Dialogue and Discourse 4(2) (2013) 185-214 doi: 10.5087/dad.2013.209

Annotation of negotiation processes in joint action dialogues

Thora Tenbrink T.TENBRINK@BANGOR.AC.UK
School of Linguistics and English Language
Bangor University
Bangor, Gwynedd LL57 2DG, UK

Kathleen Eberhard KEBERHAR@ND.EDU
Department of Psychology
University of Notre Dame
Notre Dame, IN 46556, USA

Hui Shi SHI@INFORMATIK.UNI-BREMEN.DE
FB3 Faculty of Mathematics and Computer Science
University of Bremen
28334 Bremen, Germany

Sandra Kübler SKUEBLER@INDIANA.EDU
Department of Linguistics
Indiana University
Bloomington, IN 47405, USA

Matthias Scheutz MATTHIAS.SCHEUTZ@TUFTS.EDU
Department of Computer Science
Tufts University
Medford, MA 02155, USA

Editors: Stefanie Dipper, Heike Zinsmeister, Bonnie Webber

Abstract

Situated dialogue corpora are invaluable resources for understanding the complex relationships among language, perception, and action. Accomplishing shared goals in the real world can often only be achieved via dynamic negotiation processes based on the interactants' common ground. In this paper, we investigate ways of systematically capturing structural dialogue phenomena in situated goal-directed tasks through the use of annotation schemes. Specifically, we examine how dialogue structure is affected by participants updating their joint knowledge of task states by relying on various non-linguistic factors such as perceptions and actions. In contrast to entirely language-based factors, these are typically not captured by conventional annotation schemes, although isolated relevant efforts exist. Following an overview of previous empirical evidence highlighting effects of multi-modal dialogue updates, we discuss a range of relevant dialogue corpora along with the annotation schemes that have been used to analyze them. Our brief review shows how gestures, action, intonation, perception and the like, to the extent that they are available to participants, can affect dialogue structure and directly contribute to the communicative success in situated tasks. Accordingly, current annotation schemes need to be extended to fully capture these critical additional aspects of non-linguistic dialogue phenomena, building on existing efforts as discussed in this paper.

Keywords: Annotation schemes; dialogue corpora; situatedness; joint action; dialogue modelling

 Submitted 03/12; Accepted 12/12; Published online 07/13

1. Introduction

Dialogue corpora are of great utility to a diverse set of researchers, ranging from those who intend to test theories of language use in human social situations (e.g., conversation analysis, discourse analysis) to researchers developing artificial agents that can effectively interact with humans via natural language (human-computer and human-robot interaction). While the identification of factors that affect the efficiency of language use are common to both research agendas, the latter community is particularly interested in corpora of *task-oriented dialogues* given that artificial agents are typically designed to be directed by humans for performing various tasks.

Over the years, an increasing number of task-oriented dialogue corpora has been developed, with varying task and interaction complexities. Among the most long-standing and well-known corpora are the *Map Task corpus* (Anderson et al., 1991) and the *TRAINS corpus* (Heeman and Allen, 1995) which are both based on relatively simple tasks and domains (e.g., drawing or planning routes on a two-dimensional map) with no nonverbal response types. More recent corpora (discussed in detail below) are based on more complex *situated* tasks and domains where interactants have to manipulate real-world objects, move through environments, and overall perform more naturalistic perceptions and actions.

Our focus in this paper is on annotation schemes that capture structures of negotiation processes in joint action dialogues, building on speech-act related theories (e.g., Allen and Perrault, 1980; Cohen and Levesque, 1980) to describe dialogic actions in discourse (i.e., *dialogue acts* as specialized speech acts referring to dialogue actions). This structure-oriented approach contrasts with another line of research on automatic dialogue control, which has produced important results for negotiation processes using *Information State* systems. These focus primarily on belief states on the part of each interactant rather than structural patterns, following Grosz and Sidner (1986). A typical target for research in this area concerns information seeking dialogues (e.g., Bohlin et al., 1999; Larsson and Traum, 2000), addressing the negotiation of alternative solutions (although some applications in multimodal scenarios have been proposed as well, e.g., Lemon et al., 2002). In contrast, joint action scenarios do not primarily focus on the exchange of information but rather on achieving shared goals that require physical actions. In such settings, the negotiation of alternative solutions and the conveyance of facts are often less important compared to updating common ground with respect to perceptions, actions, and action outcomes. We will argue that this intricate interplay between perceptions, actions, and linguistic exchanges requires fine-grained annotation schemes in order to model the organization of the dialogue adequately.

In Section 2, we thus start with a discussion of dialogue aspects that are relevant for annotations of task-oriented dialogue, starting from Clark's influential view (Clark, 1996; Clark and Wilkes-Gibbs, 1986) of dialogue as a *joint activity*. In the case of task-oriented settings, the dialogue itself is embedded within a larger joint activity corresponding to the task's prescribed goals, and the verbal dialogue is a means by which the participants coordinate their task-relevant (non-communicative) actions. However, as our discussion highlights, the relative role of language is affected by the extent to which the interactants share their perceptions of the task domain, of the actions performed within it, and of each other. That is, shared perceptions provide a reliable basis for the interactants to update their *common ground*, which includes shared knowledge about the task state and the remaining goals. This reduces the need to rely on dialogue for the updating process; hence annotation of nonverbal communicative actions (e.g., gaze, gestures, head nods, etc.) is important for capturing the participants' coordination processes. Conversely, the less shared perception there is, the greater the

participants' reliance on dialogue for updating their common ground; this enhances the need for annotating further layers of the dialogue that convey crucial information (e.g., intonational patterns, types of referential expressions, discourse markers, syntactic complexity, types of dialogue acts, etc.).

With these considerations in mind, Section 3 reviews different annotation schemes proposed for corpora that differ with respect to the level of shared perception as well as the situatedness of the task and the complexity of the task goals. We specifically consider the usability of the annotation schemes for investigating the coordination and negotiation processes, i.e., the ease of updating common ground, in task-oriented dialogue.

Section 4 then provides a brief discussion of the layers of annotation that are required for fully capturing activity coordination and goal negotiation processes as they occur in joint action scenarios. We conclude by a brief discussion of the challenges ahead and consequences for dialogue annotation and corpus data collection.

2. Common ground in task-oriented dialogue

We adopt Clark's (1996) view of dialogue as a *joint activity*, which he defines as involving two or more participants coordinating their actions in order to achieve a mutual dominant goal. In the case of *task-oriented dialogue*, the dialogue is a joint activity embedded within a larger joint activity corresponding to the prescribed goal of the task. The task's goal is specified with respect to a circumscribed domain of objects, the actions that are to be performed on the objects, the constraints on the actions, and the participants' responsibilities in performing the actions. Thus, the dialogue is important for coordinating the participants' performance of non-communicative actions involving task-relevant objects. Achieving the task's dominant goal requires a hierarchy of nested subgoals, and corresponding joint activities. This hierarchy is central to the participants' common ground, which is the knowledge that they believe they share about their joint activity. At any given point, the common ground includes the joint activity's initial state, its current state, and the mutually known events and perceptions that have led to the current state (Clark, 1996). The updating of common ground is fundamental to the participants' coordinating both the sequencing and timing of their actions and goals. The ease or accuracy of the updating determines the efficiency with which they perform the overall joint activity, for instance, in terms of accuracy, time needed, number of subgoals that were achieved, etc.

An overall aim of analyzing corpora of task-oriented dialogue then is to identify factors that affect the participants' coordination of their joint activity, and hence the efficiency of their performing the task. Because coordination is a dynamic process that involves updating information in common ground, important factors will include those that affect the ease and/or accuracy of the updating. The greatest efficiency occurs with face-to-face interaction in which the participants share perception of each other's performance of task-relevant actions. This shared perception provides reliable and immediate updating of the task's status in common ground, with less reliance on dialogue for the updating.

Moreover, face-to-face interactions enable the use of deictic gestures and other visual cues that can focus the participants' attention and facilitate coordinated behavior. For example, the ability to monitor each others' gaze allows for joint attention processes to aid the coordination of the task-relevant actions (e.g., Bangerter, 2004; Bard et al., 1996; Brennan et al., 2008; Neider et al., 2010). In particular, a speaker's gaze at an object when referring to it can facilitate an addressee's estab-

lishment of reference (e.g., Hanna and Brennan, 2007), and the addressee's gaze at an object can provide evidence of his or her understanding to the speaker. The facilitation is reflected in less interruption in speech and fewer turns (Boyle et al., 1994; O'Malley et al., 1996), due to the availability of non-verbal signals of understanding (e.g., head nods) and the use of mutual gaze for regulating turn-taking (e.g., Kendon, 1967; Allwood et al., 2007). Substantial facilitation can be achieved even when only the interlocutors' faces are perceivable, e.g., when interactants are co-located but blocked from seeing task-relevant objects and actions (e.g., Brennan et al., 2012). However, the beneficial effects of shared face perception do not appear to extend to remote interactants who share face perception via video conferencing. In fact, video conferencing increases the number of turns (O'Malley et al., 1996). If a delay of responses is involved, this will affect the dialogue even if it is only slight (Sellen, 1995; Fischer and Tenbrink, 2003).

Since multimodal layers of interaction thus evidently affect how people perceive a situation and act within it, the structure of interaction should be fundamentally affected by the availability of multimodal channels. However, although dialogic negotiation processes have been approached from various perspectives, to our knowledge no annotation scheme has been proposed that systematically captures negotiation processes and keeps track of common ground via a comprehensive structural integration of non-verbal dialogue contributions. In the following, we examine promising efforts in relevant directions.

3. Annotation schemes for various layers of task-oriented corpora

Task-based dialogues are comprised of many layers of task-relevant information exchanges, including the critical layer of *dialogue acts*. Existing annotation schemes focusing on this aspect include DIT++ (Bunt, 2009), Verbmobil (Alexandersson et al., 1998), as well as DAMSL (Allen and Core, 1997; Jurafsky et al., 1997) and HCRC (Carletta et al., 1997). Other schemes highlight further layers of dialogues. The well-developed annotation scheme ToBI (Beckman and Hirschberg, 1994; Beckman et al., 2005) captures features of *intonation* (e.g., in the Columbia Games Corpus, Gravano, 2009), and nonverbal forms of communication such as *gestures* are captured, for instance, by FORM (Martell, 2002; Martell and Kroll, 2007) and the MPI gesture annotation (Kipp et al., 2007).

