# Ontology-Based Discourse Understanding for a Persistent Meeting Assistant

**John Niekrasz** and **Matthew Purver** and **John Dowding** and **Stanley Peters**
Center for the Study of Language and Information, Stanford University
Stanford, CA 94305, USA
{niekrasz, mpurver, jdowding, peters}@csli.stanford.edu

**Abstract**

In this paper, we present research toward ontology-based understanding of discourse in meetings and describe an ontology of multimodal discourse designed for this purpose. We investigate its application in an integrated but modular architecture which uses semantically annotated knowledge of communicative meeting activity as well as discourse subject matter. We highlight how this approach assists in improving system performance over time and supports understanding in a changing and persistent environment. We also describe current and future plans for ontology-driven robust natural-language understanding in the presence of the highly ambiguous and errorful input typical of the meeting domain.

## Introduction

This paper describes current research efforts toward automatic understanding of multimodal, multi-party discourse for a persistent personal office assistant. The assistant aids users in performing office-related tasks such as coordinating schedules with other users, providing relevant information for completing tasks, making a record of meetings, and assisting in fulfilling the decisions made in the meetings. Our focus within this enterprise is on *automatic meeting understanding* – extracting detailed information about what was discussed, what the participants' conversational actions were, what decisions were reached, and the action items assigned. The assistant monitors meetings non-interactively, although the user may interact with the system between meetings to access the extracted information for use in their other activities.

The discourse understanding system we describe operates as a component of an enduring personal assistant known as the Cognitive Agent that Learns and Organizes (CALO)[1]. The overarching goals for the system include transforming the personal workspace into an environment of semantically unified information, enduring improvement through experience, and a robustness to unexpected events and missing knowledge in a changing environment. These goals pertain not only to the personal interactions users make with the system on their personal computer, but also apply to its presence as a ubiquitous agent in the meeting room. To this end, the results from understanding the meeting-room discourse must be semantically linked with the knowledge obtained and contained elsewhere, and the system must be designed to perform better with each new meeting, despite previously unknown knowledge being frequently introduced.

Even without these additional challenges, natural multi-party meetings pose several significant problems to automatic discourse understanding:

- The unconstrained nature of the spoken interaction (including significant speaker overlap, large vocabulary, unknown speaker identity and speaker location) as well as the nature of the acoustic environment leads to significantly increased error and reduced confidence in speech recognition output, producing an approximate 30% word error rate. This propagates significant error and ambiguity to the natural-language understanding components.
- The relatively unrestricted subject domain limits the utility of constrained lexicons and grammars for interpretation, techniques commonly relied on in spoken dialogue understanding. Interpretation routines require broad-coverage lexicons and grammar, and/or shallow processing techniques which rely less on detailed syntactic and semantic information. As a consequence, the system must be able to learn new words, concepts, and patterns of interaction online, expanding its interpretation capabilities according to the observed domain.
- The participants' natural use of multiple communicative modes means much of the discourse is unimodally ambiguous. Natural human discourse exhibits temporally and spatially linked verbal and physical behavior, and both physical and virtual objects such as charts, paper documents and slide presentations will be the common subject of conversational reference. Multimodal interpretation is therefore a fundamental requirement.
- Disparate types of information coming from a multitude of sensors, software agents, and human participants, over multiple communicative modes and physical media, creates a large-scale architectural and technical challenge. Information must be shared among components and meaningfully integrated into a common representation. This requires a semantic generalization and modularization of meeting knowledge as well as a functional



[1] http://www.ai.sri.com/project/CALO

architecture for making information accessible to independent as well as interdependent components.

Given the complexity of the task and the highly ambiguous and errorful component interpretations, we believe that a critical necessity to approaching these problems is to establish a flexible, unifying *multimodal discourse ontology* coupled with a generalizable framework for sharing knowledge hypotheses between components. While this is not a complete answer to the problems posed above, it provides an essential starting-point for tackling some of the more difficult problems of understanding in a persistent, dynamic and multimodal context.

First, we put this in a functional context by describing the physical and communicative environment in which the system and user are meant to interact. Following this, we present the multimodal discourse ontology which provides a system-wide semantics for the acquired knowledge. Next, we describe our temporal knowledge-base architecture designed to handle the sharing of information in the system. We then give an overview of our current approach to natural-language and discourse understanding in this context. Finally, we sketch out an account of how online learning may be achieved through persistent use of the system over many meetings.

