure 1 shows an overview diagram of the architecture of the proposed approach. Depicted on the outside is a cycle of physical actions and situations as well as utterances made by players. Utterances and physical actions are interwoven and influence each other. The system listens to speech via its grounded language parser, which during the parsing process connects the words heard to current, past or future situations within the game. Aspects of these situations, namely the physical aspects, the discourse state and a hierarchy of possible player plans serve as possible referents during the grounding process, and are in turn updated by the grounded parser when the utterance is understood. On the other side, a situation tracker maps the event of actions occurring in the game to the situational models to let parsing occur in a fully up-to-date context at any given time. Each element of this diagram is explained in detail in the following sections.

2.1 Settings and Data Collection

I have so far primarily investigated what I will call *egocentric visually grounded language use*. All language that played a role in my projects so far was primarily about visual aspects of a scene. I collected data from human beings describing objects they saw, either real objects on a table, or virtual objects rendered on a screen. Beyond attributes of the objects such as colour and shape, I was primarily interested in how people use spatial relations and groupings in describing objects ("the green one behind the three purple ones"), and set up data collections that

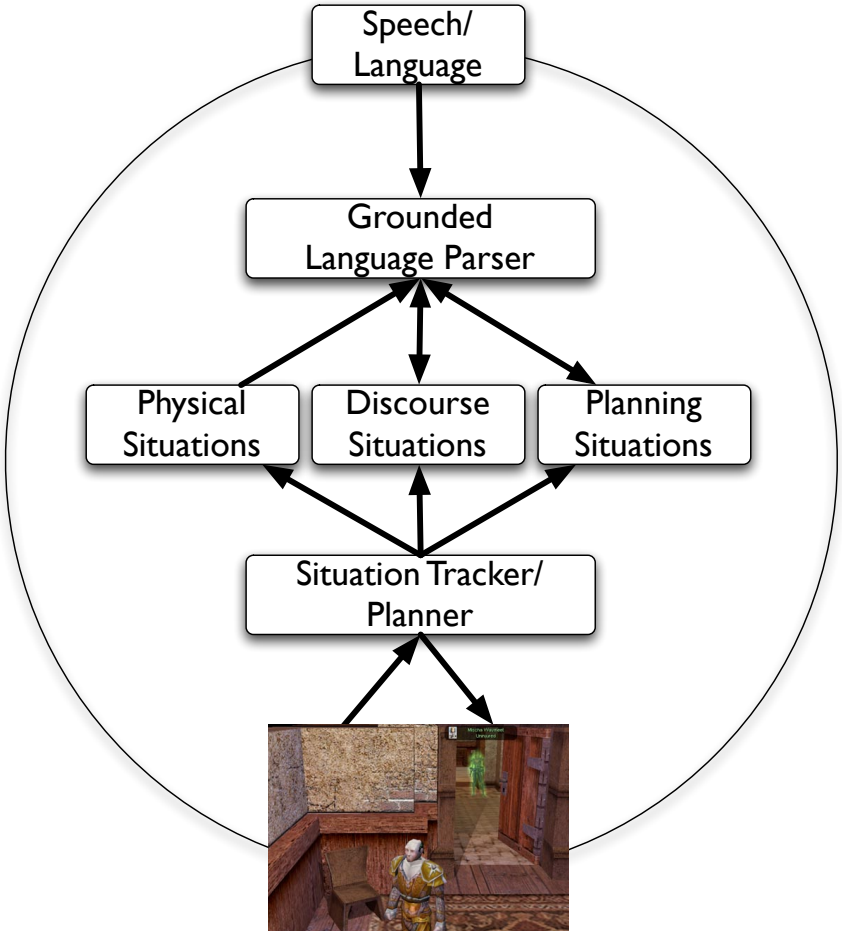

Figure 1: The architecture of the proposed system

naturally caused participants to use such constructs to distinguish amongst many otherwise identical objects. I call this visual setup, and the spatial relations participants used, 'egocentric', because they assume a specific viewpoint associated with an agent. We later transferred this viewpoint based language understanding system to Ripley, our robot, and added the ability to refer to others' viewpoints ("my left" vs. "your left"). However, the language embedding and as a result the language used are still relative to a specific agents' visual perception of the world.

One way to push this research further is to stay with the immediate physical surroundings as a situation, but to enrich the language grounding in this situation to include aspects of language use such as past and future events, plans and goals, hypothetical situations and more complete models of the physical world that include notions of touch, object topology and the function of objects. One's immediate physical surroundings, however, are only a small part of what we regularly talk about. Another, and at least at first glance somewhat orthogonal challenge, is to show how the language grounding strategies that we have developed over

the years transfer to other domains of human discourse, such as talk about one's location in space, possessions, social relations, abilities, and goals. In the research proposed here, I am aiming to show that our physical grounding philosophy can be extended to cover these different domains of human discourse without losing its important feature of being based on a continuing interaction with the embedding situation.

Just as was the case for the egocentric visual grounding of language, I propose to start the investigation using a virtual environment, and having human participants speak about this environment. A virtual environment has the advantage that it can be designed to elicit the types of language use that I am proposing to cover here, while simplifying the sensing and action problems robots face in the real world. A virtual environment is especially necessary in this case as we do not have robots that can robustly navigate an environment, be aware of their location, have objects in their possession, perform acceptable speech recognition, perform physical actions that are of use in a realistic setting and be aware of their own and their collaborators capabilities and goals. Furthermore, I have found it to be an extremely useful to first collect data from people interacting with other people in the environment of interest. This provides a baseline because we know of this data that it was at least understood by other human beings in the same situation. In many ways, embedding human players in a virtual environment has the nature of a Wizard of Oz study, where the players are both the subject and the wizards. We do not yet have an artificial agent that can understand situated language with all the facets described in the introduction. However, human beings are such agents, and by embedding them in a virtual environment that makes most aspects of their interaction and its context easily recorded, I hope to produce a set of data that allow me to design a stand-in artificial agent for any player. This agent can sense everything that the player sensed, and is thus put in the same situated language understanding situation as the player, with the only difference that it cannot actively make decisions, but rather has to follow the decisions the human player made when the data was recorded. Note that the agents still has to model and understand these decisions to be able to understand the accompanying language, it just does not decide itself. After this agent is designed and successful at following along a player's path and understanding the situated language that occurs along the way, it can be given its own decision making mechanism and let loose on the world to share it as an autonomous entity with other players.

