A Multimodal Discourse Ontology for Meeting Understanding

John Niekrasz and Matthew Purver

Center for the Study of Language and Information, Stanford University
{niekrasz,mpurver}@csli.stanford.edu

Abstract. In this paper, we present a multimodal discourse ontology that serves as a knowledge representation and annotation framework for the discourse understanding component of an artificial personal office assistant. The ontology models components of natural language, multimodal communication, multi-party dialogue structure, meeting structure, and the physical and temporal aspects of human communication. We compare our models to those from the research literature and from similar applications. We also highlight some algorithms that are used to perform automatic processing and understanding using these models and suggest elements of the ontology that may be of immediate interest to meeting annotation by human or automated means.

1 Introduction

People can communicate with great efficiency and expressiveness during natural interaction with others. This is perhaps the greatest reason that face-to-face conversations remain such a significant part of our working lives despite the numerous technologies available that allow communication by other means. Nevertheless, businesses spend millions of dollars each year conducting meetings that are often seen as highly inefficient [1], and there is great interest in researching these interactions to better understand them, create technology to facilitate them, and assist in the recording and dissemination of their content.

To do this in a manner that is truly useful to organizations and desirable to individuals, automated "meeting understanding" should encompass not only the annotation of video and audio for playback, but the extraction of relevant information at the level of semantics and pragmatics: what subjects were discussed, what decisions were made, and what tasks were assigned [2]. Because natural multi-party interactions are vastly complex, and because this information we wish to extract is equally complex, of many different types, and expressed in many different modalities, a meeting understanding system must have an *integrated* and *expressive* model of meetings, discourse, and language supporting it to effectively manage its knowledge.

For our meeting understanding system, a component of the Cognitive Assistant that Learns and Organizes (CALO), knowledge integration and expression is performed through the use of a formal ontology. Our work in the design of

this ontology parallels that which has has been termed "meeting modelling" [3], "meeting ontology" [4], or "meeting data model" [5] elsewhere in the literature. While other efforts of this kind are similar in purpose, to our knowledge, our ontology is the only implementation that (1) integrates such a wide variety of components, (2) is directly linked to a domain of understanding, and (3) uses an expressive semantics for representation and inference.

In the following sections, we present our multimodal discourse ontology (henceforth, MMDO) and describe its purpose in the CALO system. Section 2 provides a clearer problem definition in relation to similar research. In Sect. 3, we describe the ontology itself in detail. Finally, in Sect. 4, we present some of the current and potential functional uses of the ontology in performing automatic understanding and annotation.

2 Background

There are currently multiple efforts being undertaken to create systems that observe, organize, facilitate, or otherwise understand meetings automatically. Each effort has brought forth distinct proposals for models of meetings and their associated data. Many commonalities may be found between these models, while in some cases, differing motivations and requirements have caused new approaches to be taken.

One nearly universal motivation is the support of user-level applications. [5] proposes a model for meetings and meeting data intended for a meeting browsing web tool; [3] describes a generic model for corpus-based multimodal interaction research supporting remote conferencing and virtual simulation; [4] describes an ontology of collaborative spaces and activities for meeting argumentation structuring, navigation, and replay. Our ontology is designed similarly to support user-level applications including a meeting browser with search, summary, and playback capabilities and a proactive assistant for relevant document retrieval during the meeting. Additionally, system testing will be carried out via a set of user-level queries, encoded using the ontology and based on common user-level requirements, similar to those obtained in user studies such as [6] and [2].

In addition, the MMDO is also designed to facilitate inter-process communication within an adaptive automatic discourse and natural-language understanding architecture, which requires the modelling of concepts that may *not* play a role for the user. Any information generated by individual components, e.g. the speech recognizer or natural language parser, must be specified in the model in order to be communicated system-wide, increasing the ontology's complexity and requiring that it take into account constraints imposed by the functioning of system components.

The MMDO is also closely linked with other ontologies that support CALO's other functions, such as event calendaring, email and contact management, and task monitoring. These concepts and knowledge about them are the very subject matter of the meetings we wish to automatically understand, requiring our ontology to elegantly connect to representations of discourse subject matter.