## The User Environment

A typical meeting discourse of the kind supported by CALO will be a short (15–30 minute) meeting between approximately 3–8 participants. The participants may engage in any number of common meeting-room activities including short slide presentations describing ongoing work, the planning of project tasks and milestones, briefings about completed work, the making of important decisions, or assigning action-items for post-meeting fulfillment. These activities are likely to be realized in many communicative modes including the use of a whiteboard to draw project plans, explicit reference to elements in physical or virtual documents, and elementary verbal interaction.

Beyond the meeting room, the system is ubiquitous in the user's computing environment, handling email, calendar, and other components of a user's working information landscape. This forms a persistent and personal basis of interaction with the user, providing a critical level of information reliability that is not available from the non-interactive multi-party meeting domain. For this reason, we must use and produce knowledge which is framed in the semantics of the interactive system in order to make the results of meeting understanding less error-prone and useful to the user.

As a starting point for capturing the necessary information to begin to automatically understand meeting activities, the meeting room and users' laptops are instrumented with an array of sensing devices including close-talking microphones, laptop- and whiteboard- mounted stereo cameras, far-field microphone arrays, and electronic whiteboards. The system also makes use of a specially-designed 360-degree table-top video camera (Rybski *et al.* 2004b).

The multiple data streams coming from these physical sensors are in turn segmented, abstracted, and integrated into discrete, symbolic physical activities and gestures using a range of robust and adaptive tracking and recognition agents (Patil *et al.* 2004; Torre *et al.* 2005; Demirdjian & Darrell 2002; Ruddarraju, Haro, & Essa 2003). Automatic speech recognition is performed on close-talking audio channels using Carnegie Mellon University's Sphinx[2] recognition engine, and sketching and handwriting gestures are recognized using the Oregon Graduate Institute's Charter gesture recognizer. This initial layer of activity recognition forms the first level of semantically-annotated system knowledge.

The next step is to accumulate this knowledge into a discourse-relevant semantics and to tie it directly to the rest of the system knowledge. The next section describes the ontology we have constructed to perform this task.

## A Multimodal Discourse Ontology

Our approach to discourse understanding is centered around an ontology of multimodal discourse (MMD ontology). An ontology, as widely defined in the field of knowledge engineering, is an "explicit specification of a conceptualization" (Gruber 1993). It is used to concretely define a semantics for knowledge in the system. In our MMD ontology, the conceptualization is that of concepts and relationships in communicative actions performed during multimodal discourse: from the automatically sensed physical actions through to higher level interpretations of these actions and their relations to each other in discourse. The MMD ontology is a coalescence of a great number of distinct yet interwoven features of human discourse, many of which are derived from theories of language and communication, e.g. (Mann & Thompson 1988; Asher & Lascarides 2003; Kunz & Rittel 1970; Davidson 1980), as well as the technical capabilities of system components, each of which we describe later.

As a first step toward managing the complexity of the ontological framework, we must make broad distinctions between major components. The first distinction we make is that which separates discourse activity from discourse content or subject matter. The brief description given above of the MMD ontology's domain does not include any notion of what is being talked *about*, and this is entirely purposeful. As a basic requirement, our understanding process must of course have the means to understand discourse as being about *something*. However, it is not the responsibility of a general system like this one to define what the *something* might be. Rather, the MMD ontology defines a conceptual template for modeling discourse activity which can then be attached to a specific domain-dependent interpretation scheme and strategy. In the following paragraphs, we describe the elements of this generic template, reserving discussion of domain-dependent aspects and natural-language interpretation for a subsequent section. For a detailed account of the use and modularization of ontologies in dialogue systems, see (Flycht-Eriksson 2004).

[2] http://cmusphinx.sourceforge.net/html/cmusphinx.php

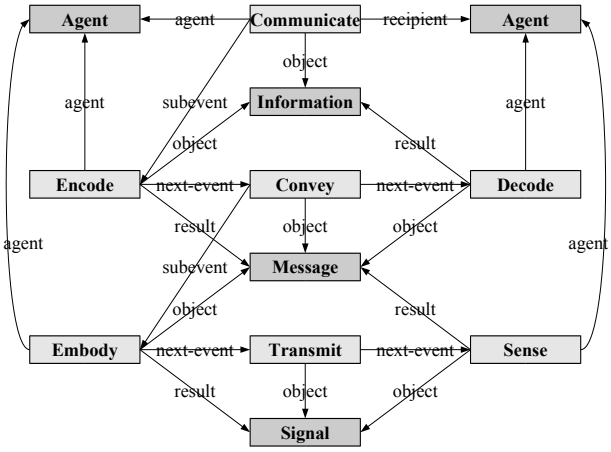


Figure 1: The CLib Communication Model

**The Core Ontology** The CALO system's numerous ontologies, of which the MMD ontology is only one, are designed around a core Component Library[3] (CLib) (Barker, Porter, & Clark 2001). The CLib is an assembly of both atomic and complex semantic modeling components, each representing a type of entity, event, role, or property. The CLib is encoded in the Knowledge Machine[4] knowledge representation language, with a subset being automatically transformable into the popular Web Ontology Language (OWL) of the Semantic Web[5] framework. The concepts which make up the MMD ontology are therefore derived from the instantiation and combination of CLib components. Each of these components are assigned a common-sense textual name which we refer to using `monospace` characters.