The virtual environment I have prepared for data collection in the last few months is that of online, graphical multiplayer role-playing games. These games embed the player (via his or her avatar) in a rich graphical world that includes a myriad of objects, rooms, doors, creatures, other players and problems to be solved. Each avatar has pre-assigned capabilities and a resulting clearly defined role (for example, that of being a fighter who is strong and capable in combat vs. that of being a thief who can hide, pick locks and disarm traps). As a result, each character also enters in interesting social relations with the other characters,

which develop further as the game progresses. Finally, this environment is ideal because it comes with tools that allow one to shape the physical surroundings of the characters, the objects they will find and have to interact with, and the creatures and quests they will encounter. Players regularly make use of typed language when playing these games. Figure 2 shows a sample of this type of language collected from several players playing a popular game module.

```
player 3 : elrlilia - stay still!
player 3 : did you get it?
player 2 : you want me to drink this?
player 3 : yes
player 2 : isn't it better saved for battle?
player 1 : I can rest in here
player 3 : or - we could rest
player 2 : let's go rest
player 1 : I am the king of the castle
player 2 : whew
player 2 : fabio; did you loot all those chests?
player 3 : one chest
player 2 : weapons racks?
player 3 : two weapon racks - some arrows, not much
player 2 : ok
player 1 : anybody heavier than this bolder -
              please step onto the platform
player 2 : gimme a sec...
player 3 : do it do it!
player 1 : wish I could
player 1 : anyone got something heavy to throw in?
player 3 : maybe we need to convince somebody GIB
player 3 : BIG
player 2 : we missed a room up top
player 3 : have you guys been down this corridor?
player 2 : i checked the corpse, just gold
```

Figure 2: Gaming Language Sample

I have instrumented a popular game of this type, Neverwinter Nights (shown in Figure 3), so that I can collect not only the text users type (like the sample above), but also their movements and actions, such as item pick-ups and drop-offs, doors opened and levers pulled. Furthermore, the game world can be scanned for object and room locations. The data will therefore consist of a complete record of the game situation, physical changes to the situation, player actions and player text messages. In addition to the online collection that only includes typed text, I

Figure 3: The Neverwinter Nights graphical role-playing game

will also perform some in-lab data collection where I will record players' time synchronized speech instead of text messages.

Having collected the data, I will split them into a training set that I will use to build and train the understanding system and a test set, that I will set aside to test the system once it is complete. I will annotate the training data in several ways. I will identify grammatical constituents in the data and bracket them to train the natural language parser, discussed below. For each constituent, I will also mark whether it refers to an object in the game environment, and, if so, to which one. This includes objects such as sub-tasks of quests to be solved (e.g. the unlocking of a door that bars the way onwards) or abilities of characters ("Can you unlock this door?"). I will also mark how each constituent refers to its referent (e.g. by egocentric spatial relation ("the one in front of me") or allocentric spatial relation ("the one in the North-East corner") and what type of action (or speech act) the whole utterance constitutes (e.g. informative description ("there's a chest here") or request for help ("can you open this chest for me?")). What types of reference and types of speech acts cover the data will only be clear after looking through the training data, just as I did in my previous egocentric visual reference projects. Once the system is built, I will annotate the testing data in the same way for evaluation purposes.

2.2 Speech Recognition

In previous work I have used my group's own speech recognizer, the sphinx 4 speech recognizer as well as manually transcribed speech. In each case, however, the input to the language understanding system was a single string of text, namely the best guess of the speech recognizer given the acoustic signal, or the manual transcription if no speech recognizer was used. This leads to brittleness in a live system like Ripley, because a mis-recognized utterance cannot be corrected at later stages of understanding. Rather, the system either signals a failure, or performs an inappropriate action if the words it the speech recognizer produced could be interpreted.

In the proposed research I will use the Sphinx 4 speech recognizer which produces probabilistic lattices as output. As will be explained in the following sections, accepting probabilistic lattices as input to the natural language parser will allow me to chain a probabilistic n-best list of interpretations of the acoustic model all the way to the semantic grounding in the current situation. As a result, all levels of processing can refine the decision as to what the speaker should be taken to have said. For example, while the speech recognizer might produce a lattice that has "toes the boar" as its most likely sentence given no other information, a pass through the parser might produce "close the boar" as a more likely path through the lattice, whereas the grounding in the situation at hand might correct this once more to read "close the door" (this example is for illustration purposes, as both the speech recognizer and the parser might well correct the utterance earlier due to the statistics of words and syntax they maintain).

I will use both the text training data collected and the transcribed speech training data to construct trigram language models for the speech recognizer using standard tools.

2.3 Language Grounding

The data collected from people describing objects in cluttered scene to each other yielded a set of *descriptive strategies*, combinations of linguistic patterns and visual features, that speakers employed in combinations to unambiguously identify referents. I identified the main strategies covering most of the collected data, and designed and implemented corresponding modules that attach to the rules of the natural language grammar as well as the lexical entries [20].