The discussion in this section will start with a brief description of the *Human Communication Research Centre* (HCRC) annotation scheme and the *Dialogue Act Markup in Several Layers* (DAMSL) scheme. They were initially developed for the Map Task corpus (Anderson et al., 1991) and the TRAINS corpus (Heeman and Allen, 1995), respectively, but have also been applied to other corpora, e.g., the Maze Task (Kowtko et al., 1993) and Switchboard (Jurafsky et al., 1997). While both schemes illustrate the kinds of dialogue act classifications that reflect coordination of communicative actions, they differ with respect to the communication levels they recognize: the HCRC scheme distinguishes transaction, game, and move levels; DAMSL uses communication management, task management, and task levels. Based on an extensive study of coding outcomes using these two annotation schemes, Stirling et al. (2001) provide a detailed analysis of the extent to which the schemes can capture relations between dialogue structure and prosody patterns. Since we are concerned with joint action tasks in the current paper, our focus lies on individual dialogue moves and task-relevant dialogue acts as part of the negotiation process. Also, we focus on the types of phenomena that can potentially be highlighted by annotation schemes, rather than attempting a direct comparison and evaluation based on extensive recording procedures.

The following subsections provide descriptions of a non-exhaustive list of corpora and annotation schemes. We present examples of how the speakers update common ground, and discuss these in light of annotations of dialogue acts, intonation, and gestures. The corpora were selected in part because they are well-known (in the case of the Map Task and TRAINS corpus), and in part because they contribute new types of annotation, or new insights concerning the specific phenomena relevant for joint action tasks. We conclude the description of each corpus and annotation example with a brief evaluation of its usability for describing updating mechanisms for common ground in task-oriented scenarios.

3.1 Map Task Corpus and the HCRC Coding Scheme

The Map Task (Anderson et al., 1991) represents a widely used paradigm that has been employed in various versions. Typically it involves two participants who both have a map, which the other person cannot see. The maps contain line drawings of labeled objects (e.g., rocks, bridge, mountain, etc.) which can serve as landmarks. The maps are not identical and the participants are informed of this. Both maps are marked with a start point and an end point. One person's map has a line indicating a route from the start point to the end point. The person with this map is the instruction Giver, and the other person is the Follower. The task is for the Follower to draw the same route on his or her map as the one indicated on the Giver's map by following the Giver's instructions.

The HCRC annotation scheme[1] (Carletta et al., 1997) consists of three interdependent levels of coding. The basic dialogue acts are coded as conversational moves, which indicate the communicative purpose of an utterance. Each utterance is coded with a single move, which is classified as either initiation or response. Initiation moves include instructions (INSTRUCT), explanations (EXPLAIN), and questions (QUERY-YN, QUERY-W) and two types of requests: confirmation of accurate understanding (CHECK) and confirmation of a readiness to continue (ALIGN). The initiation moves elicit response moves, which include acknowledgments (ACKNOW), replies to questions (REPLY-Y, REPLY-N, REPLY-W), and clarifications (CLARIFY). The game level identifies the set of moves that fulfill the purpose of an overall initiation move. For example, an instruction may initiate a game that is ended by an acknowledgment of the completed action. Games can be embedded within games, as in the case when an instruction move is followed by a question-reply sequence (game) reflecting the need for clarifying information before the instruction can be completed. The transaction level distinguishes sets of games that concern performing a task-relevant action from those that concern managing the task (e.g., reviewing transactions, which refer to previously completed actions, and overview transactions, which refer to the overall task). The normal transactions correspond to subtasks (segments) of the overall task. The identification of the subtasks requires taking into account the task-relevant actions performed by the participants. In the case of the Map Task, the transactions are labeled with the beginning and ending points on the map that correspond to the segments of the route that are incrementally drawn on the map by the Follower.

Table 1 gives an example of a normal transaction consisting of a dialogue sequence that corresponds to the Follower marking the first segment of the route on his or her map. The transaction involves two instruction games and an embedded check game in which the Follower checks their understanding of the Giver's preceding instruction. INSTRUCT, ACKNOW, CHECK, REPLY-Y, CLARIFY are moves.

1. More information about the corpus and the annotation can be found at `http://groups.inf.ed.ac.uk/maptask/`.

ID	Speaker	Utterance	Game	Move
utt2	Giver	starting off we are above a caravan part	Game 1 (INSTRUCT)	INSTRUCT
utt3	Follower	mmhmm		ACKNOW
utt4	Giver	we are going to go due south straight south and then we're going to g– turn straight back round and head north past an old mill on the right hand side	Game 2 (INSTRUCT)	INSTRUCT
			Game 3 (CHECK), embedded 1	
utt5	Follower	due south and then back up again		CHECK
utt6	Giver	yeah		REPLY-Y
utt7	Giver	south and then straight back up again with an old mill on the right and you're going to pass on the left-hand side of the mill		CLARIFY
utt8	Follower	right okay		ACKNOW

Table 1: An example dialogue from the Map Task Corpus.

Usability for task-oriented scenarios. Three hierarchical levels of dialogue structure can be distinguished in this annotation scheme: conversational moves, games, and transactions. The dialogue's structure closely reflects the task's goal structure because the dominant goal involves minimal action, which is performed by the Follower (i.e., the Follower draws a line indicating a route on the map). The Follower's incremental drawing of the route is included as a layer of annotation, which is used to segment the dialogue at the transaction level of coding. This annotation also includes instances when the Follower crossed out a portion of the route and drew a new one, reflecting the correction of inaccurate information in common ground. The online corpus also has other layers of linguistic annotation (e.g., disfluencies, part-of-speech tagging) and a layer of gaze annotation for a subset of participants. The latter indicates whether the participant was looking at the map or at the other participant during the task. However, the gaze annotation has not been used to investigate the dynamic process of coordinating the joint actions. Rather, the effects of gaze have been investigated by comparing overall differences in various aspects of the dialogue annotations (e.g., number of turns, number of words, frequency of types of disfluencies) in eye-contact vs. no eye-contact conditions (Branigan et al., 1999; Boyle et al., 1994; Bard et al., 2007).

3.2 The TRAINS corpus and DAMSL annotation scheme

The TRAINS corpus (Heeman and Allen, 1995, 1994) consists of dialogues between two participants in a problem-solving task involving shipping goods by train to various cities. The participants were informed that the goal was to build a computer system that can aid people in problem-solving tasks (see Allen et al., 1995; Sikorski and Allen, 1997; Traum, 1996). One participant had the role of system User and the other had the System role. The participants were asked to act normally rather than mimic an automatic system. They were given a map showing engines, box cars, and factories, and their locations along a single rail track. The User was given various tasks (e.g., construct a plan for shipping a box car of oranges to Bath by 8 AM). The System had further information that was important for the User's task such as timing and engine capacity. No actions were directly executed in the task. Since the knowledge about the task was partially shared and information was distributed between participants, the TRAINS corpus contains rich negotiation dialogues and frequent initiative shifts.

ID	Speaker	Utterance	Dialogue act
utt1	User	We could use the train at Dansville.	OPEN-OPTION
utt2	System	Could we use one at Avon instead?	OPEN-OPTION, HOLD(*utt1*)
utt3	User	No, I want it for something else.	ASSERT, REJECT(*utt2*)
utt4	System	How about the one at Corning then?	OPEN-OPTION, HOLD(*utt1-utt3*)
utt5	User	Okay.	ASSERT, ACCEPT(*utt4*)
utt6	System	Okay.	ACCEPT(*utt1-utt5*)

Table 2: An example dialogue from the TRAINS Corpus.

The DAMSL scheme (Core and Allen, 1997; Allen and Core, 1997) involves three orthogonal layers of coding. The layer that codes the communicative purpose or functions of utterances (dialogue acts) has two categories: forward-looking functions and backward-looking functions (similar to the initiation-response dichotomy in the HCRC annotation scheme). Forward-looking functions affect the subsequent interaction as in the case of questions (INFO-REQUEST), instructions (ACTION-DIRECTIVE), suggestions (OPEN-OPTION), statements (ASSERT), etc. The backward-looking functions concern the utterance's relation to the previous discourse, e.g., answers to questions (ANSWER), signals of agreement (ACCEPT) or understanding (ACKNOWLEDGE), etc. An important feature of the coding at this level (contrasting with the HCRC scheme) is that utterances may simultaneously perform more than one function. The second layer is the information level, which distinguishes utterances that are about the performance of task-relevant actions (task), from those that concern planning, coordinating, or clarifying the task-relevant actions (task-management), and those that concern the management of the dialogue itself (e.g., acknowledgments, opening, closings, repairs, signals of delays, etc.). The third layer (communicative status) distinguishes complete utterances from incomplete, abandoned, or unintelligible utterances. Table 2 shows an example from the TRAINS corpus (adapted from Allen and Core, 1997)[2] that illustrates the dialogue acts at the communicative purpose layer, which is the most relevant layer for the discussion of functional dialogue structures.

The TRAINS scenario differs from the Map Task in that responsibilities and knowledge are divided between the speakers. Although the User is responsible for deciding the planned actions involving which trains pick up, move, and deliver cargo, the System can suggest plans as well as reject plans proposed by the User due to constraints which are known to the System, but not to the User. We provide two examples to illustrate how this particular aspect affects dialogue structure. Table 2 shows a portion of a dialogue from the TRAINS corpus that has two embedded subdialogues. The dialogue begins with an OPEN-OPTION tag, which is a forward-looking function corresponding to a suggested course of action. The first embedded subdialogue is tagged by HOLD(*utt1*), which is a backward-looking function that indicates a suspension of a response to the suggestion in the first utterance. The suspension is due to the System suggesting an alternative course of action (OPEN-OPTION), which is rejected by the User in *utt3*. The second embedded subdialogue results from the System suggesting a second alternative course of action in *utt4*. This suggestion is accepted by the User in *utt5*, with the System's acceptance in *utt6* closing the task that began with *utt1*.