**The CLib Communication Model** The CLib contains a generic three-layer model of communication (hereafter the Communication Model), which we take as our starting point for the design of the MMD ontology. Figure 1 (courtesy of Ken Barker) shows its essential classes and relationships.

This multi-layer structure allows us to distinguish three essential component levels of the overall act of communication (the `Communicate` event) – essentially those of domain, linguistics and physics. At the top-most layer, the `Information` being communicated is rooted in the domain model, consisting of a statement about CLib classes or particular instantiations thereof. In the middle layer, that information becomes `Encoded` into some linguistic/symbolic `Message` in a particular language (e.g. spoken English, iconic gesture, or orthography). A `Message`'s elements may then carry any number of classes of features, including semantic, syntactic, diagrammatic, or phonological features. The bottom layer shows that the message must simultaneously be `Embodied` into some physically deliverable form (a perceivable `Signal`).

[3]http://www.cs.utexas.edu/users/mfkb/RKF/clib.html

[4]http://www.cs.utexas.edu/users/mfkb/RKF/km.html

[5]http://www.w3.org/2001/sw/

**Multimodality** We extend this model to include three fundamental modes of communication, distinguished by the physical `Medium` through which the communication is transmitted. The first is `Sound`, which includes both verbal and nonverbal spoken communication. The second is `Light`, through which visually perceived three-dimensional gestures (such as pointing and nodding) are communicated. The last is `Ink`, which carries typically two-dimensional diagrammatic and handwritten gestures.

The addition of multimodality to the Communication Model can be thought of as adding a third dimension to the diagram in Figure 1 at the linguistic and physical layers, with `Messages` and `Signals` also being distinguishable by mode. This is not required at the topmost layer, where overall meaning combines the various modalities and is anchored in the essentially a-modal domain model.

**Sensors and Physical Awareness** The sensors present in the meeting room are the system's only direct connection to the physical world, but they are also a part of that world and subject to perception themselves. For this reason, `Sensors` are modeled as `Spatial-Entities` in the environment, able to `Sense` signals in one of the previously described media. These `Sensors` produce `Recordings` which contain the lowest-level data available to the system: discretely-sampled continuous variables such as acoustic pressure levels, spatial coordinates for arm and head positions, and whiteboard pen locations.

This fine-grained physical-level data is not modeled directly by the MMD ontology, nor used directly by the discourse understanding system;[6] instead, it is used by dedicated physical awareness agents which recognize the parts of the recordings which might correspond to `Signals` carrying information-bearing `Messages` – i.e. the parts which might be relevant to discourse communication. These messages take the form of symbolic representations of physical events such as `Deictic-Pointing` gestures (Demirdjian & Darrell 2002), `Sitting` or `Standing` events (Patil *et al.* 2004), head-pose and gaze connections between participants (Ruddarraju, Haro, & Essa 2003) and hypothesized speech transcripts (or speech recognizer output lattices). It is this level of representation that is then combined with linguistic information to give a full model of communication (see below).

**Linguistics and Segmentation** Another important element of the MMD ontology is that which defines the relationship between the physical and symbolic levels of the Communication Model. Here, we establish the connection between perceivable events and the symbolic structural elements they encode. Such relationships are instantiated by communication in any mode, through the use of semi-mode-dependent linguistic, diagrammatic, or gestural languages.

To exemplify the interdependencies at this level, consider a simultaneously verbal and handwritten reference to an object in the domain of discourse, such as a project task with

[6]However, maintaining access to the raw `Recording` data is still important for post-meeting playback functionality.

the name "demo" (see (Kaiser 2005) for a detailed account of automated learning of vocabulary during such phenomena). The requirement of the model here is to create a mechanism for unifying not only the underlying referent, but also the common vocabulary which has been used to refer to it. The symbolic component of both `Messages` at use here is the `Word` "demo", which is realized both as `Orthographic-Units` and `Phonetic-Units`, depending on the mode-type of the `Embody` activity carried out. While the `Message` is modeled by a constituent structure containing constituent classes such as `Sentences`, `Phrases`, and `Words`, at the lowest level these are realized as mode-dependent temporal or spatial segmentations of `Sound`, `Light`, or `Ink` signals.