Conceptually, I treat lexical entries like classes in an object oriented programming language. When instantiated, they maintain an internal state that can be as simple as a tag identifying the dimension along which to perform an ordering, or as complex as multidimensional probability distributions. Each entry also has a function interface that specifies how it performs semantic composition. Currently, the interface definition consists of the number and arrangement of arguments the entry is willing to accept, whereas type mismatches are handled during composition rather than being enforced through the interface. Finally, each entry can

contain a semantic composer that encapsulates the actual function to combine this entry with other constituents during a parse. The lexicon may contain many lexical entries attaching different semantic composers to the same word. For example, "left" can be either a spatial relation or an extremum. The grammatical structure detected by the parser (see the next Section) determines which compositions are attempted in a given utterance.

During composition, structures representing the objects that a constituent references are passed between lexical entries. I refer to these structures as *concepts*. Each entry accepts zero or more concepts, and produces zero or more concepts as the result of the composition operation. A concept lists the entities in the world that are possible referents of the constituent it is associated with, together with real numbers representing their ranking due to the last composition operation. A composer can also mark a concept as referring to a previous visual scene, to allow for anaphoric reference. It also contains flags specifying whether the referent should be a group of objects or a single object ("cones" vs. "cone"), and whether it should uniquely pick out a single object or is ambiguous in nature ("the" vs. "a"). These flags are used in the post-processing stage to determine possible ambiguities and conflicts.

The chart parser incrementally builds up rule fragments in a left to right fashion during a parse. When a rule is syntactically complete, it checks whether the composers of the constituents in the tail of the rule can accept the number of arguments specified in the rule. If so, it calls the semantic composer associated with the constituent with the concepts yielded by its arguments to produce a concept for the head of the rule. If the compose operation fails for any reason (the constituent cannot accept the arguments or the compose operation does not yield a new concept) the rule does not succeed and does not produce a new constituent. If there are several argument structures or if a compose operation yields several alternative concepts, several instances of the head constituent are created, each with its own concept.

The framework as it stands makes a number of simplifying assumptions with regards to the domain of reference of an utterance, and the speaker's intentions: the domain is always the set of immediately visually perceivable objects (and groups thereof), whereas the intention was always assumed to be one of pure description and reference. In the version of this framework that runs on Ripley, we have extended the possible intentions of the speaker to include commands to perform various actions on single objects. One of the primary goals of the research proposed here is to loosen both assumptions.

To extend the existing framework to cover multiple domains in parallel, it is necessary to augment both the features measurable, as covered in the last section, and the grounding performed during parsing. In dropping the assumption that all utterances are descriptions or simple action requests involving objects currently physically present, it becomes necessary to determine which domain an utterance refers to, and what the speakers intention in producing the utterance is. There are

two sources of information which can be used to make these determinations: the words said, and how they tie to measurable features of the current world state, and a discourse and intention history that may bias the understanding system one way or another. Here, I only cover the grounding of words in the currently measurable world state, whereas intention tracking is left for Section 2.5.

In my work so far, it is already the case that words can be grounded in several distinct ways, such as "left" exercising a different grounding in "to the left of" versus "the left one". However, due to the one-domain assumption made, words like "one" could be assumed to be referring to visible physical objects only. Given the multi-domain nature of the language collected in the gaming environment proposed, words must now be able to tie to entities in one or more domains. Some words may only tie to one or two domains, such as "left" being primarily concerned with ego- and allocentric spatial language. Others, such as "one" or "I" are domain-transcending. Each utterance, however, can be taken to have at least a primary domain of reference - "do you have the sword?" falls clearly in the domain of possession, whereas "There's a chest in the South-East room" is an allocentric description of the physical world. Thus, not only will I build new grounding and semantic composition modules to cover the new domain, but the grounding process will have to estimate several things for each utterance interpreted:

1. what is the domain of reference for this utterance?
2. which objects in the domain does this utterance refer to?
3. what proposition is being expressed?
4. what speech act is the speaker engaging in, and why (in terms of the speakers goals being tracked)?

Note that each utterance may have several distinct set of referents, especially when it involves transitive verbs. Finding answers to these questions involves tying words to possible referents, just as in the egocentric visual grounding case, and composing the semantics of words, which applies constraints that narrow down the possible sets of referents and intentions. I believe that my existing framework will extend to cover these new demands.

To support the probabilistic integration with speech recognition, parsing and intention tracking levels, I have recently unified the modules that tie words to the world by using Tenenbaum's example generalization algorithm, which is quickly trained from very few examples and gives an estimated probability of how likely a new example belongs to a class of previous named examples [50]. Thus, at the end of the parsing and grounding process, the system will produce probabilities over possible domains, referents and intentions that integrate all information considered by the system.

As the game language sample above shows, the language used by players is highly elliptic, contextual and unconstrained. As in the egocentric visual case, it

is unlikely that I will be able to design a system that covers all language used. However, by designing a more constrained gaming world that lets the system be informed not only about the physical layout of objects in the world, but also give it a good idea as to players' possible goals at any given time. It should be possible to cover the majority of utterances in this setting, which would mark a significant step forward in situated language understanding.

2.4 Parsing and Grounded Semantic Composition

As mentioned before, I have so far used a deterministic bottom-up chart parser. A bottom-up parser was suited to the often ungrammatical utterance people produce in freeform object descriptions, because it produces all possible sub-parses of an utterance even if the utterance as a whole cannot be parsed. The resulting chart is a rich analysis of the utterance heard, and can be analysed for contradictions, ambiguities and underspecification. The parsing process itself also doubles as a convenient driver for the language grounding process, under the assumption that the constituents produced by the parser are units that can be bound to the grounding process, be it via visual reference to objects ("the green ones") or modification of other grounding processes ("my left" vs. "your left").

There are two drawbacks to the types of parser used in the work so far. The minor drawback is one of efficiency, well known about pure bottom-up parsers, but exaggerated in our case because parsing drives grounded semantic interpretation. A bottom-up parser produces all sub-parses of an utterance, even those that cannot be part of higher level syntactic structures due to the nature of the grammar. While this has an advantage because in our case we cannot be sure that an utterance as a whole will actually be grammatical according to the grammar, it is still true that many constituents are produced during the parsing process that could be eliminated by specifying that the system needs to parse at least to, say, a noun phrase or a verb phrase to be able to act on an utterance at all.