Another driving factor in our design is the system's upper ontology. All ontologies in the CALO system are designed using the Component Library (CLib) ontology [7], a library of generic atomic and complex concepts, each representing a type of entity, event, role, or property. While we will not describe implementation specifics in this paper, the reader should be aware that CLib and CALO's component ontologies, including the MMDO, are implemented by the CLib maintainers in the Knowledge Machine language [8], an expressive frame-based knowledge representation language with first-order logic semantics.

Our design of the MMDO, following the motivations presented above (see [9] for comparable set of motivations in the design of a dialogue act taxonomy), is meant to remain flexible and generic. In many cases models are purposefully underspecified to support further theory development. In others, system requirements have prompted full specification of models that may change to accommodate a more generic architecture. We will now turn to describing the core ontology that is a foundation for the MMDO.

2.1 Upper Ontology

The CLib [7] serves as the CALO system's upper ontology. Its components are designed to be reusable and composable by non-experts and therefore take inspiration from natural language, causing its concepts to remain relatively intuitive to users. The principal division in the library is between *Entities* (things that are) and *Events* (things that happen). Events are divided into *States* and *Actions*, where states are relatively static and brought about or changed by actions. In addition, a *Role* is something an entity *is* in the context of an event. Composition is then achieved through the use of *relations* between components and *properties*. Every concept in the MMDO described below is designed through composition and relation to these and other previously defined components.

2.2 The CLib Communication Model

The CLib ontology includes a Communication Model (CM), a model of communication and knowledge exchange between agents. It includes three layers, representing the physical, symbolic, and informational components of individual communicative acts (the *Communicate* event); the events in these three layers typically occur simultaneously, transforming the communicated domain-level *Information* into an encoded symbolic *Message*, from this message into a concrete physical *Signal*, and back again (see Fig. 1, where dashed lines divide the layers). *Events* are depicted as ovals and *Entities* are depicted as darker rectangles. The arrows signify *relations*. The three layers may be interpreted as aligning with the layers of joint action described in [10] at which communicative grounding takes place. To complete the first layer, there must be *attention*; for the second, *identification*; and for the third, *understanding*.

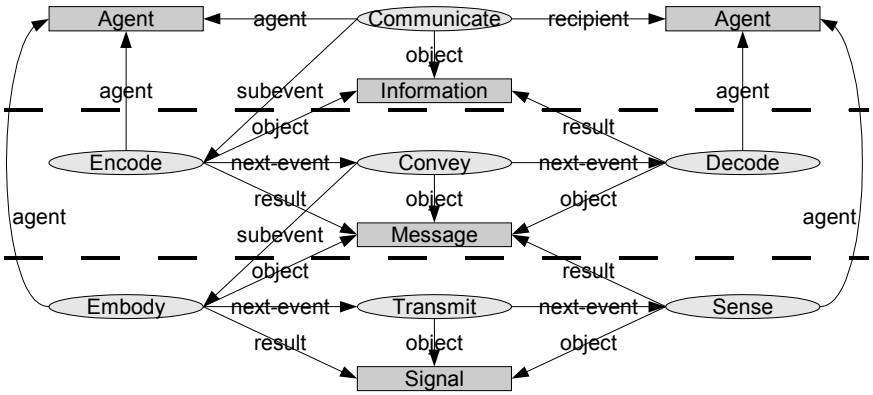

Fig. 1. The CLib Communication Model

As a foundation for further development of the MMDO, we posit a functional interpretation of the *Communicate* event that is appropriate for multi-party human dialogue. Namely, the *Communicate* event is taken to serve the role of an atomic *dialogue move*: a temporally contiguous communicative action with a possible interpretation and contextual significance, along the lines of what may be called a *speech act*, *communicative act*, *dialogue act*, or *conversational act/move* in the literature. Its role in the ontology will serve as its formal definition.

3 The Multimodal Discourse Ontology

We now turn to describing the details of the MMDO. We present the ontology in three parts proceeding conceptually from local to global elements. First, Sect. 3.1 describes extensions to the CM required to apply its internal model of communicative acts to natural multimodal communication; Sect. 3.2 then goes beyond these internals to describe the discourse model that connects communicative acts together and that defines their relationship to individual participants in a multi-party discourse; finally, Sect. 3.3 describes our model of the meeting activity and its relationship to the participants, the discourse, and the meeting environment.

