Dialogue Structure Annotation for Multi-Floor Interaction

David Traum[1], Cassidy Henry[2], Stephanie Lukin[2], Ron Artstein[1]
Felix Gervitz[3], Kimberly A. Pollard[2], Claire Bonial[2], Su Lei[1],
Clare R. Voss[2], Matthew Marge[2], Cory J. Hayes[2], Susan G. Hill[2]
[1]USC Institute for Creative Technologies, [2]U.S. Army Research Laboratory, [3]Tufts University
Playa Vista, CA 90094, Adelphi, MD 20783, Medford, MA 02155
traum@ict.usc.edu

Abstract

We present an annotation scheme for meso-level dialogue structure, specifically designed for multi-floor dialogue. The scheme includes a *transaction unit* that clusters utterances from multiple participants and floors into units according to realization of an initiator's intent, and *relations* between individual utterances within the unit. We apply this scheme to annotate a corpus of multi-floor human-robot interaction dialogues. We examine the patterns of structure observed in these dialogues and present inter-annotator statistics and relative frequencies of types of relations and transaction units. Finally, some example applications of these annotations are introduced.

Keywords: dialogue structure annotation, human-robot interaction, multiparty dialogue

1. Introduction

We present an annotation scheme for meso-level dialogue structure (Traum and Nakatani, 1999), specifically designed for multi-floor dialogue. The scheme includes both a *transaction unit* for clustering utterances from multiple participants and floors that contribute to realization of an initiating participant's intent, and *relations* between individual utterances within the unit. While there are standard annotation schemes for both dialogue acts (Bunt et al., 2012) and discourse relations (Prasad and Bunt, 2015), these schemes do not fully address the issues of dialogue structure. Of particular interest to us, and not previously addressed in other schemes, are cases in which the units and relations span across multiple conversational floors. Dialogues can be characterized by distinct information states (Traum and Larsson, 2003). These include sets of participants, participant roles (e.g. active, ratified participant vs. overhearer), turn-taking or floor-holding, expectation of how many participants will make substantial contributions at a time (Edelsky, 1981), and other factors. Often distinct dialogues with different information states are going on at the same time. There are a number of ways in which such dialogues can be related to each other, including:

- having the same purpose but distinct participants, e.g., teams competing in a trivia contest to come up with the answer first.
- co-located such that participants in one can observe and possibly comment on the other, such as groups of people sitting at different tables at a restaurant.
- having one or more (but not all) participants in common, where such participants are *multicommunicating* (Reinsch et al., 2008), e.g., someone in a meeting is texting with one or more people outside the meeting.

In the multicommunicating case, the multiple dialogues that a multicommunicator is part of might involve completely separate topics or be more closely related, such that satisfaction of the goals of one depends on actions in the other. For example, a question arising in a meeting might be conveyed and answered over the text channel. We use the term *multi-floor dialogue* to refer to cases in which the high-level dialogue purposes are the same, and some content is shared, but other aspects of the information state, such as the participant structure and turn-taking expectations, are distinct. Situations of distributed decision-making and action are quite common, e.g., in a restaurant where some people take the customer's order and others make the food, or in military units, where orders are relayed through the chain of command. In some cases, where all parties can hear all communication, we can view this as multi-party dialogue within a single floor, but in other cases not all the communications are available to all participants – this is a case of multi-floor dialogue. We are particularly interested in capturing the latter case.

In the next section, we present the annotation scheme. This is applied to a corpus of human-robot dialogue (section 3), and the scheme is shown to have high inter-annotator reliability (section 4). Our objectives for the annotation are two-fold. First, we seek to explore how multi-floor dialogue works and characterize different kinds of multi-party, multi-floor contributions. This is addressed by analysis of dialogue annotated with this scheme and the kinds of patterns of interaction that are observed (see section 5). Second, we use data from the corpus annotated with this scheme to serve as training and evaluation data for creating automated multicommunicators (see section 6).

2. Annotation Scheme

We annotate two aspects of Dialogue Structure at the *meso-level* (bigger than a single speaker-turn, but smaller than a complete dialogue activity) (Traum and Nakatani, 1999). First, we look at *intentional structure* (Grosz and Sidner, 1986), consisting of units of dialogue utterances that all have a role in explicating and addressing an initiating participant's intention. Second, we look at the relations between different utterances within this unit, which reveal

Expansions	relate utterances that are produced by the same participant within the same floor.
Responses	relate utterances by different participants within the same floor.
Translations	relate utterances in different floors.