**Dialogue and Argumentation** Broadening our perspective of the Communication Model, we take individual `Communicate` acts to be atomic elements (dialogue moves) in our model of discourse. We then incorporate these moves within a model of dialogue history much like that of (Lemon & Gruenstein 2004): dialogue state is modeled as a tree, with individual moves forming the nodes, and the connections between nodes being the antecedent relation between the moves; separate branches of the tree are separate (although possibly simultaneous) conversational threads (sequences of antecedent-related moves). However, we depart from this model in two ways: firstly, by our incorporation of the various possible dialogue move types as subclasses in the MMD ontology and corresponding specification of their associated semantic and pragmatic constraints as ontological properties; and secondly, by the addition of a dimension of rhetorical and argumentative structure.

We classify dialogue moves along two nominally independent dimensions: their immediate short-term effect on the dialogue state, and their sometimes longer-term rhetorical or argumentative function. To represent the instantaneous state of the discourse, we add an `info-state` slot to the `Communicate` class, whose value can be seen as similar to the notion of *information state* of e.g. (Bohlin (Ljunglöf) *et al.* 1999).[7] This includes information about currently salient referents (for anaphora resolution) and currently relevant propositional information.

Specifically, the short-term effects are modeled using (Ginzburg forthcoming)'s Question-Under-Discussion (QUD) model: specific classes of move are seen as introducing or removing questions from a stack, represented as a `qud` slot in the current information state. This way, we can articulate within the MMD ontology itself the constraint that a `Query` move must introduce its question as the topmost (most salient) question in `qud`; or that a direct `Answer` move must express a proposition that can be unified with an antecedent `qud` question. Similarly, `qud` information from the antecedent move can be used for the resolution of elliptical fragments.

To our model of dialogue moves, we add a notion of both rhetorical and argumentative structure (see e.g. (Asher & Lascarides 2003; Mann & Thompson 1988)), as one of our primary interests is to model decision-making and its outcomes, such as assigned tasks (a.k.a. action-items). In the model, dialogue moves may also be classified according to their argumentative function, via their relation to an `iun` slot in the information state, implementing a version of (Larsson 2002)'s Issue-Under-Negotiation model. Argumentative threads are seen as pertaining to particular `Issues`, modeled as questions on the `iun` stack (e.g. `Introduce` moves introduce new issues, `Proposals` introduce possible alternative answers thereto, `Acceptances` or `Rejections` remove those alternatives). Again, these effects and/or preconditions on the move types are expressed directly as properties of their subclasses in the ontological model. Currently, we treat the two notions of discourse structure mentioned above as independent dimensions of the dialogue tree: e.g. a dialogical `Answer` might function rhetorically as `Proposal`, `Rejection`, `Acceptance` or others.

**Collaborative Behavior and Negotiation** Beyond a move-by-move account, however, meetings exhibit longer-term negotiative and argumentative patterns which present an extremely difficult challenge to automatic understanding. Due to their psychological roots but unclearly-defined semantics, an account of meeting structure at this level is both extremely difficult yet extremely useful. In support of this essential (and perhaps ultimate) goal for the understanding system, we specify a model to capture the semantics of these long-term negotiative structures.

We use a model based on the Issue-Based Information System (IBIS) put forth in (Kunz & Rittel 1970) and exemplified in systems such as Compendium (Bachler *et al.* 2003) and techniques such as *Dialogue Mapping* (Conklin *et al.* 2001). These models are critical for deriving meaningful user-level structure from the discourse, turning the meeting into a useful shared-memory resource. We derive our conceptualization in great part from the AKT reference ontology[8] and the meeting-oriented additions made in (Bachler *et al.* 2003). These include notions of meeting `Artifacts` – physical or virtual information-bearing documents – and long-term negotiative behaviors around them, such as the reading of an `Agenda`, assigning `Action-Items`, and following up on `Decisions`. These objects are instantiated through composition of the rhetorical and argumentative structures described above.

**Topics and Discourse Phases** Finally, at the most abstract level, we take a single meeting to be subdivided into `Discourses`, representing its major phases or topics (examples of distinct `Discourses` might be `Discussions` on separate agenda topics, or slide `Presentations`). There are assumed to be no discourse-level connections between these phases; within them, all communicative acts are seen as being interconnected to some degree. A `Discourse` can therefore be represented as a dialogue

[7] Models of information state usually incorporate a history of dialogue moves as well as the records we describe here – in our model, this is available directly from the dialogue move tree itself.

[8] http://www.aktors.org/publications/ontology/

move tree with a single root node: all moves within this discourse must be part of this tree.