The more complex example in Table 3 further illustrates the negotiation of shared information in this joint task. We take a closer look at the underlying intentions behind dialogue acts as far as they

2. See `http://www.cs.rochester.edu/research/cisd/resources/damsl/` for the annotation details.

Id	Speaker	Utterance	Dialogue act
utt7	User	the orange warehouse where I need the oranges from is in Corning	ASSERT
utt8	System	right	ACCEPT(*utt7*)
utt9	User	so I need is it possible for one of the engines would it be faster for an engine to come from Elmira or Avon	INFO-REQUEST
utt10	System	uh Elmira is a lot closer	ANSWER(*utt9*), ASSERT
utt11	User	what time would engine two and three leave Elmira	INFO-REQUEST
utt12	System	um well they're not scheduled yet	ANSWER(*utt11*), REJECT(*utt11*)
utt13	System	but we can send them at any time we want	ANSWER(*utt11*), OPEN-OPTION
utt14	User	okay	ACCEPT(*utt13*)
utt15	System	uh so + if we sent them right away it'd get there at at um at two a.m.	OPEN-OPTION, OFFER, ASSERT
utt16	User	at Corning	HOLD(*utt15*), INFO-REQUEST
utt17	System	yeah	ANSWER(*utt16*), ASSERT
utt18	User	and how long would it take to get from Corning to Bath	INFO-REQUEST, COMMIT
utt19	System	uh two hours	ANSWER(*utt18*), ASSERT
utt20	User	how long would it take to load the oranges from the warehouse into the engine	INFO-REQUEST, COMMIT
utt21	System	uh well we can't load oranges into an engine we need a boxcar to load them into	ANSWER(*utt20*), REJECT(*utt20*), ASSERT
utt22	User	mm-hm	ACCEPT(*utt21*)
utt23	User	so can I dispatch an engine and a boxcar from Elmira simultaneously to Corning	INFO-REQUEST, OFFER
utt24	System	uh yeah yeah	ANSWER(*utt23*), ASSERT
utt25	System	we can uh connect an engine to the boxcar and then take have the engine take the boxcar to Corning	ACCEPT, COMMIT, ASSERT
utt26	User	so it'll be two hours to Corning	ASSERT
utt27	System	right	ACCEPT(*utt26*)

Table 3: An example dialogue from the TRAINS Corpus, showing the negotiation of shared information.

can be inferred from the speakers' information status. The purpose of the ASSERT dialogue act in *utt7* is to convey the information that oranges are needed (which is unknown to the System), while at the same time establishing the orange warehouse in Corning as common ground. Both aspects of this dual-purpose assertion are accepted by the System in *utt8*. *utt9* contains a request for timing information, which is simply responded to by the System in *utt10*. The following request in *utt11* by the User signals a misconception of the kinds of information available to the System, leading to a rejection in *utt12*. In *utt20*, the User poses an information request containing a commitment to the future action "to load the oranges from the warehouse into the engine" in a way that signals (via presupposition) a misconception of the task. This aspect, which is not captured directly by the annotation, is clarified by the System in *utt21*, leading to a dialogue utterance that is simultaneously a rejection of the request and an assertion of facts, which in the following leads to new considerations and requests for information on the part of the User. Altogether, the dialogue is dominated by the need to confirm and exchange information based on what is mutually known or needs to be conveyed to the dialogue partner.

The DAMSL annotation reflects the negotiation identified here to some extent via the following emerging structures: ASSERT and ACCEPT (e.g., *utt7* and *utt8*, *utt26* and *utt27*), INFO-REQUEST and ANSWER (e.g., *utt9* and *utt10*, *utt18* and *utt19*), and INFO-REQUEST and ANSWER + REJECT as a less well established type of adjacency pair (*utt11* and *utt12*, *utt20* and *utt21*) that can be traced back to the fact that information is distributed across participants.

Usability for task-oriented scenarios. The DAMSL coding scheme is clearly devised for precisely the kind of scenario found in the TRAINS corpus, which involves no shared action or perception, distributed resources of information on both sides, and a joint task. In contrast to the HCRC scheme, the three layers used in DAMSL are not hierarchical, but orthogonal. Rather than providing further detail about a type of verbal interaction, they highlight three different aspects of the same verbal interaction (forward or backward looking functions, information level, and communicative status). Like HCRC, all of the annotation pertains to the verbal interaction, rather than taking further aspects into account such as the checking of information from the database. However, since most of the System's utterances in Table 3 rely on the information solely available in the System's database, such a database check must have been a frequent action on the part of the interactant playing the part of the System. This affects the development of the dialogue since a substantial part of the dialogue consists of the updating of common ground based on information from external resources available only to one of the interactants, accessed in reaction to the other interactant's contributions. In other scenarios, existing beliefs may be negotiated in a balanced way, or information updates may be based on changes in the real world.

3.3 Schober: Accounting for the addressee's abilities

In various papers, Schober (1993, 1995, 1998, 2009) addresses how speakers modify their spatial perspectives for the benefit of their addressee with respect to visually shared scenes (seen from different perspectives). Schober's approach is to provide a systematic analysis along with illustrative examples, where annotation is focused on the conceptual aspect of perspective taking (which is of less concern for our current purposes). The dialogue in Table 4 is an example from Schober (2009, p. 31) and involves speakers with mismatched abilities. We have added a DAMSL-type annotation to this example in order to highlight the communicative acts. The task (a referential communication task) was to identify one out of four dots arranged around a sketched airplane shown on a screen.

ID	Speaker	Utterance	Dialogue act
utt1	Director	Okay my plane is pointing towards the left	ASSERT
utt2	Matcher	Okay	ACCEPT(*1*)
utt3	Director	And the dot is directly at the tail of it	ASSERT
utt4	Director	Like right at the back of it	ASSERT
utt5	Matcher	Okay mine is pointing to the right	ACKNOWLEDGE(*3,4*), ASSERT
utt6	Director	Oh yours is pointing to the right	ACKNOWLEDGE REPEAT-REPHRASE(*5*)
utt7	Matcher	Yeah	ACCEPT(*6*)
utt8	Director	So your dot should be on the left	ASSERT
utt9	Director	Because my dot is on the right	COMPLETION(*8*)
utt10	Director	In back of it	ASSERT
utt11	Director	So your dot should be at the left	REASSERT(*8*)
utt12	Director	At the back of it right	REASSERT(*10*)
utt13	Matcher	Yeah	ACCEPT(*8,10,11,12*)
utt14	Director	Yeah	ACCEPT(*13*)
utt15	Matcher	But if it is the same - but if it - the same dot-right? Wait a minute, if my - your plane is pointing to the left *[something] - *	ASSERT, CORRECTION, *comm. manag.*, REPEAT-REPHRASE(*1*)
utt16	Director	*My* plane is pointing to the left	REASSERT(*1*)
utt17	Matcher	Mm-hm	*comm. manag.*
utt18	Matcher	And that dot and the dot that's highlighted is the one all the way in the back of it	ASSERT
utt19	Matcher	Like behind the tail	ASSERT
utt20	Matcher	Yes, so so my dot is gonna be	ACCEPT(*18,19*), ASSERT
utt21	Director	So my dot is on the right	REASSERT
utt22	Director	And yours should be on the left right	REASSERT, COMPLETION(*20*) INFO-REQUEST
utt23	Matcher	Yeah	ACCEPT, ANSWER(*22*)
utt24	Director	Okay *so your - *	ACKNOWLEDGE(*23*), ASSERT
utt25	Matcher	*Right behind the tail* okay	COMPLETION, ACCEPT(*24*)
utt26	Matcher	Okay	*comm. manag.*

Table 4: An example dialogue from Schober (2009).

Director and Matcher had different views on this scene, but the other person's view was indicated on the screen. However, this fact did not keep participants from negotiating each other's view in order to reach their discourse goal.

In this example, Director and Matcher jointly try to agree on which dot is the correct one. This requires some negotiation, to which both interactants contribute although only the Director has access to the information about the goal dot. The Matcher's contribution is to provide information about the state of understanding, as if thinking aloud for the benefit of the Director, who can use this information to support the thought process. In particular, the Director states the direction of the airplane (from his or her point of view) in *utt1* in order to describe where the dot is located (*utt3*). Instead of mentally rotating the airplane based on this description and marking the dot on the display, the Matcher merely states his or her own view on the scene in *utt5*. Subsequently, the Director describes the location of the dot from the Matcher's perspective (*utt8* to *utt12*). The Matcher exhibits confusion in *utt15* and tries to use the object-centered view to understand the location of the dot. Then the Director again uses different perspectives (*utt21* and *utt22*) to describe the dot, which is accepted by the Matcher. This iterative exchange, required by this particular pair of

speakers to establish common ground related to their different perception of the scene, is in contrast with the following very simple exchange, which achieves the same subgoal:

Director: It's behind the plane.
Matcher: Okay.

The negotiation effects are reflected in our DAMSL-type annotation by frequent cases of REASSERT, REPEAT-REPHRASE, and COMPLETION, along with many cases of ASSERT contributed by both speakers. Thus, a simple instruction is not deemed sufficient by the speakers in this case; the main point of negotiation is to find out how the two views can be matched so as to interpret the instruction correctly. This (conceptual) task is jointly achieved by both interactants, so that the Matcher can follow the instruction of marking the goal dot (which is a simple action once the conceptual matching process has been completed). Accordingly, neither instruction nor action are ever explicitly mentioned throughout this dialogue.

Usability for task-oriented scenarios. This dialogue corpus is particularly well-suited for highlighting how interactants achieve common ground by taking each other's view on the scene into account. The negotiation of spatial perspectives is tightly integrated in the overall task. By highlighting dialogue acts, the annotation in Table 4 reflects the structure of the perspective clarifications needed to achieve common ground about spatial locations. However, a thorough understanding of the interactants' intentions and conceptual perspectives is only possible by accounting for the associated pictures as perceived by the participants.

3.4 Airplane: Continuing each other's thoughts

The Collaborative Research Center SFB 360 *Situated Artificial Communicators* in Bielefeld (Germany) (Rickheit and Wachsmuth, 2006) explored a scenario in which participants had to build a toy airplane from its parts[3]. While this scenario was used to address diverse challenges in the artificial intelligence area including human-robot interaction, we focus here on a human-human dialogue setting involving language-based instruction. Participants were separated by a screen and could not see each other. One of them, the Instructor, had a completed toy airplane, while the other (the Constructor) had a set of separate parts. Thus, the setting involved sufficient complexity of actions to involve a high amount of negotiation. Poncin and Rieser (2006) discuss a brief dialogue example in much detail in order to establish how speakers manage to negotiate actions, and in particular, how the Constructor completes some of the Instructor's utterances, shown in Table 5 (annotations by original authors, augmented by the dialogue acts in DAMSL for purposes of comparison).