Table 1: Top Level Corpus Relations

how the information state of participants in the dialogue is updated as the unit is constructed. To fully understand the intentional and interactional structure, it is also necessary to include aspects of micro-level meaning (e.g. dialogue acts) and macro-level meaning (e.g. dialogue purpose); however those aspects are beyond the scope of the current paper.

We call the main unit of intentional structure a *transaction unit*, following Sinclair and Coulthard (1975) and Carletta et al. (1996). A transaction unit (TU) contains an initial message by one speaker and all subsequent messages by the same and other speakers across all floors to complete the intention. For example, a transaction may consist of an instruction initiated by one participant in one floor that is relayed by a multicommunicator to another floor, and then performed by another participant of the second floor, in addition to various sorts of feedback between pairs of participants. For a TU we focus on the lowest level of dialogue in which intentions are fulfilled across speakers. In some kinds of dialogues, particularly complex negotiations or problem-solving, intentional structure can be recursive, such that the purpose of one segment partially contributes to the purpose of a higher-level segment (Grosz and Sidner, 1986). Other types of dialogues have a flatter structure, including transactions that contribute to an overall dialogue purpose, but with few, if any, levels in between. Each utterance-level message is assigned to at most one TU, and the TU is defined by the set of constituent utterances. At most points in a dialogue, there is only one active TU, however there are occasions where there are multiple active TUs, when a new one is started before the previous one has been completed.

We also model the internal structure of TUs as *relations* between pairs of utterances within the unit. Each relation is annotated by coding a relation-type and an antecedent for each utterance after the first in a transaction. Thus, each transaction can be viewed as a tree structure, with the first utterance as root (having no relation-type or antecedent annotations). While relations often exist between an utterance and multiple previous utterances, to simplify the annotation, we code only the most direct, recent such relation. This practice is common for many annotation efforts, e.g. the "code-high" principle from (Condon and Cech, 1992). In the future, we plan to use inference rules to derive some "indirect relations" from what has been annotated.

We have developed a taxonomy of relation types based on how a new utterance is connected to its antecedent. At the highest level, we distinguish relations by the combination of the participants who produced the utterances (often called "speakers", even if the communication was not using speech) and floors that the utterances are part of (Table 1).

Each of these types has one or more subtypes. For expansions, we indicate how the intention is expressed across multiple utterances. Utterances that add additional content are termed *continues*, and utterances that remove or replace some content are called *corrections*. Utterances that do neither, but reiterate some content, are termed *summarizations*. Finally, utterances that consist primarily of explicit discourse markers that link a preceding utterance to a following one are termed *link-next*.

processing:	positive feedback (Allwood et al., 1992) at the perception level, but lack of feedback at higher levels.
acknowledgement:	positive feedback at the understanding level, indicating *grounding* (Clark and Schaefer, 1989; Traum, 1994), with subcategories indicating attitudinal reaction, commitment and performance status of an instructed action.
clarification:	negative feedback of understanding, with subcategories representing different strategies for repair.
question-response:	the antecedent is a question and the response indicates understanding and some attempt to address the question,
reciprocal response:	response indicates the same or similar content as the antecedent e.g. reciprocal greetings.
3rd turn feedback:	a response to a response - often an evaluation of the response.
other:	a response not fitting the other categories.

Table 2: Response Relation Types

For translations, there is one subtype for each source and target floor combination. Thus, a dialogue with two floors would have two translate relations, while one with three floors would have up to six. We also include two other types: *quotation* and *comment*, where some content is conveyed across floors but not the same illocutionary force as the original.