3.1 Extending the Communication Model

At the level of individual communicative acts, our model uses the CM as a starting point, but requires several extensions to take into account both the constituent structure of natural language and the multimodal, multi-party nature of meeting dialogue.

Multimodal Communication. The basic CM assumes a one-to-one mapping across its three layers, neglecting the multimodal co-expression of speech

and gesture that is found in natural conversation [11] (e.g. simultaneous verbal and gestural reference, as in "Can you pass me that [point] cup please?"). To model multimodal communication, we extend this to multiple media via the CLib concepts of *Medium* and *Language*, where a *Signal* must be transmitted over some *Medium* and a *Message* must be encoded in some *Language*. For a single *Communicate* event, we now allow the *Encode* action to produce a multiplicity of *Messages*, each in their own *Language*, which each generate their own physical signal in some *Medium*. Speech is characterized as employing any *SpokenLanguage* such as *SpokenEnglish* and the medium of *Sound*; writing of text employs a *WrittenLanguage* such as *WrittenEnglish* and the medium of *Ink*; natural human gesture employs the language of *HumanGesture* and the medium of *Light*.

Additionally, the basic association of physical-layer events with various media are encoded as definitional axioms for subclasses of the *Embody* event such as *Speak*, *Draw*, and *Gesticulate* (*Hear*, *Read*, and *See* are encoded as subclasses of *Sense* for the sensory half of the model). By asserting these latter physical-layer events independent from the symbolic or informational layers, they may optionally serve to represent events like coughing or accidental ink-marks that are produced in the appropriate mode but determined to be without linguistic or communicative function.

Constituent Structure. Despite our addition of a dimension supporting multimodality to the CM, there remains a single symbolic entity (a *Message*) between the physical signal and the domain interpretation for each mode. In extending our model to natural language, and in particular when providing a basis for automatic NL processing, we of course require a more complex representation which includes not only the multiple layers of utterance representation in the CM but also their internal constituent structure (representations of individual words and phrases within utterances). While keeping to the CM model, we therefore take *Messages* as our equivalent of *signs*, with lexical, syntactic, semantic, phonological, or semaphoric (gestural) representations expressed as properties thereof.

Our framework follows that of the General Ontology for Linguistic Description [12], positing a recursively-defined *LinguisticUnit*, which is the building-block of *Messages* and is a *Message* itself. Units can then be built into constructions through composition, generating a *LinguisticConstruction* (a collection of units forming its own unit), a *LinguisticConstituent* (one of two or more units that form a construction), and a *LinguisticAtom* (a unit that is not a construction). These generic classes are realized through medium- and language-specific subclasses, allowing information in all modalities to be expressed in the same framework. For written and spoken language, these specific subclasses include *Word* and *Sentence*, together with sub-lexical units such as spoken *Phonemes* and written *OrthographicUnits*. For graphical representations such as whiteboard diagrams, they include atomic and compound *DiagramObjects*.

For gestural communication, they include units such as *DeicticGesture* and *IconicGesture*, modelling the set of gestures termed "semaphoric" in [11].

Physical Embodiment and Signal Segmentation. If we are to be able to replay particular constituents for analysis, or to train processing components (e.g. speech recognizers) based on their observed realizations, this linguistic constituent structure must be linked to a parallel structure in the layer of physical signals, and we therefore elaborate the CM one step further. We take the *Embody* event to be composed of *subevents* that realize the individual constituents of the *Message*, resulting in temporal sub-constituents of the overall *Signal*. This provides us with an event-based (temporal) representation for the physical realization of linguistic constituents, allowing a representation for language-based signal segmentation of audio, ink, and video, a common task and important requirement for linguistic and multimedia annotation (see [13] for a discussion).

Semantics. In the case of gestural acts such as *DeicticPoint*, knowledge of its referent is enough to fully characterize the *Information* component in the communicative model. Units of natural language, however, are semantically more complex and need to be annotated for meaning at their multiple constituent levels. In the MMDO, this is handled by each linguistic constituent (including the *Message* as a whole) potentially having a *logical-form* component, allowing us to express not only the propositional content of the constituent, but also the referential content of individual words and phrases where suitable. This component may be expressed in the semantics of the CLib ontology and its component domain ontologies, allowing direct linking to the system's knowledge base. Additionally, given the high levels of noise due to the speech recognizer errors and ungrammatical speech that are prevalent in multi-party dialogue, full propositional semantic annotation will usually not be possible for the highest-level *Message*, but by taking a robust fragment-parsing approach within a Davidsonian semantics [14], this representation allows us to posit event, entity, and role representations wherever possible, while leaving other entities or roles underspecified.