From the example in Table 5, it is obvious that the negotiation of actions can lead to a high involvement by the participants with less knowledge, who make informed guesses about the possible steps of action – even to the extent that they complete the Instructor's sentences. According to Poncin and Rieser (2006) this is only possible because of the information embedded in the directives along with a high amount of shared common ground based on the previous actions and background knowledge. Moreover, they point to the important contribution of prosody in the interpretation of the speakers' joint achievement in this exchange. In particular, "Cnst's completion 'copies' the global, rising-falling pitch movement of Inst's preceding utterance", and "Inst's repair of the proposed

3. The corpora without annotation are available at `http://www.sfb360.uni-bielefeld.de/transkript/`.

ID	Speaker	Utterance	Annotation	Dialogue act
utt1	Inst	So, jetzt nimmst du *Well, now you take*	Inst's proposal	ACTION-DIRECTIVE
utt2	Cnst	eine Schraube, *a screw,*	Cnst's acceptance of Inst's proposal and her proposal for a continuation	ACCEPT(*1*) OFFER
utt3	Inst	eine orangene mit einem Schlitz *an orange flat-head*	Inst's non-acceptance of Cnst's proposal and his repair by extension	REJECT-PART(*2*) ACTION-DIRECTIVE
utt4	Cnst	Ja. *Yes.*	Cnst's acceptance of Inst's repair	ACCEPT(*3*)

Table 5: Example dialogue from the Airplane scenario.

completion is realized using some kind of contrasting prosody" (Poncin and Rieser, 2006, p. 730). This analysis is one of few that aim to capture prosodic phenomena thoroughly (but see also Purver and Kempson, 2004). Although the authors only discuss two specific examples, they claim (based on an investigation of the whole corpus) that the phenomenon is widespread and can be generalized. To our knowledge, no systematic analysis or annotation scheme capturing these effects is available. The DAMSL dialogue acts that we added in the right column of Table 5 capture the outcome, not the process of updating common ground itself.

Usability for task-oriented scenarios. This interaction scenario raises an important issue that future annotation schemes will need to address, namely, which aspects in the joint task will lead to shared common ground that is sufficiently established to enable the less informed person to complete the more informed person's sentences. As part of this process, the contribution of prosodic features is crucial for the update of common ground based on the immediate recognition and integration of subtle prosodic cues.

3.5 Dollhouse: Pro-active cooperation by an instructee

An extensive corpus collected by Coventry, Andonova, and Tenbrink (first cited in Tenbrink et al., 2008) involved pairs of participants who were asked to furnish a dollhouse. Only Directors had full visual information about positions of objects; Matchers were given the task of placing the objects into an empty dollhouse based on the Directors' instructions. Since Directors could not see the Matchers' dollhouses and actions, information needed to be communicated verbally. Since the task did not involve shared action, it could be assumed that Matchers listened to and followed the Director's commands without much negotiation. However, as shown by Tenbrink et al. (2008), this was not the case; Matchers were, in fact, rather active in discussing spatial locations despite their role as recipient of information. In particular, the Matchers' contributions of new spatial content could serve to clarify a global aspect of the current situation, to disambiguate an ambiguous description by explicitly mentioning options, or to specify an object's position. The latter could be achieved by relating it to an(other) object already placed, or by suggesting another spatial term to describe the spatial relationship. Thus, Matchers actively participated in the joint effort of furnishing the dollhouse.

Table 6 shows an example from the Dollhouse corpus; we provide annotations according to the DAMSL scheme. The example illustrates the engagement of the Matcher in the identification and placement of an object (here: a washbasin) in the dollhouse. At the beginning, in *D311-37* there is a clarification question about the Director's current spatial focus, since the Director started a new dis-

ID	Speaker	Utterance	Dialogue act
D311-36	Director	dann is' das nächste Ding du hast ähm *then the next thing is you have um*	ASSERT
D311-37	Matcher	noch immer im selben Raum? *still in the same room?*	INFO-REQUEST
D311-38	Director	genau. *correct.*	ANSWER(*D311-37*)
D311-39	Director	links neben der Toilette hast du diese blaue Stellwand, diese Zwischenwand. *to the left next to the toilet you have this blue movable wall, this partition.*	ASSERT
D311-40	Matcher	ja. *yes.*	ACCEPT(*D311-39*)
D311-41	Director	und dahin, praktisch da links daneben davon is' das Waschbecken angesiedelt. *and behind it, basically to the left beside it, the washbasin is located.*	ACTION-DIRECTIVE
D311-42	Matcher	das Waschbecken mit dem Spiegel? *the washbasin with the mirror?*	INFO-REQUEST
D311-43	Director	exakt. *exactly.*	ANSWER(*D311-42*)
D311-44	Director	und mit diesem Handtuchhalter daneben. *and with this towel rail beside it.*	ACTION-DIRECTIVE ASSERT
D311-45	Matcher	ja genau. *yes, exactly.*	ACCEPT(*D311-44*)
D311-46	Matcher	und das kommt ähm wohin? *and ah where to put it?*	INFO-REQUEST
D311-47	Director	in die Lücke, in die Lücke zwischen den beiden Wänden an der linken Hauswand, unter das Dach praktisch. *in the gap, in the gap between the two walls on the left outside wall, under the roof basically.*	ANSWER(*D311-46*), ACTION-DIRECTIVE
D311-48	Director	okay? *okay?*	INFO-REQUEST
D311-49	Matcher	ach ach so in die Lücke da. *oh oh so into the gap there.*	ASSERT OFFER
D311-50	Director	genau, in die Lücke. *right, into the gap.*	ACCEPT(*D311-49*) ACTION-DIRECTIVE
D311-51	Matcher	wo die Wand is'? *where the wall is?*	INFO-REQUEST
D311-52	Director	exakt, zwischen die beiden Wände. *exactly, between the two walls.*	ANSWER(*D311-51*) ACTION-DIRECTIVE

Table 6: Example dialogue from the Dollhouse Corpus.

course subgoal. Rather than asking an open question, the Matcher makes an informed guess, which is confirmed in *D311-38*. Then the Director establishes common ground via an assertion about the current status (*D311-39*). This serves as the basis for the actual object placement description in *D311-41*. *D311-42* reveals uncertainty on the part of the Matcher concerning the identity of the object to be placed; again the Matcher makes an informed guess, which is confirmed and further enhanced by the Director in *D311-43* and *D311-44*. This subdialogue about the object's identity is closed by the Matcher's acceptance in *D311-45*, who then returns in *D311-46* to the previously started negotiation of the object's placement. Again, the Matcher is actively involved by partially repeating (thereby confirming) instructions, or by making informed guesses as in *D311-51*. The Director, in turn, elicits confirmation that the Matcher is still on track, as in *D311-48*.

This high level of engagement of the Matcher may be traced back to two specific features of the scenario: that (1) the Matcher actively needs to manipulate real world objects when furnishing the dollhouse, which requires sufficient understanding of the instruction to perform the associated action, rather than just identifying one of various possibilities without further consequences in the given task setting; and that (2) each single object placement represents a complex subgoal within the overall task of furnishing the dollhouse (objects need to be identified, placed, and oriented correctly in relation to the other objects within the same house). This overall complexity provides ample opportunity for the Matcher to integrate his or her knowledge about previously reached subgoals with the actual goal of placing a particular object.

Usability for task-oriented scenarios. Similar to the Airplane scenario discussed in the previous subsection, the Dollhouse scenario allows for the less informed person (the Matcher) to contribute to the continuous update of common ground by making relevant suggestions. In particular, their perception of the scene allows for a number of spatial inferences to be integrated towards a reasonable interpretation of the Director's intentions. Paralleling our observations for Schober's scenario in subsection 3.3, the dialogue-act-based annotation captures the underlying structural elements on a superficial level, but does not actually account for the perception-based negotiation efforts contributed by the interaction partners. To capture this latter aspect, their different perceptions on the spatial scene need to be considered.

3.6 NavSpace: Actions as dialogic contributions

The NavSpace corpus (Tenbrink et al., 2010) serves as an example of the effects of the Instructor's ability to perceive task-relevant actions performed by the Instructee. The interlocutors were guiding a wheelchair avatar through a schematized spatial environment shown on a computer screen. Only the Instructor knew about the goal position of the avatar, and only the Instructee could move the avatar, using a joystick. The Instructee was either a human or a dialogue system; Instructors were always humans. Communication was achieved via a text-based chat interface. Participants in the human-human condition shared the same view on the scene including the avatar's position, although they were not co-located and could, therefore, not use gestures or facial expressions to communicate intent. While the dialogue system was only equipped to respond (affirmatively) by saying "OK" and moving the avatar, or (in the case of problems) by asking a clarification question or rejecting the instruction, the human Instructee's responses were more complex and varied.

In contrast to all other scenarios described so far, this setting allows for actions as responses to instructions, which are frequently found in the NavSpace corpus, particularly in the human-human interaction situation. Unlike the dialogue system, humans did not regularly say "OK" before moving

Id	Speaker	Action	Utterance	Dialogue act
10-52-16	Instructor		weiter gerade aus *continue straight ahead*	ACTION-DIRECTIVE
10-52-17	Instructee	action		ACCEPT(*10-52-16*)
10-52-26	instructor		stopp *stop*	ACTION-DIRECTIVE
10-52-35	Instructee	action		OFFER
10-52-58	Instructor		rückwärts zum vorigen raum links *backwards to the previous room on the left*	REJECT(*10-52-35*) ACTION-DIRECTIVE
10-53-2	Instructee	action		ACCEPT(*10-52-58*)
12-48-26	Instructor		eine weiter *one further*	ACTION-DIRECTIVE
12-48-30	Instructee		bis zur ersten oder zweiten biegung? *until the first or the second turn?*	INFO-REQUEST HOLD(*12-48-26*)
12-48-32	Instructee	action		ACCEPT(*12-48-26*)
12-48-35	Instructee		ok	ACCEPT(*12-48-26*)
12-48-35	Instructee	action		ACCEPT(*12-48-26*)
12-48-36	Instructor		zweite *second*	ACTION-DIRECTIVE ANSWER(*12-48-30*)
12-48-56	Instructee	action		ACCEPT(*12-48-36*)

Table 7: Example dialogues from the NavSpace Corpus.

the avatar, and human Instructors interrupted actions if the move direction was incorrect. Neither DAMSL nor the HCRC annotation scheme cover action in lieu of (verbal) dialogic contributions since they were created for different kinds of corpora and research questions. Therefore, in order to annotate dialogue corpora with actions, some extensions are required. Here we propose to introduce an action layer parallel to the task layer in DAMSL, which (like language-based dialogue acts) may serve forward-looking and backward-looking functions. The dialogue acts conveyed by actions are less rich than in language-based utterances. The most frequent action acts with forward-looking functions are OFFER and REQUEST, and with backward-looking functions ACCEPT and HOLD. Table 7 gives examples.