We also annotate several types of responses, many of which in turn have sub-types. The main types of responses are summarized in Table 2. These cover positive and negative feedback on contact, perception, understanding, and attitudinal reaction (Allwood et al., 1992), as well as perfomance status and other relevance relations. The *acknowledgement*, *clarification*, and *question-response* relation types have multiple sub-types, as shown in Table 3. Acknowledgements all indicate a claim or demonstration of understanding of the antecedent. However, the subtypes also indicate the status of an instructed action: whether it has been started or completed, or whether the responder thinks it can or will be done. Clarifications all indicate a lack of ability to fully understand and act on an instruction, with subtypes indicating problems with receiving a message, the message being incomplete, requesting clarification, or providing a repair elicited by another participant. Question responses can be either answers or non-answers that address a question but do not directly provide an answer.

acknowledgment	(listed in order of likelihood of action being successfully performed)
ack-done	ack that a command or prior planned act has been completed successfully
ack-doing	ack that the speaker understands the command and is starting to do it
ack-wilco	ack of a command and promise to do it in the future (includes acceptance with something like "ok")
ack-understand	express or show understanding without commitment to action or agreement. Includes repetitions of what was said, affirmative cue words like "uh-huh".
ack-try	ack of a command and promise to try to do it (but not necessarily)
ack-unsure	ack of understanding of a command, expressing uncertainty about whether it can/will be done. Not clearly an ack-cant or ack-try, but also more than ack-understand because of some explicit statement of doubt about possibility or future action.
ack-cant	expression that the previous command was understood but can't be executed.
clarification	
req-clar	request for clarification – indicates that something in the prior utterance was not clear, and asks the other speaker to do something about it (such as answer a question or confirm a trial). E.g. a command was not specified well enough to be unambiguously confirmed and carried out.
clar-repair	providing a clarification to a prior utterance , after prompting by another (other-initiated self-repair).
missing info	indicates a specific part of the antecedent was not interpretable well enough to act on, but not requesting further action (e.g "I don't know which object you are referring to"). The other party has the option of whether to clarify-repair or move on and do something else.
nack	indicates that the antecedent could not be understood well enough to act on, but not explicitly requesting action (e.g. "no copy" or "I don't understand")
req-repeat	request to repeat a prior utterance
clar-repeat	providing (other-initiated self) repeat, after prompting to repeat with a req-repeat
question-response	
answer	an answer to a question, other than a clar-repair or clar-repeat.
Non-Answer-Response (NAR)	addresses question without providing an answer. E.g. explains why an answer won't be given, or the question is not relevant, or a helpful suggestion of how the requested information might be arrived at.

Table 3: Response Sub-Relations

3. Initial Domain Application: Distributed Human-Robot Interaction

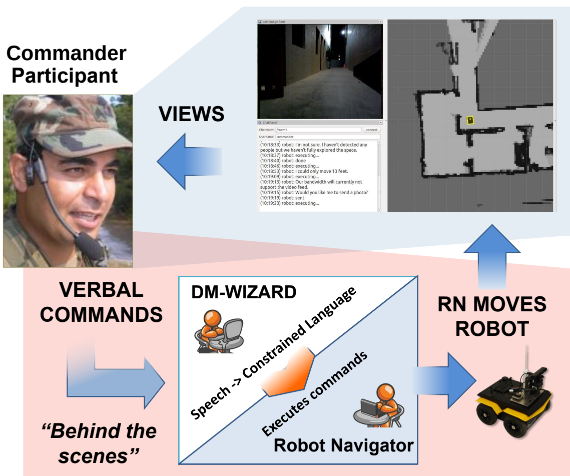

Figure 1: Domain Application Experimental Setup: Human-Robot Interaction with Wizards

We first apply this annotation scheme to a corpus of human-robot interaction, taken from a project with a long-term goal to create an autonomous robot intelligence that can collaborate with remotely located human participants on exploration and navigation tasks. In the initial versions, a human "Commander" tasks the robot verbally, and gets feedback via multiple modalities, including text messages, a live 2D-map built from the robot's LIDAR scanner, and still photos captured from the robot's front-facing camera. In order to collect sufficient information about the type of language used by a Commander (Marge et al., 2017), and provide training data to support development of appropriate language processing components, the development of the autonomous human-robot interaction begins with a series of "Wizard of Oz" experiments (Marge et al., 2016; Bonial et al., 2017), where the robot is controlled by two wizards, with an internal communication floor, distinct from the floor used by the Commander to communicate with the robot. The wizards include a Dialogue Manager (DM-Wizard, or DM) who handles communication to the Commander and "speaks" via text messages (Bonial et al., 2017) and a Robot Navigator (RN-Wizard, or RN) who teleoperates the robot based upon commands issued by the Commander relayed by the DM (Figure 1).