The second, more serious, drawback of using a non-probabilistic parser is that alternatives produced by the parser are hard to compare to each other, and that probabilities produced by either the speech recognizer or the semantic grounding system cannot easily be folded into the actual parsing process. In our system, the parser currently works from a non-probabilistic best hypothesis from the speech recognizer, and the only influence of the semantic grounding on the actual parsing process is in the case that no possible referents are produced, in which case the current parse tree being pursued is discarded. Furthermore, in interpreting the chart produced by the parser, I currently use the semantics produced by the grounding engine together with heuristics involving coverage of a parse within the utterance to decide between different parses of the same utterance. However, this does not allow for a an good decision between different interpretations that is equally informed by the acoustic signal, the syntactic likelihood and the semantic binding of the candidates.

We have recently implemented an probabilistic Earley parser that works both top-down and bottom up simultaneously. In addition to producing parse trees of sentences, this parser computes the probability of each parse given the grammar, where the probabilities encoded in the grammar can be learned from the data I will collect. This parses solves both of the problems described above, while retaining the advantages of a bottom-up chart parser. The probabilistic lattice produced by the speech recognizer can be flattened to resemble the bottom part of a probabilistic chart, specifying which words were heard at certain times, together with probabilities of their occurrence. The probabilistic Earley parser works directly off this pre-initialized chart, incorporating the speech recognizer's probabilities directly into the parsing process without performing a separate parse for each hypothesis encoded in the lattice. During the parsing process, this parser computes the probability of each constituent produced. Given the the grounding engine already computes probabilities representing the likelihood with which each constituent refers to objects and actions in the world, these probabilities can also be taken into account when calculating the overall probability values for the various candidate parse trees. The final decision between different interpretations is thus made easier and more informed, as it incorporates information from all understanding modules in a rigorous manner.

2.5 Situation Tracking

So far, I have discussed language understanding as a problem of tying words to various representational levels of the current situation. In my work to date, I simplified the understanding problem to only deal with the current static situation, or at most one previous situation. In the proposed multiplayer game setting, however, it will be of utmost importance to track many aspects of the situation as it develops, and to maintain a history of salient past situations. I use *situation* here much like Barwise and Perry use the term [4]. In the game environment, I envision the situation to include at least the

physical situation including the player's location, the current time, visible objects and current possessions

discourse situation including the words uttered, the addressee(s), previous discourse turns and their types, and possible targets for anaphora

planning situation including goals and subgoals currently held by players

All three of these aspects of the total situation are intricately linked and can inform each other. All three are also not easily observable, for even though the perception problem is simplified in the game world, knowing the distances objects are away from the player does not tell us which object he or she has actually seen, is currently paying attention to, or is salient in the current situation. Note also that both utterances and non-linguistic actions should play into our tracking of the

situation: utterances' references and speech act types can be disambiguated given the current situation as described in previous sections, whereas their interpretation should update our belief about the current situation. Similarly, non-linguistic actions such as player movement or object manipulation can be interpreted in the light of the current situation and inform beliefs about future situations. Finally, to resolve anaphora and deictic references in utterances and to estimate hierarchies of plans and discourse levels it is not only necessary to track the current situation, but to search and integrate a history of situations.

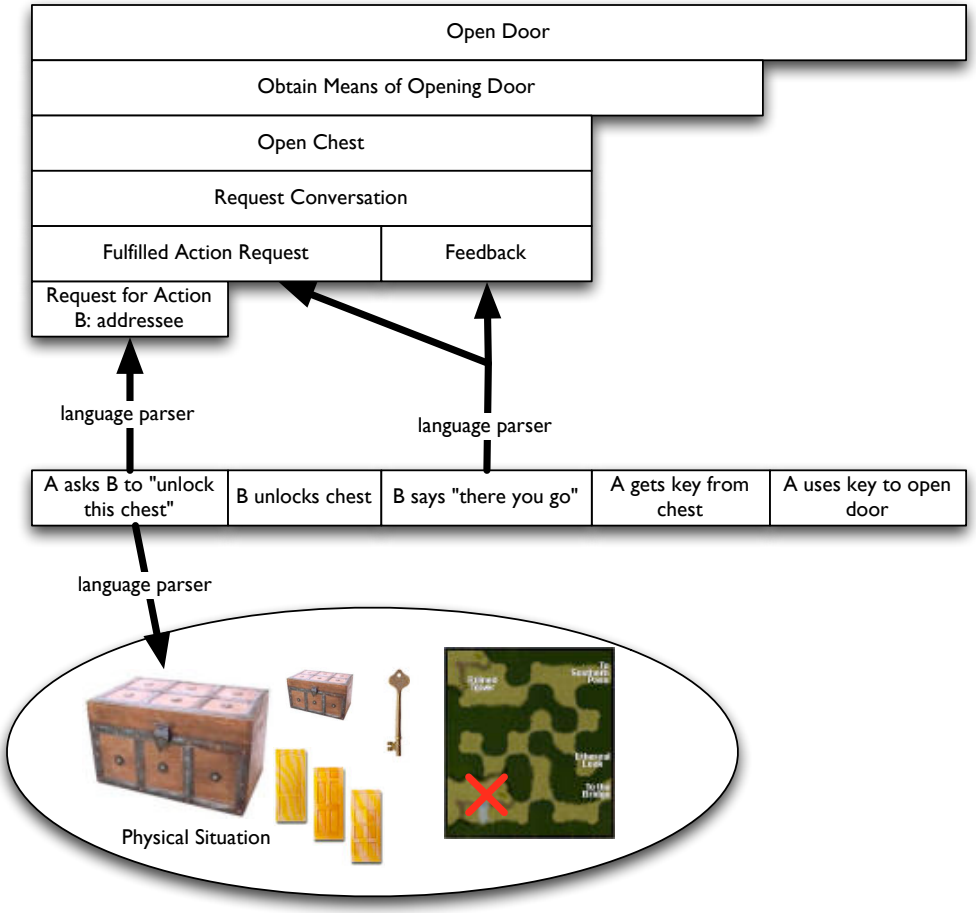

Figure 4: A sample parse of hypothetical game events

To build a high level hierarchical model capable of tracking the current and past situations, I propose to first abstract the raw temporal data (movements, utterances, events) into a higher level description that only contains possibly salient events such as entering a new room, nearing an object, pulling a lever. Furthermore, I propose to parse this event timeline once more with a probabilistic context free grammar parser. The grammar for this parser is largely dictated a priori by the game's design. For example, a game scenario might have the top level