In the first example in Table 7, instructions given in language (*10-52-16* and *10-52-58*) are accepted via actions (*10-52-17* and *10-53-2*, respectively). Furthermore, the Instructee uses an action to provide an offer in *10-52-35*, which is rejected by the Instructor in *10-52-58*. In the second example in Table 7, a request for additional information (*12-48-30*) is posed in parallel to a movement action that accepts the previous instruction (*12-48-32*). In *12-48-35*, acceptance is signaled in parallel via language and action.

Usability for task-oriented scenarios. This scenario illustrates the importance of actions as proper contributions to the dialogue. As shown in the example dialogues in Table 7, action and language are frequently interleaved temporally, and they provide meaningful information to advance the dialogue individually or jointly. Generally, if actions are perceptually accessible to both dialogue partners (independent of whether they are physically located in the same room), actions contribute directly to the dialogue processes and structures, just as utterances do. Actions can be used directly to update common ground, via reactions that are appropriate at a specific state in the dialogue. Furthermore, both speech and actions constitute common ground and can be referred to in later utterances by both

dialogue participants. Annotation schemes clearly need to account for these effects, and we have provided here a first suggestion for how this can be accomplished in the DAMSL scheme.

3.7 CReST: Lexical and intonational expressions of uncertainty

The Indiana Cooperative Remote Search Task (CReST) Corpus (Eberhard et al., 2010) is a corpus of approximately 8 minute dialogues recorded from 16 dyads performing a cooperative search task in which one person (Director), who was located in a room away from the search environment, remotely directed the other person (Searcher) through the environment by a hands-free telephone connection. The environment consisted of six connected rooms leading to a long winding hallway. Neither the Director nor the Searcher was familiar with the environment prior to the task. The Director guided the Searcher through the environment with a map. The environment contained a cardboard box, 8 blue boxes, each containing three colored blocks, 8 pink boxes, and 8 green boxes. The locations of the cardboard box, blue boxes, and pink boxes were shown on the map; however, 3 of the 8 blue boxes' locations were inaccurate, and the participants were informed of this. At the beginning of the experimental session, the Director and Searcher were told that the Searcher was to retrieve the cardboard box and empty the blocks from the blue boxes into it. The Searcher also was to report the locations of the green boxes to the Director, who was to mark them on the map. They were told that instructions for the pink boxes would be given sometime during the task. Five minutes into the task, the Director and Searcher were interrupted and the Director was told that each of the 8 blue boxes contained a yellow block, and the Searcher was to place a yellow block into each of the eight pink boxes. In addition, they had three minutes in which to complete all of the tasks. A timer that counted down the remaining time was placed in front of the Director. The dyads' performance was scored with respect to the number of boxes out of 24 for which the designated task was completed. The average score was 12 with a range of 1-21. The dyads' verbal interactions were recorded along with video recordings of the Searcher's movement through the environment and the Director's marking the map with the green boxes. The verbal interactions were orthographically transcribed and annotated for conversational moves using the HCRC scheme.

In the CReST corpus, the Director's map, which was a two-dimensional floor plan, was a less reliable source of knowledge about the task domain (search environment) than the Searcher's direct perceptual experience of that domain. The disparity in the reliability of the knowledge was reflected in the Directors producing twice as many requests for information (i.e., QUERY-YN, QUERY-W, CHECK, ALIGN) than the Searchers. In addition, about a third of the Directors' unsolicited descriptions of new elements in the environment, coded as EXPLAIN moves, conveyed uncertainty via hedging expressions (e.g., "I think", "it looks like", "there should be", etc.). These utterances were identified by coding them as EXPLAIN/HEDGED. Examples are given in the top half of Table 8.

Notice that the Searcher's acknowledgments to the EXPLAIN/HEDGED moves were affirmatives (i.e., "yes", "right"), which confirm the accuracy of the common ground. There also were instances in which the Directors' EXPLAIN moves appeared to convey uncertainty by ending with rising intonation, similar to questioning intonation. The location of the rising intonation was indicated with a question mark in the transcriptions. An example is shown in the bottom half of Table 8. Specifically, *utt7* was coded as EXPLAIN/QUERY-YN because its declarative form and its unsolicited description of new elements in the environment are consistent with the EXPLAIN code. However, the final rise in intonation was consistent with a request for confirmation of the accuracy of the description, or the QUERY-YN code. Ordinarily, the latter code would be assigned, but a consideration of the larger

ID	Speaker	Utterance	Dialogue act
utt43	D3	okay	READY
utt44	D3	and straight in front of you should be: filing cabinets	EXPLAIN/HEDGED
utt45	S3	yes	ACKNOW
utt44	D3	okay	READY
utt46	D3	so: between the second cubicle on the right and the filing cabinets there should be: kind of like a space to walk through	EXPLAIN/HEDGED
utt47	S3	right	ACKNOW
utt48	D3	so go though there	INSTRUCT
utt49	S3	kay	ACKNOW
utt5	D4	and through the first door	INSTRUCT
utt6	S4	okay	ACKNOW
utt7	D4	and you'll come to like a platform with some steps?	EXPLAIN/QUERY-YN
utt8	S4	yes	ACKNOW
utt9	D4	and you're gonna wanna turn to the right?	INSTRUCT
utt10	S4	yes	ACKNOW
utt11	D4	and go straight ahead through that door	INSTRUCT
utt12	S4	yes	ACKNOW

Table 8: Example dialogues from the CReST Corpus for EXPLAIN/HEDGED and EXPLAIN/QUERY-YN moves.

context made the interpretation of the final rise in intonation ambiguous. Specifically, like *utt7*, *utt9*, which is an INSTRUCT move, ended with the same rising intonation, whereas *utt11*, which is also an INSTRUCT move, ended with falling intonation. This pattern is consistent with "list intonation", which occurred when Directors or Searchers gave a complex instruction or description in installments. In the example in the table, *utt7 - utt12* constitute a segment of a larger dialogue in which the Director directs the Searcher through three connected rooms and down a hallway to retrieve a cardboard box. The first two utterances, *utt5* and *utt6*, end the first segment in which the Searcher was directed from the first room to the second room. Thus, *utt7 - utt12* involved directing the Searcher to the third room. Like a yes-no question, the rising intonation at the end of each installment is followed by a pause for an acknowledgment from the addressee. Like backchannels, the acknowledgments produced in this context perform a continuer function, i.e., *I hear you, please continue* (e.g., Gardner, 2002). Furthermore, the typical forms for this function are "mhm", "uh huh", and "yeah", which also are associated with a confirmatory function. Like the examples of EXPLAIN/HEDGED moves in the top half of the table, the EXPLAIN/QUERY-YN example is followed by the affirmative acknowledgment "yes", which may indicate the Searcher's sensitivity to the possible request.

Usability for task-oriented scenarios. The EXPLAIN/HEDGED and EXPLAIN/QUERY-YN moves in the CReST corpus illustrate the reliance on dialogue for updating common ground and coordinating joint actions in a scenario where the interactants communicated remotely. In the case of the EXPLAIN/HEDGED move, the Director's hedged description of an aspect of the environment reflected his or her reliance on a map, which was a less reliable source of information compared to the Searcher's direct perceptual experience of the environment. The EXPLAIN/QUERY-YN move demonstrates how intonation can be ambiguous with respect to whether it reflects the communica-

tive action being performed (i.e., a request for confirmation) or the dialogue structure (i.e., an installment in a sequence of moves for completing a segment of the task, c.f. with "uptalk"). This distinction might be captured in a layer of finer-grained annotation of the intonation, such as the ToBI labeling scheme (Beckman and Hirschberg, 1994) (see below). However, regardless of whether the Director's rising intonation was intended to be a request for confirmation, the Searcher's affirmative acknowledgment provided this confirmation, allowing the common ground to be updated accordingly (e.g. Safarova, 2006).

3.8 Intonation and turn-taking in the Columbia Games Corpus

The Columbia Games Corpus (Gravano, 2009) is a corpus of 12 dyadic conversations recorded during two collaborative tasks, namely games played on separate computer screens. The games involved locating cards based on spoken communication, and lasted on average about 45 minutes. The participants were seated in a room, facing each other so that their own computer screen could not be seen by the other participant. Additionally, there was an opaque curtain separating them so that they could not see each other. In such a scenario, speakers have to rely on spoken language and intonational cues to keep track of the conversation and manage common ground while constantly recurring to the perceptual information shown on their separate screens.

Ford and Thompson (1996) showed that, while most intonationally complete utterances are also syntactically complete, about one half of syntactically complete utterances are intonationally incomplete, signalling the continuation of an ongoing speaker turn. This demonstrates the prominent role of intonation for dialogue structure. Gravano (2009) aimed at identifying the precise ways in which these turn taking processes operate on the basis of subtle intonational cues conveyed during speech. The annotation in the Columbia Games Corpus includes self-repairs, non-word vocalization (laughs, coughs, breaths, and the like), and a detailed intonation analysis using the ToBI labeling scheme (Beckman and Hirschberg, 1994). ToBI consists of four tiers: an *orthographic* tier, which includes the orthographic transcription of the recording, a *break index* (BI) tier, a *tonal* tier, and a *miscellaneous* tier, which can be used to mark disfluencies, etc. The break index tier annotates the end of each word for the strength of its association with the next word. The tonal tier describes a phonological analysis of the utterance's intonation pattern. This annotation use two distinct tones, H(igh) and L(ow), with additional diacritics to mark pitch accents or downstep.