Communicative Roles. The basic CM contains a simple representation for the relations that individuals have to a communicative act. They are either the *recipient* or *agent* of the events in the model. For natural multi-party conversation, this is overly simplistic. People may be overhearers of acts even though they are not the direct addressees; and the intended addressee of an utterance or gesture may be the entire group (e.g. lecturing), a subset (e.g. third-party talk), or an individual. The basic model will therefore not support algorithms for addressee detection (and subsequently turn-taking and initiative management in an interactive system). We therefore add *Addressee* and *Overhearer* to the set of *Roles* that a *Person* may play in a *Communicate* event.

3.2 Modeling Discourse Structure

The extensions described so far are restricted to individual communicative acts. This section describes further extensions that allow us to express relations between these acts, providing an integrated model of a *Discourse* event and its structure.

Dialogue Structure. Our notion of discourse structure is expressed by considering individual *Communicate* events as *dialogue moves*, expressed via membership of particular subclasses and with their interrelation expressed via the properties associated with these subclasses. Following e.g. [15], we class moves at more than one nominally independent layer. At the most fundamental level, we consider only a move's effect on the immediate short-term context, and use the *generic* act level of MALTUS [15] (compatible with the MRDA scheme [16]). This includes the basic acts *Statement*, *Question*, *Backchannel* and *Floorholder*, but not more intentional acts such as e.g. *propose*, *challenge* (see below).

However, rather than simply label moves, we use their *antecedent* property to express discourse structure directly, relating each move to its antecedent. At this level we restrict moves to having a single antecedent, but allow multiple moves to share the same antecedent; this results in a tree structure (following [17]) able to express not only simple adjacency pairs but multiple possibly simultaneous threads represented by the branches of the tree. We take each tree to be a *Discourse*, a structurally related set of individual *Communicate* acts, required to be semantically or pragmatically coherent via constraints on their structural relations.

These constraints on the classes of move that can serve as each others' antecedents can of course be expressed directly by constraints on the *antecedent* property associated with those classes (e.g. answers must have queries as antecedents, backchannels must have antecedent moves with different speakers). However, our intention is to model not only the move structure of the discourse, but its effect on the emerging context, and so we combine this approach with a notion of *information state* and constraints on its update. This allows us to express the information-state update approach familiar in dialogue processing ([18] among others) directly within the MMDO, rather than requiring a separate dialogue management module or rule set. As set out below, we believe this is advantageous for automated processing and learning, allowing multiple constraint types to be considered simultaneously. The exact constraints will depend on the model of information state used: in an obligation-based model an *Ask* move can be associated directly with the introduction of an addressee's obligation to address the question; in a question-based model it can be associated with the direct introduction of a new question under discussion [18]. Importantly, including these fine-grained semantic constraints does not commit us to a bottom-up approach, building semantic interpretations and using them to derive move type; on the contrary, standard dialogue move classifiers can be used to hypothesize move types, and the information state constraints used to influence or disambiguate semantic interpretation.

Argumentation and Decision-making. At a higher level of abstraction, we also allow for a coarser-grained level of structure intended to model the argumentative and decision-making processes of meeting discourse (embodying a notion similar to that of "rhetorical relations" or "discourse structure" in the analysis of text) such as the raising of issues and the proposal, defense, rejection and acceptance of alternative solutions to the issue [19]. We do not regard it as either practicable or desirable to assign this structure at the level of individual utterances (the level of individual *Communicate* acts assumed in the dialogue move structure of the previous section). Instead, raising issues or proposing alternatives is a function often performed by segments of multiple utterances. A single coherent proposal sequence might consist of multiple atomic statements and questions, and it will be most useful to users to report it in this way. We therefore posit *Communicate* events that can have multiple *Encode* subevents, spanning those events which characterize dialogue moves. These higher-level acts of communication characterize steps in a negotiative process such as *Propose*, *Reject*, and *Accept*, each acting on an *Issue* which is represented using the domain ontology in the same manner as the logical form content of dialogue-level communicative acts.