This Wizard of Oz communicative setting thus involves multi-floor dialogue: three participants (Commander, DM-Wizard, and RN-Wizard), two floors (Commander to DM, called "left", and DM to RN, called "right") and four distinct message streams. The DM-Wizard is multicommunicating and can translate from one floor to the other. When the DM translates Commander speech to the "right" floor, this is called translation-r, and when they translate from the RN to the "left" floor, this is termed translate-l. The RN and the Commander cannot speak to one another directly.

4. Corpus and Annotation

A total of 60 dialogues (up to 20 min. each) were collected from 20 Commander participants. Aligned transcripts were produced for all dialogues in the form of Tables 4–7, by transcribing Commander and RN speech, and aligning with

	Left Floor		Right Floor		Annotations		
#	**Commander**	**DM→Commander**	**DM→RN**	**RN**	**TU**	**Ant**	**Rel**
1	move forward three feet				1		
2		ok			1	1	ack-wilco
3			move forward 3 feet		1	1	translation-r
4				done	1	3	ack-done
5		I moved forward 3 feet			1	4	translation-l

Table 4: Example Minimal TU. The ack- prefix indicates a type of acknowledgement.

	Left Floor		Right Floor		Annotations		
#	**Commander**	**DM→Commander**	**DM→RN**	**RN**	**TU**	**Ant**	**Rel**
1	face west				1		
2	and take a photo				1	1	continue
3			face west, photo		1	2*	translation-r
4		executing...			1	2*	ack-doing
5				image sent	1	3	ack-done
6		sent			1	5	translation-l

Table 5: Example Extended-Link TU. The ack- prefix indicates a type of acknowledgement. * indicates that the most direct antecedent is part of a sequence connected by expansions (e.g. utterances 1 and 2 are both being translated or responded to, but 2 is the most recent antecedent).

DM text messages. The right three columns include annotations of dialogue structure, with Transaction Unit (TU), Antecedent (Ant), and Relation-type (Rel) for each.

Because all participants in the dialogues are guiding the robot to accomplish a small set of search and exploration tasks, many of the TUs in this corpus involve instructions initiated by the Commander, which are translated into a simplified form by the DM and passed along to the RN-Wizard, who then carries out those instructions by teleoperating the robot. We call TUs that include only these components and acknowledgements (and translations of acknowledgements) "Minimal TUs" – an example of which is shown in Table 4, or "Extended-link" TUs, like Table 5, depending on whether they include a single instruction or a sequence of multiple instructions. However, there are also units involving questions and clarifications, and other types of dialogue moves, when dialogue ensues to repair any Commander instructions that are unclear (perhaps due to garbled speech), ambiguous (as to a referent in the physical environment), or impossible (given the constraints of the physical environment). Table 6 includes two example TUs involving questions and responses. Table 7 shows another example of two TUs, however in this case one involves a repair, and the second one commences before the first one has been completed.

4.1. Inter-annotator Reliability for this annotation scheme and corpus

Inter-annotator reliability was calculated separately on three different markables: antecedents, relation types, and transaction units. An initial sample of 3 dialogues (482 utterances) was annotated by up to 5 coders using an early version of the coding manual; this was followed by several iterations of refining the coding manual, after which a second sample (1 dialogue, 314 utterances) was annotated by 6 coders. Results of the agreement tests are in Table 8.

After each round of agreement testing, disagreements were reconciled, and the annotation guidelines were revised to clarify issues of disagreement. In the initial version, there were confusions about how many distinct instructions were part of the same transaction unit. The guidelines were revised to make clear that continuations of an instruction would be considered part of the same TU only until the DM started acting on them (with feedback and/or a translation). Subsequent Commander instructions would be seen as starting a new TU, even if they follow semantically from the previous instruction. For example, in Table 7, the instruction in line 6 to take another picture is seen as new TU, even though the previous TU has not been completed. Another added guideline allowed the antecedent to be designated as a sequence (via a *) rather than just a single utterance.