Given the expected low level of accuracy of speech recognition and therefore parsing, we recognize that the segmentation of the meeting into `Discourses` is unlikely to be achievable by recognizing relations between moves from their semantics (or other internal properties) alone. We therefore pursue automatic `Discourse` segmentation based on shallower features of the discourse such as global lexical cohesion and speaker activity changes (following e.g. (Galley *et al.* 2003)) as well as gesture and other multimodal activities (Banerjee & Rudnicky 2004; Rybski *et al.* 2004a), and we model these correspondingly in the MMD ontology. This adds further constraints to the possible hypotheses of the pragmatic integration agent (see below) by requiring moves within these `Discourse` spans to be related (backing off to relating a move to the root node if no other relation can be inferred).

## *KronoBase*: A Temporal Knowledge Base

In addition to the ontology specification described in the previous section, we have developed a persistent temporal knowledge base system called *KronoBase*, which is used for the exchange of information gathered from the perceptual and interpretive activities performed by the components during the meeting. It supports use of the OWL form of the MMD-inclusive CLib ontology, allowing assertions to be made on a single statement-by-statement basis, and querying through the use of the RDQL[9] query language, supported by the Jena Semantic Web framework API[10].

The role of *KronoBase* is both as a repository of knowledge collected by the component agents and as a manager of meta-information about the knowledge itself. We expect the majority of asserted knowledge to be speculative or incomplete (as produced from the viewpoint of individual components). *KronoBase* maintains this speculative information in the form of probabilities and underspecified logical structures, allowing later learning via reinforcement or supplementary information. In addition, it maintains reference to the source and time of the assertion and the context in which it was asserted, thus enabling access to a complete history of the knowledge state. This results in a generic framework for persistent, collaborative interpretation.

## Discourse Interpretation

Having specified both a generic template ontology for modeling multimodal discourse and an architectural mechanism for sharing this information persistently between components, we now must address the problem of performing discourse interpretation, with the ultimate goal of generating a semantic analysis of the discourse subject matter that is compatible with CALO's central domain ontology (an exhaustive model of concepts pertaining to the user's office environment) and useful to the user in their interaction with the system between meetings. In this section, we describe some of our proposed techniques for doing this in the context of highly ambiguous and errorful input.

### Natural Language Processing

In the meeting environment, the relatively free subject domain prevents the use of a constrained grammar for semantic interpretation. Instead, we use a general broad-coverage grammar (based on generic domain-independent lexical resources) to perform deep parsing where possible, and back off to shallow chunk parsing otherwise. Both functions are performed using the Gemini parser (Dowding *et al.* 1993). Following (Swift 2005), we build a large noun lexicon using Comlex (Grishman, Macleod, & Meyers 1994) to provide syntactic information, and we use WordNet (Fellbaum 1998) to provide corresponding semantic class information. VerbNet (Kipper, Dang, & Palmer 2000) is then used to provide syntactic and semantic information for verbal predicates, including semantic selectional restrictions for their arguments. Once some modifications have been made to link the WordNet noun class hierarchy with both the hierarchy used in VerbNet (EuroWordNet – (Vossen 1997)) and the CLib ontology, this provides us with an overall broad-coverage grammar; and importantly, the parser output includes semantic logical forms whose sortal information is directly related to the CLib ontological classes (see (Dzikovska, Swift, & Allen 2004) for a related approach).

To outline this process in greater detail, consider a transcription posited by a speech recognizer agent as an annotation of a `Signal` in the knowledge-base. The parser will attempt to produce a corresponding logical form (LF) (in fact, usually several alternative hypotheses corresponding to ambiguities in both parsing and speech recognition), asserting this as a `Message` instance (related to the signal via its `Transmit` event). The possible top-level `Communicate` event is not posited at this stage, but left to the pragmatic integration agent, where other modalities are available (see below).

The logical form representation we mention above uses a Davidsonian event-based semantics (Davidson 1980) with the thematic roles defined in VerbNet; this allows the output of successful full sentential parses and that of shallow chunk parsing to be compatible, and allows a straightforward translation into an ontological representation. Quantifier scope is left underspecified by use of a QLF representation (see (Alshawi 1992)). A full parse of a sentence "Move the milestone to April" would be given a representation as follows:

$$\begin{aligned}&\exists e.[move(e)\\&\quad\wedge Agent(e, aterm(addressee))\\&\quad\wedge Theme(e, qterm(the, x, milestone(x)))\\&\quad\wedge Destination(e, aterm(april))]\end{aligned}$$

This event-based representation can be translated directly to an instance of the CLib ontological class `Move` with its associated slots – again, this step is not taken here, but as part of multimodal integration as described below. In cases where unknown words are present in the input, and a full predicate-argument structure cannot be created, the system