3.3 Modeling the Meeting Activity

The previous sections describe a bottom-up discourse model, assembling a pragmatically unified *Discourse* structure out of interrelated *Communicate* events. However, meetings are not just discourse; they may include non-communicative activities (e.g. note-taking, waiting for all to arrive) and multiple discourses (e.g. simultaneous side conversations, dialogues separated by breaks for equipment setup). The MMDO therefore models a *Meeting* as an independent class of collaborative *Activity*, an event that has a collection of component *subevents*, the majority of which are *Discourses*. Our only restriction on the subevents is that they occur in one location over a contiguous period of time. As well as the bottom-up characterization, we can therefore also segment *Meeting* and *Discourse* activities in a top-down, coarser-grained way.

Coarse Segmentation. User studies such as [6] demonstrate that a temporally-coarse characterization of a meeting can help users to extract information from annotated meeting records. Automatic coarse segmentation of meetings has correspondingly been the subject of much research, but approaches differ widely in the concepts of *segment* used. One approach is to segment according to "group actions", recognizing physical group activities using speech and/or multimodal features of the discourse [20–22]. The taxonomies used combine a high-level analysis of discourse type (e.g. monologue and discussion) with physical actions of the participants (e.g. presence at the whiteboard and note-taking). In earlier work [23, 20], the taxonomy included activities based on an argumentative dimension of the discourse (e.g. consensus and disagreement), though these do not appear in later analysis. [5] suggest a similar set of "meeting activities" but include

a wide variety of other concepts like voting, multiple simultaneous discussions, and silence. A contrasting approach [24] suggests a simple taxonomy contrasting multi-party, multi-directional exchange of information with uni-directional exchange, to attain high coverage and low ambiguity. In addition, segmentation can also be driven by content – e.g. [25] incorporate lexical features to segment discourse by topic.

It is clear from this variety of segmentation methods that no single segmentation nor taxonomy of segments is objectively optimal. Nevertheless, each type of segmentation is likely to provide a useful means for meeting browsing, summary and information retrieval. Therefore, rather than identifying a single taxonomy of segment classes in the MMDO, we have adopted the aims of high coverage, low ambiguity, and high inter-annotator agreement highlighted in [24] and [9], and have identified a number of nominally independent dimensions over which either a `Meeting` or `Discourse` can be usefully segmented and classified.

At a coarse-grained level, a `Meeting` may be segmented along the dimensions of *physical state* and *agenda state*. *Physical state* depends only on the physical activities of the participants (for example, all participants being seated around a table, vs. one being at the whiteboard while the rest are in their seats). *Agenda state* refers to the position within a previously defined meeting structure, whether specified explicitly as an agenda (providing a list of classes) or implicitly via the known "rules of order" for particular formal meeting types.

At a similar level of granularity, `Discourses` may be segmented along the dimensions of *information flow* and *topic*. *Information flow* describes the general discourse type (e.g. is the subject matter open for discussion with participation by several parties, or is there a one-directional flow as in a presentation or briefing) [24]. *Topic* then describes the coherence of the theme or semantic content of the discussion (we expect this to align significantly with the agenda state for some meeting types). We anticipate that both of these dimensions will be useful for browsing and summarization of meetings, and have produced annotations and initial algorithms to support doing this automatically [26]. We also anticipate that finer-grained segmentations of `Discourses` may be useful, for example according to *floor-holding* activity, and include this ability in the MMDO. Annotation at this level is currently being investigated.

Participant Roles and Segment Classes. In each of the above dimensions, segments may then be classified and participants assigned roles in those events. While we have yet to define a comprehensive set, we provide some potential examples to clarify.

In the dimension of physical state, a frequent suggestion in the literature is for a segment class of "presentation" or "whiteboard" [20–22]. In our model, the *physical state* of being at the whiteboard is represented independently of an *information flow* dimension. Thus, for the segment in the latter dimension, the roles of `InformationProvider` and `InformationConsumer` are specified (see [24]); while the segment in the former dimension will require a single role of

one person at the whiteboard, characterizing it independently as a whiteboard activity.