A major contributor to remaining disagreement in the second test was the case where compound instructions are both presented and translated in multiple utterances. In this case, the second translation was both a translation of the second instruction as well as a continuation of the first translation. The initial guidelines indicated that the most recent relation was to be annotated, but in this case it was felt that translations were more important to capture than continuations, and marking translations of individual utterances (where appropriate) was more informative than translations of whole sequences, so the guidelines were updated to consider the most important recent relation. An example is shown in Table 9. Here, the whole sequences of 6,7,10

	Left Floor		Right Floor		Annotations		
#	**Commander**	**DM→Commander**	**DM→RN**	**RN**	**TU**	**Ant**	**Rel**
1	how many window openings do you see in front of you				1		
2		three			1	1	answer
3	do you see a yellow flashlight				2		
4		processing...			2	3	processing
5		I'm not sure			2	3	answer
6		If you describe an object, you can help me to learn what it is.			2	3	non-answer response

Table 6: Example including two QA TUs

	Left Floor		Right Floor		Annotations		
#	**Commander**	**DM→Commander**	**DM→RN**	**RN**	**TU**	**Ant**	**Rel**
1	move to where you see the first cone				1		
2		I'm not sure which object you are referring to. Can you describe it in another way, using color or its location?			1	1	request-clarification
3	move to the cone on the right a red cone on the right				1	2	clarification-repair
4			move to face the cone on the right		1	3	translation-r
5		executing...			1	3	ack-doing
6	take another picture				2		
7				done	1	4	ack-done
8		done			1	7	translation-l
9			image		2	6	translation-r
10				image sent	2	9	ack-done
11		sent			2	10	translation-l

Table 7: Multifloor example: two partially interleaved TUs (Repair and minimal)

Markable Type	Agreement		Distance Metric
	Initial	Second	
Antecedents	0.72–0.82	0.78	Nominal[a]
Relation Types	0.77–0.82	0.89	Nominal[a]
Transaction Units	0.48–0.70	0.93	MASI[b]

[a]Krippendorff (1980) [b]Passonneau (2006)

Table 8: Inter-annotator agreement (Krippendorff's α)

and 5,8, 9 are acknowledgements and translate-r of the sequence of 1,3, respectively. So 10 and 9 are both continuations within their local sequence as well as direct translation and acknowledgment of 3.

We expect that agreement would improve following the final guidelines, but as it was already fairly high, we did not do a final test.

5. Analysis of Corpus Annotation

Table 10 shows the distributions of relation types and major subtypes in the annotated corpus. The high percentage of translations and relatively low percentage of expansions indicate a high degree of multi-communicating and relatively low complexity in intra-turn discourse structure, respectively.

We also examined the tree-structures of TUs, which reveals that 644 unique TU patterns make up all the 2230 observed TUs in the collected corpora. These patterns were classified into the following taxonomy of TU types: Minimal, Extended-Link, Repair, Question-Answer, and Other.

Minimal TUs consist of a single instruction from the Commander that is well formed and that the DM passes on in a single instruction to the RN. Minimal transactions include a single translation-r, an acknowledgment back to the Commander, a successive acknowledgment from the RN, and finally a translation-l back to the Commander. Minimal TUs

	Left Floor		Right Floor		Annotations		
#	**Commander**	**DM→Commander**	**DM→RN**	**RN**	**TU**	**Ant**	**Rel**
1	go one hundred eighty degrees				1		
2		done,sent			0	0	translation-l
3	and take a picture				1	1	continue
4		ok			1	3*	ack-understand
5			turn 180		1	1	translation-r
6		I will turn around 180 degrees			1	1	ack-wilco
7		and. . .			1	6	link-next
8			then. . .		1	5	link-next
9			send image		1	3	translate-r
10		I will send a picture			1	3	ack-wilco
11		turning. . .			1	3*	ack-doing
12				uh done and sent	1	9*	ack-done
13		done, sent			1	12	translation-l

Table 9: Example TU containing complex instructions conveyed across floors in parts. Note that line 2 is from the previous TU, but its antecedents are not shown here so it is labelled 0.