goal of finding a certain artifact. To find this artifact, the players have to make their way past a partially ordered sequence of puzzles, each of which is a constituent in the top level rule, for example TOPGOAL → PUZZLE1 PUZZLE2 PUZZLE3. Each puzzle in turn has known elements, for example PUZZLE1 → OPEN_DOOR(DOOR1). Note that goals refer to other objects, and that it is important that other elements of the current situation change rule probabilities. For example, a specific door might be opened by a key that must be found first, or one of the players might be able to bash in a non-metal door. The probabilistic parse is thus not situation-independent [36]. Figure 4 illustrates the utility of such a parse on a sequence of hypothesized game events. In this sequence, player A is looking for means to open a door blocking the way, and asks player B to unlock a chest, in which he finds the key to open the door. A parse of the events yields top level goals and subgoals as well as discourse segments and participants. The language parser described before ties words to referents either in the game setting ("chest") or in the history items encoded in this event parse, such as the discourse and action event in which player B fulfilled player A's request. Modelling the event sequence in this way thus provides a rich substrate for grounding utterances not only in the immediate game world, but also in items seen in the past, the player's goals and conversational elements.

The probabilistic parsing framework once more allows for robust treatment of the uncertainty inherent in listening to player's utterances and interpreting their actions. It coherently incorporates evidence from the event stream and the language parser, and provides suitable priors and abstract entities to disambiguate utterances and find appropriate referents. While general probabilistic plan recognition is a hard and often intractable problem, the advantage in the setting proposed here is that many of the higher level plan items, such as which steps the players need to do in which order to make it through the game, are known a priori and can be encoded in the grammar used for event parsing. This a priori structure together with easy sensing of the game setting and independence assumptions in the parsing process such as those in [36] should make the problem tractable in this specific setting.

Note that the framework proposed in this section can be co-opted for planning in addition to plan recognition, making it possible to turn the whole system from a pure recognition and modelling system into one that directly interacts with the world and other players.

2.6 Evaluation

Once I have designed the system based on the training data, I will run it on the testing data. At this point, the testing data will be marked up manually, just like the training data was before, with transcripts of the speech, bracketed grammatical constituents, speech act tags and referents in the game world, discourse states or player's plans (i.e. paths through the plan recognition grammar). I will then

evaluate accuracy in terms of percentage of correct speech act classification, object referents, and discourse and plan classification and referents. I will analyse failures in detail. Depending on the nature of the data, it may well be necessary to have several independent human judges mark up parts of the data establish a baseline performance in the human interpreter case.

Another form of evaluation will measure the improvements over the various stand-alone components of the system: How much is speech recognition performance improved by performing situated understanding? How much is parsing performance improved? These questions can be answered by running each component separately on the test data and comparing the result with that of running the full understanding system on the same data.

Secondly, I will build a computer-controlled character for Neverwinter Nights that is driven by the language understanding framework proposed here. It will be limited in its range of actions compared to human players, because aspects like in-game puzzle solving ability or language and speech generation are not the focus of this research, but would be necessary for a human-equivalent artificial player. However, an artificial character can still serve as an in-game aid to players. Using this role, I will record more data, this time of human players adventuring together with their computational servant. This will let me evaluate new elements of the system, such as command understanding rate as well as the usefulness of the framework when making its own decisions.

3 Contributions

It is my hope that this thesis will produce the first system that can be considered to be performing deep speech understanding in the sense that it interactively connects an acoustic speech signal to words, sensed external objects, possessions and speaker's plans and intentions.

Specifically, I hope to show

- that games are a viable platform to study situated speech understanding
- that speech understanding is more easily performed by considering and modelling the shared situational context of an utterance, including the physical situation, discourse context and speaker's plans
- that the non-linguistic situation an utterance occurs in disambiguates many aspects of an utterance including referents and the speaker's purpose
- the richness of meanings of indexicals, especially those referring to agents, such as "I"
- an implemented framework that performs deep situated speech understanding, performing better than any of it's individual components by combining their activity

4 Background and Related Work

There are several research areas that cover aspects of the problem at hand, or address related problems. In the following, I discuss these areas and the most relevant works in them.

4.1 Relevant Linguistics and Philosophy

I have already hinted at a view of the work proposed here as a continuous, computational and interaction-based view of Barwise and Perry's situation theory of natural language [4]. By emphasizing the continuous situational modelling and interaction aspect of speech understanding, I align my work in spirit with writers like Smith and Bickhart [48, 5]. Their thoughts about how a subject can come to acknowledge the existence of objects at all come in at a much more fundamental level than the one the work proposed here addresses. However, they emphasize the importance of seeing representation not as a passive encoding of an external world, but as an interaction process that involves action as much as sensation. I believe this view and criticism transfers directly from Smith's metaphysical worries and Bickhart's logical incoherence arguments to a framework such as the one proposed here. Seeing the problem not as one of encoding the world in a suitable way, but as one of interacting with the world in ways similar to other language users, one's focus shifts from encodings to the importance of goal-based behaviours at all levels, which is a shift that influence the design of a language understanding system.