[9] http://www.w3.org/Submission/RDQL/

[10] http://jena.sourceforge.net/

backs off to producing fragments whose role relation to the predicate is unknown:

$$\{\exists e.[move(e)],\\ qterm(the, x, milestone(x)),\\ aterm(april)\}$$

We can now hypothesize possible methods of combination for these fragments in a process of semantic construction governed by the lexical and domain ontologies, following e.g. (Milward & Beveridge 2004; Ludwig, Bücher, & Görz 2002): the same constraints used in the full sentence grammar can determine that $milestone$ and $april$ are of suitable semantic classes to play the $Theme$ and $Destination$ roles of a $move$ event. Missing arguments can be left underspecified (as with the $Agent$ role here, and in fact the $Source$ role in both examples), as can the roles played by available fragment arguments where the event itself is unknown or uncertain.

This underspecified and possibly ambiguous LF is then passed to a discourse integration module for further disambiguation and pragmatic interpretation (including both dialogue move type determination and referential interpretation at the level of the domain model), as described below. Note that until this point, the central use of the ontologies has allowed the natural-language understanding component to be to a large degree domain-independent: lexical entries, names, concepts and their combinatoric possibilities are all specified within the lexical ontologies.

### Multimodal Fusion & Pragmatic Interpretation

These underspecified LFs can now be combined with the other available sources of information: firstly, the parallel semantic representations produced by understanding agents working in other modalities; secondly, the discourse model (i.e. the MMD ontology) and the discourse history (i.e. the current state of the knowledge-base); and finally the domain ontology itself. The constraints provided by the domain and MMD ontologies allow us to examine possible combinations of these information sources while checking for consistency, both move-internal (semantic) and move-external (pragmatic).

Given an example such as the partial fragment interpretation above, this integration can help fully specify the propositional information being communicated. The linguistic information gives us an instance of the `Move` class and tells us that the $Theme$ thereof is of type `Milestone`, but leaves its identity underspecified; however, a simultaneous pointing gesture to a point (of the `Milestone` class) on a projected project diagram being can supply us with this. The same might be true for the $Source$ and $Destination$ roles. Similarly, while the $Agent$ role is specified to be the addressee, the actual identity may not be known – simultaneous eye gaze (to an entity of the correct `Person` class) may provide it.

The discourse model will also provide constraints: recognition of this move as being of type `Command` will be associated with constraints both on the roles (e.g. that the agent be the addressee – the `recipient` of the overall `Communicate` event – as already hypothesized by the grammar) and on the semantic content (that the semantic LF be an imperative). In future, it may even be possible to rule out certain semantic interpretations via consistency with the context (e.g. with a command, checking that the commanded action has not already been carried out), but this is currently beyond *KronoBase*'s inference capabilities.

Note that gesture integration must be to a certain degree domain-dependent: while certain general principles will be available (e.g. that the object of a pointing gesture can be hypothesized to fill a thematic role which is unifiable with whatever constraints are specified by the semantic LF; that the object of a `EyeGaze` event can be hypothesized to fill the addressee role) to domain-specific rules associated with particular activities (e.g. leftward gestures might be interpreted as conveying temporal motion backward in the project chart domain).

## Learning

This principled approach to semantic/pragmatic representation and multimodal integration allow us to combine knowledge sources to enable the system to learn from experience. There are two main ways in which this can occur: firstly, use of information from one modality to inform another; and secondly, use of context and dialogue history to inform the understanding agents.

**Multimodal Learning** The use of the ontology as a central unified semantic representation can allow learning directly. An out-of-vocabulary name can provide only an underspecified semantic representation; but a simultaneous deictic pointing gesture can provide the necessary reference. Unification of the two during the pragmatic integration of the overall dialogue move effectively provides the hypothesis that the name refers to the indicated entity – if this is fed back to the lexicon, ontology, and speech recognizer, new entries may be created that allow the name to be correctly parsed and resolved in future (see (Kaiser 2005; Kaiser *et al.* 2004)).

**Experiential Learning** New words, names, concepts and facts can also be learned via the history of the communicative context. If understanding routines posit underspecified representations for unknown (or uncertain) items, any subsequent successful pragmatic integration will provide a certain degree of further specification, as the constraints of the MMD and domain ontologies are applied. As more instances of these unknown items appear and are integrated into the knowledge base, the partial information will be further and further constrained, filling in more detail (in frame-based ontological terms, moving from superclass to subclass), and thus allowing gradual learning over time (with improved understanding as the new information is provided to the understanding agents). This can of course incorporate cross-modal feeding – detection of a new face might cause a highly underspecified representation of the new individual to be introduced; subsequent discussion of or addressing to an object called "John" (perhaps accompanied pointing or

eye gaze) can add new assertions associating the name with the person.