Type	**Subtypes**	**#**	**%**
Translation		4287	39
	translation-r	2355	21
	translation-l	1911	17
	comment	21	< 1
	quotation	0	0
Expansion		1583	14
	continue	1175	11
	link-next	337	3
	correction	50	< 1
	summarization	20	< 1
Response		5193	47
	acknowledgment	3998	36
	done	2015	18
	doing	1357	12
	wilco	592	5
	understand	34	< 1
	try	15	< 1
	unsure	14	< 1
	can't	11	< 1
	clarification	569	5
	req-clar	266	2
	clar-repair	237	2
	missing info	36	< 1
	nack	20	< 1
	repeat	8	< 1
	processing	315	3
	question-response	212	2
	answer	84	1
	non-answer	11	< 1
	other	48	< 1
	3rd turn feedback	37	< 1
	reciprocal response	14	< 1

Table 10: Corpus Relation frequency

make up 48% of all TUs. An example of a minimal TU is shown in Table 4. Another example is the second TU in Table 7.

Extended-Link TUs consist of more than one well-formed instruction from the Commander that the DM passes on in one or more instructions to the RN. Common examples of this include cases where a Commander asks the robot to move to a particular landmark and take a picture of it. An example is shown in Table 5. Extended-Link patterns make up 26% of all TUs.

Repair TUs contain an instruction that requires a clarification; the instruction is not actionable (e.g., it is not well-formed, or missing information) and an exchange must occur to rectify it. 9% of TUs are a repair and successfully resolve the conflict (TU 1 in Table 7 is an example of a successful repair TU). 2% of TUs involving a repair were not resolved, and might have been abandoned or a new and unrelated instruction was issued.

Question-Answer TUs contain a question and an answer or other response that sometimes involves conference between each floor, but does not involve repairing an instruction. Question-Answer TUs are different from the Repair TU; Repair TUs and Question-Answer TUs may contain a question, answer, reciprocal response or 3rd turn feedback, but a Question-Answer TU does not involve malformed instructions, thus requiring no clarification. Table 6 shows two examples of a QA TUs. In 1% of TUs, after a question is answered, the answer, as an instruction, is passed to the RN to complete the action, for example, the DM asks if the Commander would like a picture, and they respond "yes". In 4% of TUs, a question is answered but not passed as an executable instruction to the RN, as in the instance of a Commander asking about the robot's capabilities.

Other TUs that do not fall into these categories include instructions that were abandoned or interrupted and do not contain any questions or repairs (11% of TUs).

Classification and examination of TUs show consistency throughout the task for relaying well-formed instructions, with 74% of TUs being Minimal or Extended-Link, indicative of instructions that can be translated to the RN and then executed. It is encouraging to observe that the majority of malformed instructions in Repair TUs are successfully resolved and a command is executed by the RN (82% of TUs classified as Repair). On the other hand, the majority of questions in QA TUs do not involve an executable instruction (82%), suggesting that after new information is received from another speaker, the Commander comes to a decision about what further instructions should be issued or should be abandoned and re-issued as new instructions.

6. Applications of Annotated Data

The dialogue structure annotations have been used for two purposes so far: characterizing types of instructions, and automating the dialogue manager. In (Marge et al., 2017), TUs, were used to help characterize the content of instructions people would formulate to a robot before it had a chance to respond. The initial TU instruction and all expansions before any feedback were called *instruction units*, and used to contrast use of absolute vs. relative coordinates. This formulation was used to gain insight into which aspects of these initial instructions changed during the course of participants interaction with the robot.

The approach toward automating the DM involves using the annotated corpus data to train a statistical text classifier in the NPC Editor platform (Leuski and Traum, 2011). To process the data for input to the classifier, we extracted all utterances from the annotated corpus that were produced by the DM. These included the Translation relations (translation-l, translation-r) and the various Response relations (Acknowledgements, Clarifications, and Question-Responses). This gave us a large training set which contained the input-response pairings that were processed by the DM in our experiments. An example training pair for the translation-r relation is “move to the cone on the right a red cone on the right” → “move to face the cone on the right” (see Table 7). In this way, the commander’s actionable instruction was translated to the RN to carry out. In cases where the instruction was not actionable (e.g., “Move forward”) the mapped response was often a clarification request directed to the commander (e.g., “How far would you like me to move forward?”). After training on these pairings, the classifier learned to translate commands and provide appropriate feedback to many of the input utterances.