A further important influence on the work proposed here are notions of consciousness and the self in Philosophy of Mind. Dennett points out the complexity hidden behind a word like 'consciousness', and argues that it is a collection of processes that can be explained [13]. I believe it is time to commit to building complex systems with parts interacting with each other and with the outside world in rich ways if our aim is to have truly intelligent and deeply understanding machines emerge. In the context of the work proposed here, I hope to also elucidate the richness of human concepts related to consciousness and labelled by words like "I" and "you". Some writers in philosophy have acknowledged some of the complexity behind establishing meanings for these words in terms of, for example, egocentric and allocentric spatial locatedness [22, 17]. In my view, there are still many more facets of these concepts remaining to be explored, and I hope that the platform proposed here will shed some light on the different senses "I" takes in different situations and domains.

4.2 Speech Recognition

Speech is a very natural and spontaneous means of human expression, often superior in convenience and naturalness to typed input. At the same time, speech is

a far noisier input modality than typed text, both due to the problems in capturing and analysing audio, and due to its online spontaneity, allowing speakers to delay, rephrase and underspecify while still being understood by listeners. Listeners are able to understand such noisy speech not only due to their sophisticated acoustic and linguistic processing capabilities, but also due to the fact that all speech occurs embedded in a situation that allows listeners to perform disambiguation and understanding. Speech that is hard or impossible to understand when heard without situational grounding becomes easy process when the situation is shared between speaker and listener. The research proposed here focuses on analysing the embedding of language in a situation on multiple levels, and thus it is natural to use speech as an input medium to gain its advantages in human machine communication and show that situational embedding helps greatly with improved understanding of noisy spontaneous speech.

The most common approach to speech recognition today entails the use of Hidden Markov Models (HMMs) to estimate word string probabilities from acoustics [28, 37]. Improving speech recognition via situational embedding implies that the speech recognizer cannot be treated as a black box that turns an acoustic signal into a stream of words. The tight coupling approach considers syntactic (and, by extension, semantic) information while interpreting the acoustic signal [25]. This is also the approach we have used in previous work to integrate models of visual attention and semantic knowledge directly into the language model used for speech recognition. Modern speech recognizers, however, give access to the compactly represented n-best resulting word strings with associated probabilities. In the sequential coupling approach, parsing and semantic interpretation are performed on this n-best lattice representation [11]. In my work I plan to favour this loose coupling approach as it allows for the revising of estimates made by the speech recognizer without the added overhead of implementing an interpretation system integrated into the recognition process.

An important backdrop for the work discussed here are also call routing systems that perform speech recognition and a form of speech understanding in that they carry on a dialog with the called and take routing actions due understanding the user's speech. While not situated in the sense proposed here, these systems are robust and widely deployed in their limited domain. Most interestingly, the semantic associations for words and utterances are automatically learned from speaker data [19], which will also be the case for elements of the system proposed here.

4.3 Speech and Language Grounding

Winograd's SHRDLU is a well known system that could understand and generate natural language referring to objects and actions in a simple blocks world [52]. Like our system it performs semantic interpretation during parsing by attaching short procedures to lexical units [29]. However, SHRDLU had access to a clean

symbolic representation of the scene, whereas the system discussed here works with a continuous virtual world and reasons over many domains in addition to the base physical layer. Furthermore, we intend our system to robustly understand the many ways in which human participants verbally interact in a situated game setting, whereas SHRDLU was restricted to sentences it could parse completely and translate correctly into its formalism.

Word meanings have been approached by several researchers as a problem of associating visual representations, often with complex internal structure, to word forms. Models have been suggested for visual representations underlying colour [26] and spatial relations [38, 39]. Models for verbs include grounding their semantics in the perception of actions [47], and grounding in terms of motor control programs [3, 31]. Object shape is clearly important when connecting language to the world, but remains a challenging problem in computational models of language grounding. In previous work, we have used histograms of local geometric features which we found sufficient for grounding names of basic objects (dogs, shoes, cars, etc.) [42]. This representation captures characteristics of the overall outline form of an object that is invariant to in-plane rotations and changes of scale. Landau and Jackendoff provide a detailed analysis of additional visual shape features that play a role in language [27]. For example, they suggest the importance of extracting the geometric axes of objects in order to ground words such as "end", as in "end of the stick". Shi and Malik propose an approach to performing visual grouping on images [46]. Their work draws from findings of Gestalt psychology that provide many insights into visual grouping behaviour [51, 14]. Engbers et al. give an overview and formalization of the grouping problem in general and various approaches to its solution [16].

Our model of incremental semantic interpretation during parsing follows a tradition of employing constraint satisfaction algorithms to incorporate semantic information starting with SHRDLU and continued in other systems [23]. Most prior systems use a declaratively stated set of semantic facts that is disconnected from perception. Closely related to our work in this area is Schuler's, who integrates determination of referents to the parsing process by augmenting a grammar with logical expressions [45], much like we augment a grammar with grounded composition rules (see Section 2.4). Our emphasis, however, is on a system that can actively ground word and utterance meanings through its own interactions with the world. Even though the system described here senses a synthetic scene, it makes continuous measurements during the parsing process and keeps track of the situation even when no utterances occur. Schuler's system requires a human-specified clean logical encoding of the world state, which ignores the noisy, complex and difficult-to-maintain process linking language to a sensed world. We consider this process, which we call the grounding process, one of the most important aspects of situated human-like language understanding.

SAM [7] and Ubiquitous Talker [30] are language understanding systems that map language to objects in visual scenes. Similar to SHDRLU, the underlying

representation of visual scenes is symbolic and loses much of the subtle visual information that our work, and the work cited above, focus on. Both SAM and Ubiquitous Talker incorporate a vision system, phrase parser and understanding system. The systems translate visually perceived objects into a symbolic knowledge base and map utterances into plans that operate on the knowledge base. In contrast, we are primarily concerned with understanding language referring to a continuously sensed and modelled world.