Note that this loose specification of new entries, followed by subsequent further specification by experience is made possible not only by the system's persistence between meetings, but also by its non-interactivity during a meeting: as there is no requirement to act on or respond to each human utterance immediately, understanding can be temporarily underspecified until resolved (or strengthened beyond a certain probabilistic threshold) by subsequent discourse.

**Feedback** The temporal capabilities of *KronoBase*, together with its persistence between meetings, enable post-meeting interaction which can provide not only useful functionality but feedback to allow the system to learn further. We are developing a question-answering dialogue system *Meeting Reviewer* to allow a user to query information about the meeting history itself: not only what decisions were made and when, but who made them, who (dis)agreed with them, and whether they were later modified. Allowing the user to interact with and correct the system if answers are wrong can directly provide it with critical feedback which can strengthen both of the above approaches to adaptive learning.

## Future Work

With the current MMD ontology and ontology-aligned interpretation components, we must now begin investigating the application of our model to specific meeting phenomena for further understanding and learning. In particular, we hope to apply persistent distributed access to discourse information in ways which would be very difficult without it, such as using discourse structure to constrain and improve speech recognition, or to perform online interpretation in interactive discourse. The design of the MMD ontology has made possible the generic application of learning and interpretation algorithms over the knowledge base. Using corpora of ontologically-annotated meetings, and with our unified representational scheme, we hope to find and learn large-scale patterns of communication. In addition, we are also looking to develop a more efficient inter-component communication architecture around the *KronoBase* system to accommodate the large amounts of data being generated, as well as an enhancement of its reasoning capabilities (perhaps through direct integration with CLib's native Knowledge Machine language). We are also currently investigating techniques for user-level annotation and evaluation of the system.

## Summary

In this paper we have presented research toward ontology-based understanding of discourse in meetings, and have described our current implementation of this strategy. In particular, we have pointed out the necessity for an integrated but modular approach which uses semantically annotated knowledge of communicative meeting activity as well as discourse subject matter. We have described current and future plans for robust natural-language understanding in the presence of highly ambiguous and errorful input. We have also described how this approach assists in improving system performance over time and supports understanding in a changing environment.

## Acknowledgments

The authors would like to thank the numerous researchers and engineers involved in making this research possible, and who have supported the integration of their components with the MMD ontology, including but not limited to Satanjeev Banerjee, Ken Barker, David Demirdjian, Alex Gruenstein, Bill Jarrold, Ed Kaiser, Sanjeev Kumar, Vincenzo Pallotta, Paul Rybski, Ravi Ruddarraju, Yitao Sun, Mary Swift, Lynn Voss, and Regis Vincent.

This work was supported by DARPA grant NBCH-D-03-0010. The content of the information in this publication does not necessarily reflect the position or the policy of the US Government, and no official endorsement should be inferred.

## References

Alshawi, H., ed. 1992. *The Core Language Engine*. Cambridge, MA: MIT Press.

Asher, N., and Lascarides, A. 2003. *Logics of Conversation*. Cambridge University Press.

Bachler, M.; Buckingham Shum, S.; De Roure, D.; Michaelides, D.; and Page, K. 2003. Ontological mediation of meeting structure: Argumentation, annotation, and navigation. In *1st International Workshop on Hypermedia and the Semantic Web*.

Banerjee, S., and Rudnicky, A. 2004. Using simple speech-based features to detect the state of a meeting and the roles of the meeting participants. In *Proceedings of the 8th International Conference on Spoken Language Processing (Interspeech 2004 - ICSLP)*.

Barker, K.; Porter, B.; and Clark, P. 2001. A library of generic concepts for composing knowledge bases. In *Proceedings of the First International Conference on Knowledge Capture*.

Bohlin (Ljunglöf), P.; Cooper, R.; Engdahl, E.; and Larsson, S. 1999. Information states and dialogue move engines. In Alexandersson, J., ed., *IJCAI-99 Workshop on Knowledge and Reasoning in Practical Dialogue Systems*.

Conklin, J.; Selvin, A.; Shum, S. B.; and Sierhuis, M. 2001. Facilitated hypertext for collective sensemaking: 15 years on from gibis. In *HYPERTEXT '01: Proceedings of the twelfth ACM conference on Hypertext and Hypermedia*, 123–124. ACM Press.

Davidson, D. 1980. *Essays on Actions and Events*. Oxford: Clarendon Press.