While a more thorough evaluation of the system is work in progress, reasonably high accuracy on the most common commands was observed: we were able to achieve over 80% accuracy on first dialogue manager response on a held-out corpus of 6 dialogues from our corpus. The TU patterns were also used to develop dialogue manager policies to engage in these patterns, such as the type and quantity of feedback given and how to sequence feedback to the commander with translations to the robot navigator.

7. Conclusion

We have presented a new annotation scheme for meso-level dialogue structure in multi-floor dialogue. The scheme covers Transaction units that accomplish collaborative goals, sometimes across multiple floors, and relations between individual utterances in the transaction unit,. The scheme has been used to annotate an initial corpus with two floors, in a human-robot interaction scenario. We have presented statistics of the different types of relations and transaction structures present in the corpus, as well as introduced some of the ways that the annotated corpus is being used.

8. Acknowledgements

The effort described here has been supported by the U.S. Army. Any opinions, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

9. Bibliographical References

Allwood, J., Nivre, J., and Ahlsen, E. (1992). On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9.

Bonial, C., Marge, M., Artstein, R., Foots, A., Gervits, F., Hayes, C. J., Henry, C., Hill, S. G., Leuski, A., Lukin, S. M., Moolchandani, P., Pollard, K. A., Traum, D., and Voss, C. R. (2017). Laying Down the Yellow Brick Road: Development of a Wizard-of-Oz Interface for Collecting Human-Robot Dialogue. In *AAAI Fall Symposium*.

Bunt, H., Alexandersson, J., Choe, J.-W., Fang, A. C., Hasida, K., Petukhova, V., Belis-Popescu, A., and Traum, D. (2012). ISO 24617-2: A semantically-based standard for dialogue annotation. In *International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, May.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., and Anderson, A. (1996). HCRC dialogue structure coding manual. Technical Report 82, HCRC.

Clark, H. H. and Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science*, 13:259–294.

Condon, S. and Cech, C. (1992). Manual for coding decision-making interactions. unpublished manuscript, updated May 1995, available at: ftp://sls-ftp.lcs.mit.edu/pub/multiparty/coding_schemes/condon.

Edelsky, C. (1981). Who’s got the floor? *Language in Society*, 10(3):383–421.

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

Krippendorff, K., (1980). *Content Analysis: An Introduction to Its Methodology*, chapter 12, pages 129–154. Sage, Beverly Hills, CA.

Leuski, A. and Traum, D. (2011). NPCEditor: Creating virtual human dialogue using information retrieval techniques. *AI Magazine*, 32(2):42–56.

Marge, M., Bonial, C., Pollard, K. A., Artstein, R., Byrne, B., Hill, S. G., Voss, C., and Traum, D. (2016). Assessing agreement in human-robot dialogue strategies: A tale of two wizards. In *International Conference on Intelligent Virtual Agents*, pages 484–488. Springer.

Marge, M., Bonial, C., Foots, A., Hayes, C., Henry, C., Pollard, K. A., Artstein, R., Voss, C. R., and Traum, D. (2017). Exploring Variation of Natural Human Commands to a Robot in a Collaborative Navigation Task. In *RoboNLP*.

Passonneau, R. (2006). Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proc. of LREC*.

Prasad, R. and Bunt, H. (2015). Semantic relations in discourse: The current state of iso 24617-8. In *Proceedings 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, pages 80–92.

Reinsch, N. L., Turner, J. W., and Tinsley, C. H. (2008). Multicommunicating: A practice whose time has come? *Academy of Management Review*, 33(2):391–403.

Sinclair, J. M. and Coulthard, R. M. (1975). *Towards an analysis of Discourse: The English used by teachers and pupils.* Oxford University Press.

Traum, D. and Larsson, S. (2003). The information state approach to dialogue management. In Jan van Kuppevelt et al., editors, *Current and New Directions in Discourse and Dialogue*, pages 325–353. Kluwer.

Traum, D. R. and Nakatani, C. H. (1999). A two-level approach to coding dialogue for discourse structure: Activities of the 1998 working group on higher-level structures. In *Proceedings of ACL 1999 Workshop: Towards Standards and Tools for Discourse Tagging*, pages 101–108.

Traum, D. R. (1994). *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. thesis, Department of Computer Science, University of Rochester. Also available as TR 545, Department of Computer Science, University of Rochester.