In tackling issues of indexical resolution and general speech understanding via multimodal sensing, the work proposed here is also related to work in multimodal interfaces that incorporate a speech component [32]. The shared world proposed here, however, is of much greater diversity and immersiveness than that typically shared with the user, in, say, a pen based map input device. In the world proposed here, we can study many aspects of the actual embedding of people in the world, such as physical locatedness and collaboration in a spatial environment. Furthermore, our focus is not the integration of two input modalities, but the understanding of speech embedded in a rich context.

In the realm of games and language, Chapman's Sonja system addresses some of the issues raised here [10]. It's focus is less on the language understanding problem than on the use of terse instructions to an otherwise autonomous video game playing engine. It is similar to the work proposed here in that it perceives the video game situation in human-like ways and understands goals and linguistic referents in relation to this situation. However, I am primarily interested in a far more sophisticated version of the language understanding problem in a similar setting, a version that includes speech, multiple domains of reference, free-form language use and explicit goal and discourse state tracking, all of which are not present in Sonja.

We have previously proposed methods for visually-grounded language learning [42], understanding [41], and generation [40]. The main work I propose to extend here was a visually grounded system that could understand complex human descriptions of objects in cluttered scene [20]. I distinguish between aspects of this work already implemented in this previous work and proposed new elements in the preceding sections.

4.4 Natural Language Parsing and Understanding

Chart and Earley parsers are well studied parsing frameworks amenable to the task outlined here due to their ability to parse bottom-up in a partial manner [2, 15]. The probabilistic version of the Earley parser was designed specifically for applications in speech recognition using tight coupling [49, 25], but can equally be used for a sequential coupling setup [11].

SHRDLU is based on a formal approach to semantics in which the problem of meaning is addressed through logical and set theoretic formalisms. Partee provides an overview of this approach and to the problems of context based meanings

and meaning compositionality from this perspective [34]. Our work reflects many of the ideas from this work, such as viewing adjectives as functions. Pustejovsky's theory of the Generative Lexicon (GL) in particular takes seriously noun phrase semantics and semantic compositionality [35]. Our approach to lexical semantic composition was originally inspired by Pustejovsky's qualia structures.

Many of the symbolic counterparts to the ideas in my work are explored at lengths in the computational language understanding literature [2]. However, it is especially work on situation semantics that considers many of the same crucial issues of language understanding I investigate here, such as the influence of the situation an utterance occurs in on the utterance's meaning [4]. While the original work provides a rich ontology of possible influences on meanings such as roles, situation types and courses of events, it neglected actual grounding in the world when Barwise and Perry casts it as a logical framework. I share the same intuition and the same accompanying criticism of any theory of semantics, such as possible world semantics, that neglects the influence of the situation on meaning. In addition, I believe that embedding in and interaction with a world are key to human-like rich meaning, as opposed to the one-to-one mapping *anchors* that link symbols to referents in the original situation semantics theory.

4.5 Discourse Tracking and Plan Recognition

Being cast in the larger framework of time-extended interaction of several players with a virtual world and each other, it becomes necessary to acknowledge and track higher level and longer features of communication and interaction such as discourse and plans [44, 21]. Both are seen as involving hierarchies of subgoals, and it is only natural that context free grammars, which are good at capturing and recognizing time extended hierarchies with specific independence structures, have been proposed for modelling both plans and discourse phenomena. Specifically, recent work in probabilistic plan recognition has looked to probabilistic context free grammars and closely related frameworks for determining possible plan hierarchies from agents' behaviours under uncertainty [36, 8]. These improve on the previous uses of probabilistic models for plan recognition by acknowledging and exploiting the hierarchical structure of plans that make flat probabilistic plan recognition attempts using Dynamic Bayesian Networks tend towards the intractable [1, 12]. Probabilistic models are also used for certain types of discourse tracking [24, 33], and context free grammars can once more be used to model the hierarchical nature of discourse [43, 18]

5 Timeline

I plan to finish data collection and analysis by the end of summer '04, design, implementation and evaluation by the end of fall term of '04 and the thesis and

defence by the end of spring term ’05.

6 Resources

No resources beyond the standard PCs, microphones and networking facilities already available to me at the Media Laboratory are required.

References

[1] David W. Albrecht, Ingrid Zukerman, Ann E. Nicholson, and Arial Bud. Towards a bayesian model for keyhole plan recognition in large domains. In *User Modeling: Proceedings of the Sixth International Conference, UM97*, 1997.

[2] James Allen. *Natural Language Understanding*. The Benjamin/Cummings Publishing Company, Inc, Redwood City, CA, USA, 1995.

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

[4] Jon Barwise and John Perry. *Situations and Attitudes*. MIT Press, Cambridge, MA, 1983.

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

[6] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web seach engine. *Computer Networks and ISDN Systems*, 1998.

[7] M.K. Brown, B.M. Buntschuh, and J.G. Wilpon. SAM: A perceptive spoken language-understanding robot. *IEEE Transactions on Systems, Man and Cybernetics*, 6(22):1390–1402, Nov/Dec 1992.

[8] Hung H. Bui, Svetha Venkatesh, and Geoff West. Policy recognition in the abstract hidden markov model. *Journal of Artificial Intelligence Research*, 17:451–499, 2002.

[9] Josef Capek and Karel Capek. *Rossum’s Universal Robots*. Oxford University Press, 1920.

[10] David Chapman. *Vision, Instruction and Action*. MIT Press, Cambridge, MA, 1991.

[11] J.C. Chappelier, M Rajman, R. Aragües, and A. Rozenknop. Lattice parsing for speech recognition. In *Proceedings of the 6th annual conference Le Traitement Automatique des Langues Naturelles*, 1999.