The corpus was also annotated for turn-taking based on categories suggested by Beattie (1982). The annotation scheme distinguishes between overlap, interruption, butting-in, smooth switch, pause interruption, backchannel with overlap, and backchannel. According to this analysis, speakers were more likely to switch turns following subtle speaker cues such as a lower intensity level, a lower pitch level, and a point of textual completion.

Table 9 shows an example[4] with ToBI annotation, where speaker B uses contrastive stress to draw attention to two distances between symbols, the distance between the ruler and the blue crescent, and between the blue and the yellow crescent.

Usability for joint action scenarios: As illustrated in previous subsections, intonation patterns can prove crucial for the updating of common ground. The ToBI annotation scheme provides a systematic solution for capturing intonation and turn-taking aspects. Although (to our knowledge) it has not been used to identify common ground updating processes, it stands to reason that speakers

4. We thank A. Gravano for providing the example.

speaker B: but there's
BI: 1
tone: H*

speaker A: #
BI: 3p

speaker B: more space between the ruler and the blue crescent # than there is the blue crescent and the yellow crescent #
BI: 1 1 3 1 1 3 1 1 1 4 1 1 3 1 1 3 1 1 1 4
tone: H* !H- L+H* H- H* H* H* H- H* H- L* L-L%

speaker A: huh #
BI: 4
tone: H* L-L%

Table 9: Example from the Columbia Games Corpus.

use intonational clues to conclude when common ground has been reached sufficiently for current purposes, or when they need to step in so as to ask a clarification question, leading to the turn-taking patterns identified by Gravano (2009).

3.9 Basic gesture annotation

McNeill (2000) developed an annotation scheme for gestures which was used for the Rapport Corpus[5]. This corpus contains face-to-face dialogues with dyads of people who know each other well. In an example given in McNeill (2002), one person watched a Tweety and Sylvester cartoon and had to tell the story to another person. Both participants had been told that the listener would have to tell the story to yet another person. Gestures were not mentioned in the instructions. Table 10 shows the dialogue and gesture annotation. Square brackets indicate the beginning and end of the gesture, and boldface marks the gesture stroke – "the phase with semantic content and the quality of effort" (McNeill, 2002, footnote 11). The annotations specify whether one hand or both are used, and whether the gesture was symmetrical or asymmetrical in the latter case. It also describes the movement (e.g., "move down") and the number of repetitions (e.g., 2X). In the example, the speaker mostly uses gestures to illustrate motion events. In utterance (1), for example, "going up the inside" is accompanied by an upward gesture of the right hand.

Another approach to gesture annotation called FORM was used for video recordings of monologues by Martell (2002). The annotation, in the form of annotation graphs, captures different body parts in a complex procedural model (see Figure 1). FORM uses fine-grained labels for gestures such as "UpperArm.Location" and "HandandWrist.Movement". Thus, it captures a wider range of movements in a more standardized way than the scheme proposed by McNeill (2000).

Usability for task-oriented scenarios. Gestures often accompany utterances to provide supporting or additional information, which is used to establish common ground. The annotation schemes for gestures exemplified here capture the nature of gestures, which is an essential first step. However, no information can be derived about the updating function in the dialogue. Establishing the role of gestures in the context of dialogue is central for situated interaction scenarios that incorporate relevant gestures.

5. `http://mcneilllab.uchicago.edu/corpora/rapport_corpus.html`

Id	Utterance	Gesture
(1)	he tries going [**up** the in**side** of **the drai**npipe and]	1hand: RH rises up 3X
(2)	Tweety Bird runs and gets a bowling ba[ll and drops **it do**wn the drainpipe #]	Symmetrical: 2 similar hands move down
(3)	[and / a**s he s** coming **u**p]	Asymmetrical: 2 different hands, LH holds, RH up 2X
(4)	[and the **bowling b**all s coming d]]	Asymmetrical: 2 different hands, RH holds, LH down
(5)	[own he **sssw**allows it]	Asymmetrical: 2 different hands, RH up, LH down
(6)	[# and he comes **out the bot**tom of the drai]	1hand: LH comes down
(7)	[npipe and he s **got thi**s big bowling ball inside h]im	Symmetrical: 2 similar hands move down
(8)	[and h**e rolls on down**] [into **a bow**ling all]	Symmetrical: 2 similar hands move forward 2X
(9)	[ey and then **you hear a** sstri]ke #	Symmetrical: 2 similar hands move apart

Table 10: Example dialogue from McNeill (2002).

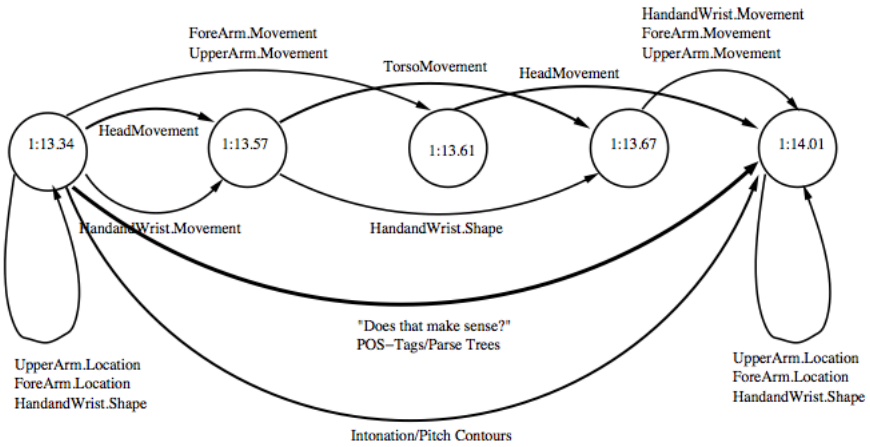

Figure 1: An example for gesture annotation in FORM (Martell, 2002).

3.10 Route gestures

Striegnitz et al. (2009) examined the use of gestures to support route directions, focusing on gestures associated with landmarks. In a close examination of five route dialogues, they identified gestures that indicated route perspective (i.e., the perspective of the person following the route), gestures consistent with a map-based survey-perspective, gestures which locate the object with respect to the speaker's actual position and orientation, and gestures that indicate the landmark's shape. All of these clearly contribute valuable spatial information, enriching the spoken language. For instance, the utterance "and it's really big" is accompanied by a gesture that indicates the landmark's *horizontal* extent. This information is not included in the dimension-neutral size term "big".

Striegnitz et al. (2009) coded the route dialogues using a DAMSL type annotation scheme, identified the gesture information separately, and related gesture types to utterance categories. Most of the gestures accompanied statements: typically those that mentioned a landmark, but also some that did not. These were further specified with respect to their role in the dialogue, such as plain statements or those that serve the function of a response, a query, an elaboration, or a redescription.

Usability for task-oriented scenarios. Although no further specification of the annotation scheme was proposed by Striegnitz et al. (2009) (nor were any dialogue examples given), the proposed scheme still demonstrates the necessity to systematically account for speech-accompanying gestures, as they play an important role in the updating of common ground.

3.11 Rolland: multimodal interaction with a mobile robot

The Rolland multimodal interaction study (first cited in Anastasiou, 2012)[6] addressed the interaction between a powered wheelchair called *Rolland* and a User who was asked to carry out a set of simple tasks with Rolland. Participants (i.e., Users) were told that Rolland could understand spoken language as well as gestures by hands and arms, but would only react through driving actions. The study was a "Wizard-of-Oz" setup, i.e., the wheelchair was remote-controlled by a human operator, unbeknownst to the participants.

Table 11 includes an example dialogue in which the User asks Rolland to come to the bedside, so that he could get into the wheelchair without much effort. We provide a DAMSL-type annotation here. The User starts by using spoken language only (until *utt140*), but remains dissatisfied with Rolland's reactions. Consequently, he adds gestures (*utt141, utt143* and *utt145*) that enhance the spoken utterances, so as to instruct Rolland more pointedly. Although still unhappy with the result, he finally accepts it in *utt147*.

Interestingly, at one point in the interaction, the gesture is inconsistent with the spoken utterance. In *utt143*, the User asks Rolland to drive a little bit backward. Along with this, however, the hand points to the goal location beside the bed – seemingly contradicting the spoken command. Arguably, the utterance refers to a more specific level of granularity (the immediate action to be performed) than the gesture, which points to the higher-level goal location. As a result, Rolland's action in *utt144* is ambiguous with respect to the performed dialogue act. On the one hand, the pointing gesture in *utt143* is accepted by the action of driving towards the bed. The spoken utterance in *utt143* (driving backwards), on the other hand, is rejected by the same action.

6. This study was carried out by Dimitra Anastasiou and Daniel C. Vale in the Collaborative Research Center *SFB/TR 8 Spatial Cognition* in Bremen, funded by the DFG. Thanks to the authors for allowing us to gain insight into their work, and cite aspects central to our current focus.

Id	Part.	Utterance	Action	Gesture	Dialogue act
utt132	User	Rolland komme bitte zum Bett hierhin *Rolland please come here to the bed*			ACTION-DIRECTIVE
utt133	User	hier wo ich sitze *here where I'm sitting*			ACTION-DIRECTIVE
utt134	Rolland		driving to bed		ACCEPT(*132,133*)
utt135	User	er ist mir auf den Fuß gefahren, okay *he drove over my foot, okay*			ASSERT ACCEPT(*134*)
utt136	User	komme etwas näher zum Bett *come a bit closer to the bed*			ACTION-DIRECTIVE
utt137	Rolland		driving to bed		ACCEPT(*136*)
utt138	User	noch etwas näher zum Bett *once more closer to the bed*			ACTION-DIRECTIVE
utt139	Rolland		driving to bed		ACCEPT(*138*)
utt140	User	weiter zu mir *further to me*			ACTION-DIRECTIVE
utt141	User	etwas näher zu mir *a bit closer to me*		two-handed configuration open hand shapes	ACTION-DIRECTIVE
utt142	Rolland		driving to bed		ACCEPT(*140,141*)
utt143	User	ein bisschen zurück fahren *drive a little bit backward*		pointing with index finger	REJECT-PART(*142*) ACTION-DIRECTIVE
utt144	Rolland		driving to bed		*ambiguous*
utt145	User	Rolland ein bisschen zurück fahren *Rolland drive a little bit backward*		pointing with index finger	REJECT(*144*) ACTION-DIRECTIVE
utt146	Rolland		driving backward		ACCEPT(*145*)
utt147	User	okay ich versuche mal so *okay I try like this*			ACCEPT(*146*)

Table 11: Example dialogue from the Rolland multimodal interaction corpus.