Demirdjian, D., and Darrell, T. 2002. 3-d articulated pose tracking for untethered diectic reference. In *Proceedings of ICMI02*.

Dowding, J.; Gawron, J.; Appelt, D.; Bear, J.; Cherny, L.; Moore, R.; and Moran, D. 1993. Gemini: a natural language system for spoken-language understanding. In *Proc. ACL 93*.

Dzikovska, M.; Swift, M.; and Allen, J. 2004. Building a computational lexicon and ontology with FrameNet. In *Proceedings of the LREC Workshop on Building Lexical Resources from Semantically Annotated Copora*.

Fellbaum, C., ed. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Flycht-Eriksson, A. 2004. *Design and Use of Ontologies in Information-providing Dialogue Systems*. Ph.D. Dissertation, School of Engineering at Linkping University. Thesis No. 874.

Galley, M.; McKeown, K.; Fosler-Lussier, E.; and Jing, H. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03)*.

Ginzburg, J. forthcoming. *A Semantics for Interaction in Dialogue*. CSLI Publications. Draft chapters available from: `http://www.dcs.kcl.ac.uk/staff/ginzburg`.

Grishman, R.; Macleod, C.; and Meyers, A. 1994. COMLEX syntax: Building a computational lexicon. In *Proceedings of COLING 94*.

Gruber, T. 1993. A translation approach to portable ontologies. *Knowledge Acquisition* 5(2):199–220.

Kaiser, E.; Demirdjian, D.; Gruenstein, A.; Li, X.; Niekrasz, J.; Wesson, M.; and Kumar, S. 2004. A multimodal learning interface for sketch, speak and point creation of a schedule chart. In *Proceedings of the 6th international conference on Multimodal interfaces*, 329–330. ACM Press.

Kaiser, E. C. 2005. Multimodal new vocabulary recognition through speech and handwriting in a whiteboard scheduling application. In *Proceedings of the International Conference on Intelligent User Interfaces*, pp. 51–58.

Kipper, K.; Dang, H. T.; and Palmer, M. 2000. Class-based construction of a verb lexicon. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000)*.

Kunz, W., and Rittel, H. W. J. 1970. Issues as elements of information systems. Technical Report WP-131, University of California, Berkeley.

Larsson, S. 2002. *Issue-based Dialogue Management*. Ph.D. Dissertation, Göteborg University. Also published as Gothenburg Monographs in Linguistics 21.

Lemon, O., and Gruenstein, A. 2004. Multithreaded context for robust conversational interfaces: Context-sensitive speech recognition and interpretation of corrective fragments. *ACM Transactions on Computer-Human Interaction (ACM TOCHI)* 11(3). (to appear).

Ludwig, B.; Bücher, K.; and Görz, G. 2002. Corega tabs: Mapping semantics onto pragmatics. In Görz, G.; Haarslev, V.; Lutz, C.; and Möller, R., eds., *Proceedings of the KI-2002 Workshop on Applications of Description Logics*.

Mann, W., and Thompson, S. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text* 8(3):243–281.

Milward, D., and Beveridge, M. 2004. Ontologies and the structure of dialogue. In Ginzburg, J., and Vallduví, E., eds., *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue (Catalog)*, 69–77.

Patil, R.; Rybski, P. E.; Kanade, T.; and Veloso, M. 2004. People detection and tracking in high resolution panoramic video mosaic. In *Proceedings of IROS'04, the IEEE International Conference on Intelligent Robots and Systems*, pp. 1323–1328.

Ruddarraju, R.; Haro, A.; and Essa, I. 2003. Fast multiple camerahead pose tracking. In *Proceedings 16th International Conference on Vision Interface*.

Rybski, P. E.; Banerjee, S.; de la Torre, F.; Vallespi, C.; Rudnicky, A.; and Veloso, M. 2004a. Segmentation and classification of meetings using multiple information streams. In *Proceedings of the Sixth International Conference on Multimodal Interfaces*.

Rybski, P. E.; de la Torre, F.; Patil, R.; Vallespi, C.; Veloso, M.; and Browning, B. 2004b. CAMEO: Camera Assisted Meeting Event Observer. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'04)*, pp. 1777–1782.

Swift, M. 2005. Towards automatic verb acquisition from verbnet for spoken dialog processing. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*.

Torre, F. D. L.; Vallespi, C.; Rybski, P. E.; Veloso, M.; and Kanade, T. 2005. Learning to track multiple people in omnidirectional video. In *Proceedings of ICRA'05, the IEEE International Conference on Robotics and Automation, 2005*.

Vossen, P. 1997. EuroWordNet: a multilingual database for information retrieval. In *Proceedings of the DELOS workshop on Cross-language Information Retrieval*.