[12] E. Charniak and R. Goldman. A bayesian model of plan recognition. *Artificial Intelligence*, 1993.

[13] Daniel Dennett. *Consciousness Explained*. Little, Brown and Company, 1992.

[14] A. Desolneux, L. Moisan, and J. Morel. A grouping principle and four applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 255(4):508–513, April 2003.

[15] Jay Earley. An efficient context-free parsing algorithm. *Communications of the ACM*, 6(8):451–455, 1970.

[16] E.A. Engbers and A.W.M. Smeulders. Design considerations for generic grouping in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 255(4):445–457, April 2003.

[17] Gareth Evans. *Varieties of Reference*. Oxford University Press, Oxford, UK, 1982.

[18] Katherine Frobes, Eleni Miltsakaki, Rashmi Prasad, Anoop Sarkar, Aravind Joshi, and Bonnie Webber. D-ltag system - discourse parsing with a lexicalized tree adjoining grammar. In *Proceedings of the ESSLLI 2001 Workshop on Information Structure, Discourse Structure and Discourse Semantics*, 2001.

[19] Allen L. Gorin, Guiseppe Riccardi, and J.H. Wright. How may i help you? *Speech Communication*, 1998.

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

[21] Barbara Grosz, Martha Pollack, and Sidner Candace. Discourse. In *Foundations of Cognitive Science*, pages 437–467. MIT Press, 1989.

[22] Rick Grush. Self, world and space: The meaning and mechanisms of ego- and allocentric spatial representation. *Brain and Mind*, 1:59–92, 2000.

[23] N.J. Haddock. Computational models of incremental semantic interpretation. *Language and Cognitive Processes*, 4:337–368, 1989.

[24] Eric Horvitz and Tim Paek. Deeplistener: Harnessing expected utility to guide clarification dialog. In *6th International Conference on Spoken Language Processing*, 2000.

[25] Dan Jurafsky, Chuck Wooters, Jonathan Segal, Andreas Stolcke, and Eric Fosler. Using a stochastic context-free grammar as a language model for speech recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 1995.

[26] Johan M. Lammens. *A computational model of color perception and color naming*. PhD thesis, State University of New York, 1994.

[27] B. Landau and R. Jackendoff. "what" and "where" in spatial language and spatial cognition. *Behavioural and Brain Sciences*, 2(16):217–238, 1993.

[28] Kai-Fu Lee. Context-dependent phonetic hidden markov models for speaker-independent continuous speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(4):599–609, 1990.

[29] George Miller and Philip Johnson-Laird. *Language and Perception*. Harvard University Press, 1976.

[30] Katashi Nagao and Jun Rekimoto. Ubiquitous talker: Spoken language interaction with real world objects. In *Proceeding of the International Joint Conference on Artificial Intelligence*, 1995.

[31] Srini Narayanan. *KARMA: Knowledge-based Action Representations for Metaphor and Aspect*. PhD thesis, Univesity of California, Berkeley, 1997.

[32] Sharon Oviatt, Phil Cohen, Lizhong Wu, John Vergo, Lisbeth Duncan, Bernhard Suhm, Josh Bers, Thomas Holzman, Terry Winograd, James Landay, Jim Larson, and David Ferro. Designing the user interface for multimodal speech and gesture applications: State-of-the-art systems and research directions. *Human Computer Interaction*, 15(4):263–322, August 2000.

[33] Tim Paek and Eric Horvitz. Dialog as action under uncertainty. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence, UAI2000*, 2000.

[34] Barbara H. Partee. Lexical semantics and compositionality. In Lila R. Gleitman and Mark Liberman, editors, *An Invitation to Cognitive Science: Language*, volume 1, chapter 11, pages 311–360. MIT Press, Cambridge, MA, 1995.

[35] James Pustejovsky. *The Generative Lexicon*. MIT Press, Cambridge, MA, USA, 1995.

[36] David V. Pynadath and Michael P. Wellman. Probabilistic state-dependent grammars for plan recognition. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence, UAI2000*. Morgan Kaufmann Publishers, 2000.

[37] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286, February 1989.

[38] Terry Regier. *The Human Semantic Potential*. MIT Press, 1996.

[39] Terry Regier and L. Carlson. Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General*, 130(2):273–298, 2001.

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

[41] Deb Roy, Peter J. Gorniak, Niloy Mukherjee, and Josh Juster. A trainable spoken language understanding system. In *Proceedings of the International Conference of Spoken Language Processing*, 2002.

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

[43] Remko Scha and Livia Polanyi. An augmented context free grammar for discourse. In *Proceedings of the Conference of the Association for Computational Linguistics*, 1988.

[44] Roger C. Schank and Robert P. Abelson. *Scripts, Plans, Goals and Understanding*. Lawrence Erlbaum Associates, 1977.

[45] William Schuler. Using model-theoretic semantic interpretation to guide statistical parsing and word recognition in a spoken language interface. In *Proceedings of the Association for Computational Linguistics*, 2003.

[46] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(22):888–905, August 2000.

[47] Jeffrey Mark Siskind. Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal of Artificial Intelligence Research*, 15:31–90, August 2001.

[48] Brian Cantwell Smith. *On the Origin of Objects*. MIT Press, Cambridge, MA, USA, 1996.

[49] Andreas Stolcke. An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, 21(2):165–201, 1995.

[50] Josh B. Tenenbaum. Bayesian modeling of human concept learning. In M. Kearns, S. Solla, and D. Cohn, editors, *Advances in Neural Information Processing Systems 11*, pages 59–68, 1999.

[51] M. Wertheimer. Laws of organization in perceptual forms. In *A source book of Gestalt psychology*, pages 71–88. Routledge, New York, 1999.

[52] Terry Winograd. *Procedures as a representation for data in a computer program for understanding natural language*. PhD thesis, Massachusetts Institute of Technology, 1970.