Usability for joint action scenarios: Although this scenario is clearly restricted by the limited interaction capabilities of the wheelchair, the example illustrates the tightly integrated yet independent role of gesture, just as in human face-to-face interaction. Gestures can elaborate and expand language to establish common ground, address different aspects, or appear as incongruent with the verbally conveyed content, leading to further dialogue structure complexities. Interlocutors may ignore gesturally conveyed content, or react verbally or non-verbally. Annotation schemes need to account for these procedures. Our DAMSL-based annotation example provides a first suggestion of how this might be accomplished.

4. Layers of Annotation

The above review of insights gained in joint task settings highlighted a range of factors that are crucial for the negotiation of shared goals in situated dialogue, showing how speakers manage to update their knowledge of the current state of the task in their common ground. In the following, we propose four potential additional layers for annotation schemes (depending on the research question at hand). Generally, for each layer, an aspect can be a *direct* contribution to dialogue if it is accessible to both interactants to the same degree, while it affects the dialogue more indirectly if such access is not shared. For example, an action can only serve as a direct response to a directive if this is also perceived by the director; if the action is not perceived as such, the actor will typically acknowledge the directive verbally in addition to acting upon it, so as to update the director's state of knowledge. Along these lines, sharedness of these layers turns out to be a major factor in any analysis of dialogue. Although we were able to use existing schemes like DAMSL to express some of these effects, there is a clear need for further extensions of annotation formalisms to be able to represent the intricate interplay between linguistic and non-linguistic interactions in joint action scenarios.

4.1 Intonation

As demonstrated by several examples in the previous subsections, intonation plays an important role in common ground updating processes. Speakers use intonational cues to determine when meaningful fragments in an utterance have been completed or when clarification questions may be asked, thus aiding turn-taking and contributing to dialogue structure. Intonation can also highlight meanings, convey the significance of an utterance, and provide feedback about the acknowledgment or rejection of a previous utterance. Critically, intonational cues are used and picked up automatically by interactants and directly affect the pragmatic implicatures and the dialogue flow. Intonational cues thus play a prominent role in spoken dialogues. When speakers cannot see each other they can compensate for missing cues conveyed by gestures or facial expressions.

We are not aware of any current dialogue scheme that combines dialogue structure annotation with intonation patterns such as those identified by the ToBI annotation scheme. Conceivably, the annotation of intonation can be pragmatically reduced to the most relevant aspects that directly affect meaning interpretations and dialogue structure. This would involve adding a further layer to an existing dialogue structure annotation scheme, with the possibility to override meaning interpretations and dialogue moves in other layers (e.g., what might otherwise be coded as "acknowledgement" could end up as a "YN-question" based on intonational information).

4.2 Gestures

Gestures can substantially supplement verbal information. This is most clearly demonstrated when they contribute spatial information, for instance in referential phrases, thereby substantially enhancing the common ground updating process by contributing aspects that may not be verbalised at all. Similar processes are active whenever speakers have visual access to their interaction partner such as in face-to-face communication. Dialogue structure annotation schemes can be straightforwardly enriched by an additional layer capturing gestures, as exemplified in Table 11 above. The level of granularity of gesture annotations will depend on the research issue at hand.

4.3 Perception of the task domain

Situated tasks involve perceptual access to the task domain. Even if perception is not shared, the task domain information that each of the speakers has access to is central to the coordination of actions and accumulation of common ground. In the Dollhouse scenario in subsection 3.5, for instance, the Matcher is able to make informed guesses about positions of objects because verbal instructions can be compared to the arrangement of objects in the perceivable scene. Similarly, many verbal contributions in the CReST scenario (subsection 3.7) directly build on the non-joint perceptions of both interlocutors and thereby affect dialogue structure. While the results of these effects are captured by dialogue structure annotations, the procedures as such can only be fully understood if the relevant perceptions are also taken into account.

We suggest adding a layer of scene perception to the annotation that can be used to capture relevant perceptual aspects that speakers draw from. In this way, the functions of particular dialogue acts can be interpreted more reliably. Moreover, in the case of non-shared scene perception, cases of miscommunication can be better identified and accounted for based on the discrepancy between the interactants' access to the task domain.

4.4 Actions

The ability to perceive task-relevant actions provides reliable and timely information for updating common ground about the status of the task. The NavSpace example in subsection 3.6 illustrates how using the same categories for coding task-relevant actions and dialogue structure captures the complementary role of actions and dialogue moves in the process of updating common ground. The interaction between actions and dialogue becomes more complicated when speakers are engaged in a joint task without directly sharing perceptual access to action outcomes, as exemplified by the CReST example (subsection 3.7). In such scenarios, speakers will communicate action outcomes in some cases, but assume that they can be inferred in other cases. As a result, common ground representations among interactants may start to diverge and become inconsistent, which will eventually result in dialogue interactions solely dedicated to resolving the inconsistencies and re-establishing common ground.

To account for these effects and reliably interpret the function of dialogue acts (e.g., to re-establish coordination and common ground), we recommend keeping track of the speakers' actions by adding a corresponding layer to the dialogue annotation. The specification of this layer (e.g., richness of action descriptions, temporal extension, etc.) will depend on the purpose of the task and dialogue analysis.

5. Conclusion

In this paper, we reviewed existing dialogue corpora across various interaction settings to investigate the different linguistic and non-linguistic aspects that affect how interactants negotiate joint actions and update their common ground in task-based dialogues. We specifically examined existing analyses and established annotation schemes in the literature to determine the extent to which they are able to identify and capture relevant aspects of action negotiation and updating of common ground. This is particularly important as interactants in joint activities will make use of any information, including perceptions and action available to them (e.g., perceptions about the task domain, gestures by other interactants, or actions on task-relevant objects).

Current annotation schemes typically fail to account for these features of situated task scenarios. In other words, while Clark's influential work (Clark, 1996; Clark and Wilkes-Gibbs, 1986) has led to a widely acknowledged view of dialogue as joint activity that involves more than just the dialogue interactions, this recognition is not yet systematically or coherently reflected in dialogue annotation schemes. For information-seeking dialogues, relevant insights have been gained about clarification phenomena (e.g., Purver et al., 2003; Rieser and Moore, 2005). However, dialogue structure analysis for situated task scenarios is more complex, as dialogue patterns differ substantially from those identified in purely language-based interaction settings.

A better understanding of the intricate processes involved in human-human dialogues as part of situated joint activities is not only central to a better understanding of human natural language interactions, but also critical for research in human-computer or human-robot interaction (e.g., Alexandersson et al., 1998; Green et al., 2006; Allen et al., 2001; Shi et al., 2010). To pursue this line of research, rich annotations of dialogue corpora are required that, in addition to linguistic annotations, include interlocutors' perceptions, intonation, gestures (where appropriate), actions, and any other relevant factors that contribute to building up and negotiating common ground. Only with these additional annotations will it be possible to determine and build computational models of the intricate interplay between linguistic and non-linguistic aspects in task-based dialogues.

6. Acknowledgments

This work was in part funded by the DFG, SFB/TR 8 Spatial Cognition, project I5-[DiaSpace], to the first and third author, and by ONR MURI grant #N00014-07-1-1049 to the second and last author. We also wish to thank Elena Andonova, John A. Bateman, Kenny R. Coventry, Nina Dethlefs, Juliana Goschler, Cui Jian, Robert J. Ross, and Kavita E. Thomas for collaboration on dialogue corpora and related issues, and Christoph Broschinski for relevant inspiration. We are also grateful to the anonymous reviewers, who made constructive suggestions that helped us to improve and focus this paper.

References

Jan Alexandersson, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Michael Kipp, Stephan Koch, Elisabeth Maier, Norbert Reithinger, Birte Schmitz, and Melanie Siegel. Dialogue Acts in VERBMOBIL-2 (Second edition). Verbmobil Report 226, University of the Saarland, Saarbrücken, Germany, 1998.

James Allen and Mark Core. Draft of DAMSL: Dialogue markup in several layers. Technical report, University of Rochester, 1997. URL `http://www.cs.rochester.edu/research/speech/damsl/RevisedManual/`.

James Allen and C. Raymond Perrault. Analyzing intention in utterances. *Artificial Intelligence*, 15 (3):143–178, 1980.

James Allen, Donna Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu, and Amanda Stent. Toward conversational human-computer interaction. *AI Magazine*, 22(4):27–37, 2001.

James F. Allen, George Ferguson, Brad Miller, and Eric Ringger. TRAINS as an embodied natural language dialog system. In *Embodied Language and Action: Papers from the 1995 Fall Symposium*. AAAI Technical Report FS-95-05, 1995.

Jens Allwood, Stefan Kopp, Karl Grammer, Elisabeth Ahlsén und Elisabeth Oberzaucher, and Markus Koppensteiner. The analysis of embodied communicative feedback in multimodal corpora: a prerequisite for behavior simulation. *Journal on Language Resources and Evaluation*, 41(2-3):325–339, 2007. Special Issue on Multimodal Corpora for Modeling Human Multimodal Behaviour.

Dimitra Anastasiou. A speech and gesture spatial corpus in assisted living. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, 2012.

Anne Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jaqueline Kowtko, Jan McAllister, Jim Miller, Cathy Sotillo, Henry Thompson, and Regina Weinert. The HCRC Map Task Corpus. *Language and Speech*, 34(4):351–366, 1991.

Adrian Bangerter. Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science*, 15(6):415–419, 2004.

Ellen Gurman Bard, Catherine Sotillo, Anne H. Anderson, Henry S. Thompson, and Martin M. Taylor. The DCIEM Map Task Corpus: Spontaneous dialogue under sleep deprivation and drug treatment. *Speech Communication*, 20(1-2):71–84, 1996.

Ellen Gurman Bard, Anne H. Anderson, Yiya Chen, Hannele B.M. Nicholson, Catriona Havard, and Sara Dalzel-Job. Let's you do that: Sharing the cognitive burdens of dialogue. *Journal of Memory and Language*, 57(4):616–641, 2007.