As a further example, in the turn-taking dimension, a single person may be said to be the *FloorHolder* for some segment of a *Discourse*, and the ontology may assert the constraint that only one person may play this role. Of course, this state will be affected by the floor-handling nature of communicative acts and constraints may be imposed on this relationship in the ontology as well.

4 Automatic Processing and Annotation

The depth and breadth of the ontology mean that it provides not only a complete basis for knowledge storage and annotation, but also a framework for communication between software agents and for machine learning across the various sources of information that those agents provide. A multi-agent system has been built (in collaboration with other project partners) that populates a knowledge base with the fundamental physical signal information (video, audio, and sketch) recorded during a meeting. Given that information, separate interpretive agents can populate the knowledge base with instances of the classes described above, building up a representation of the discourse, and perhaps using each others' assertions to learn.

At the most basic level, the Sphinx speech recognizer is used to segment the audio signal into utterances, positing instances of *Speak* events with their associated *Messages* (transcribed *Words* and *Sentences*). Video processing agents similarly posit their own *Embody* events with physical *Messages* (e.g. head nods, whiteboard-written words). At a higher level, a robust broad-coverage version of the Gemini semantic parser [27] is used to annotate spoken *Messages* with logical form fragments, and (where possible) to postulate associated *Communicate* events with their associated information content expressed via instances of events and entities in the CLib ontology.

There is now wide scope for designing and testing machine-learning agents that use the rich information available in the knowledge base to enrich or disambiguate the basic knowledge already being asserted, and to populate with the higher-level discourse structure elements described in Sect. 3.2 and 3.3. Our first step has been to learn classifiers for topic segmentation. A number of different approaches have been investigated, both discriminative (including decision trees based on lexical and discourse information such as speaker activity changes and the proportion of silence, following [25], and maximum entropy models based on simple lexical features) and generative (adapting [28] to model discourse topic shifts as changes between states in a topic-word Markov model). Results so far are encouraging, with P_k error levels against a set of human annotations approaching 30% (a similar level to that when comparing human annotator agreement, see [26]) for individual classifiers. We now plan to investigate classifier combination and boosting. We also plan to use the availability of simultaneous multimodal information to learn classifiers for speech act detection and addressee

detection (using not only prosodic and lexical information, but the semantic parser output).

Both human and automated annotation of meetings is currently being performed in this framework, though not for all components of the ontology outlined above. In the future, we expect to investigate these areas, which include principally the argumentative and decision-making aspects, semantic alignment with domain ontologies, and detection of floor-holding mechanisms and addressee detection.

Acknowledgments

The authors would like to thank Satanjeev Banerjee, Ken Barker, Jerry Hobbs, and Sanjeev Kumar for their help and suggestions toward the design of this ontology, as well as Bill Jarrold for his implementation of the design in KM. We also wish to express our appreciation to David Demidjian, Lynn Voss, Yitao Sun, and the other CALO developers who have worked with us to align their components with the ontology to create CALO, v.2! This work was supported by DARPA grant NBCH-D-03-0010.

References

1. Nicholas C. Romano, J., Jay F. Nunamaker, J.: Meeting analysis: Findings from research and practice. In: Proc. 34th Hawaii International Conference on System Sciences. (2001)
2. Lisowska, A., Popescu-Belis, A., Armstrong, S.: User query analysis for the specification and evaluation of a dialogue processing and retrieval system. In: Proc. 4th International Conference on Language Resources and Evaluation (LREC). (2004)
3. Reidsma, D., Rienks, R., Jovanović, N.: Meeting modelling in the context of multimodal research. In: Lecture Notes in Computer Science. Volume 3361. Springer-Verlag (2005) 22–35
4. Bachler, M.S., Shum, S.J.B., Roure, D.C.D., Michaelides, D.T., Page, K.R.: Ontological mediation of meeting structure: Argumentation, annotation, and navigation. In: Porc. 1st International Workshop on Hypermedia and the Semantic Web (HyperText). (2003)
5. Marchand-Maillet, S.: Meeting record modelling for enhanced browsing. Technical Report 03.01, Computer Vision and Multimedia Laboratory, Computing Centre, University of Geneva, Switzerland (2003)
6. Banerjee, S., Rose, C., Rudnicky, A.: The necessity of a meeting recording and playback system, and the benefit of topic-level annotations to meeting browsing. Submitted to INTERACT 2005 (2005)
7. Barker, K., Porter, B., Clark, P.: A library of generic concepts for composing knowledge bases. In: Proc. 1st International Conference on Knowledge Capture. (2001)
8. Clark, P., Porter, B.: KM - The Knowledge Machine 2.0: Users manual (2004) http://www.cs.utexas.edu/users/mfkb/RKF/km.html.
9. Popescu-Belis, A.: Dialogue acts: One or more dimensions? ISSCO Working Paper 62 (2005) University of Geneva.
10. Clark, H.H., Krych, M.A.: Speaking while monitoring addressees for understanding. Journal of Memory and Language **50** (2004) 62–81

11. Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X.F., Kirbas, C., McCullough, K.E., Ansari, R.: Multimodal human discourse: gesture and speech. ACM Transactions on Computer-Human Interaction **9** (2002) 171–193
12. Farrar, S., Langendoen, T.: A linguistic ontology for the semantic web. Glot International **7** (2003) 97–100
13. Ide, N., Romary, L., de la Clergerie, E.: International standard for a linguistic annotation framework. In: Proc. HLT-NAACL'03 Workshop on the Software Engineering and Architecture of Language Technology. (2003)
14. Dowding, J., Purver, M.: Trying to parse multi-party meetings. Submitted to the 6th SIGdial Workshop on Discourse and Dialogue (2005)
15. Clark, A., Popescu-Belis, A.: Multi-level dialogue act tags. In: Proc. 5th SIGdial Workshop on Discourse and Dialogue. (2004)
16. Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., Carvey, H.: The ICSI meeting recorder dialog act (MRDA) corpus. In: Proc. 5th SIGdial Workshop on Discourse and Dialogue. (2004)
17. Lemon, O., Gruenstein, A.: Multithreaded context for robust conversational interfaces: Context-sensitive speech recognition and interpretation of corrective fragments. ACM Transactions on Computer-Human Interaction **11** (2004)
18. Traum, D., Bos, J., Cooper, R., Larsson, S., Lewin, I., Matheson, C., Poesio, M.: A model of dialogue moves and information state revision. In: Task Oriented Instructional Dialogue (TRINDI): Deliverable 2.1. University of Gothenburg (1999)
19. Pallotta, V., Niekrasz, J., Purver, M.: Collaborative and argumentative models of natural discussions. In: Proc. 5th Workshop on Computational Models of Natural Argument. (2005)
20. Dielmann, A., Renals, S.: Dynamic bayesian networks for meeting structuring. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing. (2004)
21. Reiter, S., Rigoll, G.: Segmentation and classification of meeting events using multiple classifier fusion and dynamic programming. In: Proc. International Conference on Pattern Recognition. (2004)
22. McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., Zhang, S.: Automatic analysis of multimodal group actions in meetings. IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005) 305–317
23. McCowan, I., Bengio, S., Gatica-Perez, D., Lathoud, G., Monay, F., Moore, D., Wellner, P., Bourlard, H.: Modeling human interaction in meetings. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing. (2003)
24. Banerjee, S., Rudnicky, A.: Using simple speech-based features to detect the state of a meeting and the roles of the meeting participants. In: Proc. 8th International Conference on Spoken Language Processing. (2004)
25. Galley, M., McKeown, K., Fosler-Lussier, E., Jing, H.: Discourse segmentation of multi-party conversation. In: Proc. 41st Annual Meeting of the Association for Computational Linguistics. (2003)
26. Gruenstein, A., Niekrasz, J., Purver, M.: Meeting structure annotation: Data and tools. In: Submitted to the 6th SIGdial Workshop on Discourse and Dialogue. (2005)
27. Dowding, J., Gawron, J.M., Appelt, D., Bear, J., Cherny, L., Moore, R., Moran, D.: Gemini: A natural language system for spoken language understanding. In: Proc. 31st Annual Meeting of the Association for Computational Linguistics. (1993)
28. Blei, D., Moreno, P.: Topic segmentation with an aspect hidden Markov model. In: Proc. 24th Annual International Conference on Research and Development in Information Retrieval. (2001) 343–348