Geoffrey Beattie. Turn-taking and interruption in political interviews: Margaret Thatcher and Jim Callaghan compared and contrasted. *Semiotica*, 39(1/2):93–114, 1982.

Mary Beckman and Julia Hirschberg. The ToBI annotation conventions. Technical report, The Ohio State University, 1994.

Mary Beckman, Julia Hirschberg, and Stefanie Shattuck-Hufnagel. The original ToBI system and the evolution of the ToBI framework. In Sun-Ah Jun, editor, *Prosodic Models and Transcription: Towards Prosodic Typology*. Oxford University Press, 2005.

Peter Bohlin, Robin Cooper, Elisabet Engdahl, and Staffan Larsson. Information states and dialog move engines. *Electronic Transactions in AI*, 3(9), 1999.

Elizabeth A. Boyle, Anne H. Anderson, and Alison Newlands. The effects of visibility on dialogue and performance in a cooperative problem solving task. *Language and Speech*, 37(1):1–20, 1994.

Holly P. Branigan, Robin J. Lickley, and David McKelvie. Non-linguistic influences on rates of disfluency in spontaneous speech. In *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, CA, 1999.

Susan E. Brennan, Xin Chen, Christopher A. Dickinson, Mark B. Neider, and Gregory J. Zelinsky. Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition*, 106(3):1465–1477, 2008.

Susan E. Brennan, Gregory J. Zelinsky, Joy E. Hanna, and Kelly J. Savietta. Eye gaze cues for coordination in collaborative tasks. In *DUET 2012: Dual Eye Tracking in CSCW*, Seattle, WA, 2012.

Harry Bunt. The DIT++ taxonomy for functional dialogue markup. In *Proceedings of the AMAAS 2009 Workshop Towards a Standard Markup Language for Embodied Dialogue Acts*, Budapest, Hungary, 2009.

Jean Carletta, Stephen Isard, Amy Isard, Gwyneth Doherty-Sneddon, Jacqueline Kowtko, and Anne Anderson. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23 (1):13–31, 1997.

Herbert H. Clark. *Using Language*. Cambridge University Press, 1996.

Herbert H. Clark and Deanna Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22: 1–39, 1986.

Philip R. Cohen and Hector J. Levesque. Speech acts and the recognition of shared plans. In *Proceedings of the Third Biennial Conference, Canadian Society for Computational Studies of Intelligence*, Victoria, BC, Canada, 1980.

Mark G. Core and James F. Allen. Coding dialogs with the DAMSL annotation scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*. AAAI Press, Cambridge, MA, 1997.

Kathleen Eberhard, Hannele Nicholson, Sandra Kübler, Susan Gunderson, and Matthias Scheutz. The Indiana "Cooperative Remote Search Task" (CReST) Corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, Valetta, Malta, 2010.

Kerstin Fischer and Thora Tenbrink. Video conferencing in a transregional research cooperation: Turn-taking in a new medium. In Jana Döring, H. Walter Schmitz, and Olaf Schulte, editors, *Connecting Perspectives. Videokonferenz: Beiträge zu ihrer Erforschung und Anwendung*, Aachen, 2003. Shaker.

Cecilia Ford and Sandra A. Thompson. Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. In Elinor Ochs, Emanuel Schegloff, and Sandra A. Thompson, editors, *Interaction and Grammar*, pages 134–184. Cambridge University Press, 1996.

Rod Gardner. *When listeners talk: Response tokens and listener stance*. John Benjamins Publishing Co., Philadelphia, 2002.

Agustín Gravano. *Turn-Taking and Affirmative Cue Words in Task-Oriented Dialogue*. PhD thesis, Columbia University, 2009.

Anders Green, Helge Hüttenrauch, Elin Anna Topp, and Kerstin Severinson Eklundh. Developing a contextualized multimodal corpus for human-robot interaction. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, 2006.

Barbara J. Grosz and Candace L. Sidner. Attention, intentions and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

Joy E. Hanna and Susan E. Brennan. Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, 57(4):596–615, 2007.

Peter A. Heeman and James Allen. The TRAINS 93 dialogues. Technical report, Computer Science Department, The University of Rochester, 1995. URL `http://www.cs.rochester.edu/research/speech/trains.html`.

Peter A. Heeman and James Allen. Tagging speech repairs. In *ARPA Workshop on Human Language Technology*, pages 187–192, Plainsboro, NJ, 1994.

Dan Jurafsky, Liz Shriberg, and Debra Biasca. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical Report RT 97-02, Institute for Cognitive Science, University of Colorado at Boulder, 1997.

Adam Kendon. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26: 22–63, 1967.

Michael Kipp, Michael Neff, and Irene Albrecht. An annotation scheme for conversational gestures: How to economically capture timing and form. *Journal on Language Resources and Evaluation*, 41(3-4):325–339, 2007.

Jacqueline Kowtko, Amy Isard, and Gwyneth Doherty. Conversational games within dialogue. Technical Report HCRC/RP-31, Human Communication Research Centre, University of Edinburgh, 1993.

Staffan Larsson and David Traum. Information state and dialogue management in the TRINDI Dialogue Move Engine Toolkit. *Natural Language Engineering*, pages 323–340, 2000. Special Issue on Best Practice in Spoken Language Dialogue Systems Engineering.

Oliver Lemon, Alexander Gruenstein, and Stanley Peters. Collaborative Activities and Multitasking in Dialogue Systems. *Traitement Automatique des Langues (TAL)*, 43(2):131–154, 2002. Special issue on dialogue.

Craig Martell. FORM: An extensible, kinematically-based gesture annotation scheme. In *Proceedings of the International Conference on Spoken Language Processing*, Denver, CO, 2002.

Craig Martell and Joshua Kroll. Corpus-based gesture analysis: An extension of the FORM dataset for the automatic detection of phases in gesture. *International Journal of Semantic Computing*, 1 (4):521–536, 2007.

David McNeill. *Language and Gesture*. Cambridge University Press, Cambridge, 2000.

David McNeill. Gesture and language dialectic. *Acta Linguistica Hafniensia*, 34(1):7–37, 2002.

Mark B. Neider, Xin Chen, Christopher A. Dickinson, Susan E. Brennan, and Gregory J. Zelinsky. Coordinating spatial referencing using shared gaze. *Psychonomic Bulletin and Review*, 17(5): 718–724, 2010.

Claire O'Malley, Steve Langton, Anne Anderson, Gwyneth Doherty-Sneddon, and Vicki Bruce. Comparison of face-to-face and video-mediated interaction. *Interacting with Computers*, 8(2): 177–192, 1996.

Kristina Poncin and Hannes Rieser. Multi-speaker utterances and co-ordination in task-oriented dialogue. *Journal of Pragmatics*, 38:718–744, 2006.

Matthew Purver and Ruth Kempson. Incrementality, alignment and shared utterances. In *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue (Catalog)*, pages 85–92, Barcelona, Spain, 2004.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. On the means for clarification in dialogue. In Ronnie Smith and Jan van Kuppevelt, editors, *Current and New Directions in Discourse and Dialogue*, pages 235–255. Kluwer Academic Publishers, Dordrecht, 2003.

Gert Rickheit and Ipke Wachsmuth. *Situated Communication*. Mouton de Gruyter, Berlin, 2006.

Verena Rieser and Johanna D. Moore. Implications for generating clarification requests in task-oriented dialogues. In *Proceedings of the 43rd Annual Meeting of the ACL*, Ann Arbor, MI, 2005.

Marie Safarova. *Rises and falls: Studies in the semantics and pragmatics of intonation*. PhD thesis, University of Amsterdam, 2006.

Michael F. Schober. Spatial perspective taking in conversation. *Cognition*, 47(1):1–24, 1993.

Michael F. Schober. How addressees affect spatial perspective choice in dialogue. In Patrick L. Olivier and Klaus-Peter Gapp, editors, *Representation and Processing of Spatial Expressions*, pages 231–245. Lawrence Erlbaum Associates, Mahwah, New Jersey, 1998.

Michael F. Schober. Spatial dialogue between partners with mismatched abilities. In Kenny Coventry, Thora Tenbrink, and John Bateman, editors, *Spatial Language and Dialogue*, pages 23–39. Oxford University Press, 2009.

Michael F. Schober. Speakers, addressees, and frames of reference: Whose effort is minimized in conversations about location? *Discourse Processes*, 20(2):219–247, 1995.

Abigail J. Sellen. Remote conversations: The effects of mediating talk with technology. *Human-Computer Interaction*, 10:401–444, 1995.

Hui Shi, Robert J. Ross, Thora Tenbrink, and John Bateman. Modelling illocutionary structure: Combining empirical studies with formal model analysis. In A. Gelbukh, editor, *Proceedings of the 11th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2010)*, Lecture Notes in Computer Science, Berlin, 2010. Springer. March 21-27. Iasi, Romania.

Teresa Sikorski and James F. Allen. A task-based evaluation of the TRAINS-95 dialogue system. In *ECAI Workshop on Dialogue Processing in Spoken Language Systems*, pages 207–219. Springer, Budapest, Hungary, 1997.

Lesley Stirling, Janet Fletcher, Ilana Mushin, and Roger Wales. Representational issues in annotation: Using the Australian map task corpus to relate prosody and discourse structure. *Speech Communication*, 33:113 – 134, 2001.

Kristina Striegnitz, Paul Tepper, Andrew Lovett, and Justine Cassell. Knowledge representation for generating locating gestures in route directions. In Kenny Coventry, Thora Tenbrink, and John Bateman, editors, *Spatial Language and Dialogue*, pages 147–165. Oxford University Press, 2009.

Thora Tenbrink, Elena Andonova, and Kenny Coventry. Negotiating spatial relationships in dialogue: The role of the addressee. In *Proceedings of LONDIAL – The 12th SEMDIAL Workshop*, London, UK, 2008.

Thora Tenbrink, Robert J. Ross, Kavita E. Thomas, Nina Dethlefs, and Elena Andonova. Route instructions in map-based human-human and human-computer dialogue: A comparative analysis. *Journal of Visual Languages and Computing*, 21(5):292–309, 2010.

David R. Traum. Conversational agency: The TRAINS-93 dialogue manager. In Susann Luper-Foy, Anton Nijholt, and Gert Veldhuijzen van Zanten, editors, *Dialogue Management in Natural Language Systems*, pages 1–11. Universiteit Twente, Enschede, 1996.