SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 182–183.
Los Angeles, California, 21–23 September 2011.

Gestures Supporting Dialogue Structure and Interaction in the Bielefeld *Speech and Gesture Alignment Corpus* (SaGA)

Florian Hahn
CRC 673 Alignment in Communication
Bielefeld University
fhahn2@uni-bielefeld.de

Hannes Rieser
CRC 673 Alignment in Communication
Bielefeld University
hannes.rieser@uni-bielefeld.de

Abstract

We report about gestures supporting dialogue structure and interaction in the Bielefeld Speech and Gesture Alignment corpus and provide a first classification of them based on Hahn and Rieser (2009-11). Numbers will be given on the poster.

Types of Dialogue Supporting & Interactive Gestures

Our study is based on the Bielefeld Speech and Gesture Alignment corpus (SaGA) containing 25 route description dialogues generated as follows: a Router "drives" through a VR town along a route. His ride is reported to a Follower, who is expected to follow the route by himself. The data contains 5000 indexical and iconic gestures annotated in ELAN and rated (see (Lücking et al., 2010), for results) and approximately 1000 gestures supporting dialogue structure and interaction.

An important trait of the Router-Follower-situation is that it is "layered" (Clark, 1996): We have the route context using the conversational participants' (CPs) gesture spaces detailing the topical or baseline information, the larger embedding context of the experimental situation and the still more encompassing one consisting of the University and Bielefeld City. The discourse-related gestures introduced below can be grouped roughly into gestures used in turn allocation, feed-back gestures in second turn, those indicating assessment of evidence, gestures serving to highlight information, sequences of quick feed-back or monitoring gestures tied to sub-propositional contributions and, finally, truly interactive gestures exclusively social in character. All of these are accompanying speech.

Gestures related to turn allocation: Since the seminal paper of Sacks et al. (1974), valid also for dyads, we assume a regularity for turn allocation in dialogue depending structurally on the larger speech-exchange system: current speaker dominates, he selects next. If not, one of the other speakers can self-select. This option omitted, the first speaker may continue. The SaGA data show that there is more freedom in this schema leaving room for quick interrupts of other. These become acceptable for CPs if interactionally cushioned. Gestures in this class exploit the layeredness property of the situation: current speaker points to other selecting him as next. In contrast, indexing other to take the turn oneself is also a possibility. In the context of a completion current speaker may gesturally invite a contribution from other. Time being a scarce resource, current speaker may indicate a lapse should be tolerated by other and use a finger-to-lip or finger-below-lip gesture to express that. In tightly coordinated discourse there is an interesting "attack-ward-off pair": with a gesture similar to "indexing other" other may indicate that he wants to contribute at a non-turn-transition relevance place. Discouraging that, current speaker may try to fence him off with a posture using palm slanted and ASL-shape B-spread. Under pressure current speaker may give in and offer a "go ahead", palms up, directed against the domineering CP (see (Rieser, 2011)).

Feed-back gestures in second turn. Speaker of a second turn may use an iconic gesture of previ-

ous speaker in order to indicate acknowledgement or accept. As with the indexing, next speaker's gesture imitation uses a topical gesture in a discourse function. Less spectacular means can also be used in second turn, for example pointing without referring. Acknowledging an acknowledgement of second speaker by first may be done in essentially the same way (Bergmann et al., 2011).

Gestures indicating assessment of evidence. Given the Follower's route following task, it is of course vital that he get reliable information about landmarks and directions. This is a pressure on both CPs. We observed two groups of gestures to indicate reliability of information. One is conveying doubt concerning the fit of a description, usually for a landmark or one of its properties. The other one is indicating an agent's epistemic state concerning a situation and characterizing it as weaker than knowing or believing. The first one is aligned with the description in question using ASLs B-spread and a wiggle in handshape or wrist, the other one related to propositional content is a lifting of a hand out of and into a rest position with handshape B-spread accompanied by a head shake.

Gestures to highlight and to downgrade information. We take it for granted that beats are used for emphasis. However, underlining information – often an accented tone group – can also be suggested lifting a G-shaped hand, directing it against the addressee and moving it in a beat-like fashion. On the other hand we have the near-universal "brush-away" gesture indicating that information is considered to be not so relevant.

Sequences of quick feed-back or monitoring gestures tied to sub-propositional contributions. Propositions are of a Fregean design. CPs in near-to-natural task-oriented dialogue often converse quickly and in short thrusts. So we can have a Router's "don't interrupt" followed by the Follower's "let me interrupt" and, finally, the Router's acknowledgement and a "go ahead" gesture. This shows that full-blown dialogue acts do not always matter.

Truly interactive gestures exclusively social in character. To sum up: gestures accompanying turn allocation, feedback gestures in second turn and sequences of quick feed-back or monitoring gestures have to be embedded into suitable adjacency pairs and reconstructed at the level of dialogue acts. So, in order to explain a gesture's function many parameters have to be considered besides gesture morphology. The embedding speech exchange system in SaGA is a plan-based (memory-based) one on the Router's side and a plan-generating one on the Follower's side providing larger sequential structures. Gestures indicating assessment of evidence and those to highlight and to downgrade information figure at the level of dialogue acts. From these structural features we want to delineate gestures which are truly interactive such as hand and body postures to mollify someone or touching or caressing him. We have calming down and don't bother gestures. Calming down has a B-spread handshape and a slanted palm directed against the torso of other. Don't bother gestures resemble the brush-away gestures in many respects but are also directed against the other's torso.

Acknowledgments

This research has been supported by the DFG in the CRC 673 "Alignment in Communication", Bielefeld University, project B1 "Speech-gesture Alignment". Thanks to two SemDial reviewers.

References

Kirsten Bergmann, Hannes Rieser, and Stefan Kopp. 2011. *Regulating Dialogue with Gestures - Towards an Empirically Grounded Simulation with Conversational Agents*, SIGdial 2011.

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

Florian Hahn and Hannes Rieser. 2009-11. *Dialogue Structure Gestures and Interactive Gestures. Annotation Manual. CRC 673, Alignment in Communication. Working Paper*, Bielefeld University.

Andy Lücking et al. 2010. The Bielefeld Speech and Gesture Alignment Corpus (SaGA). In: Michael Kipp et al. (Eds.) 2010, *Workshop: Multimodal Corpora*, 92-98.

Hannes Rieser. 2011. *Gestures Indicating Dialogue Structure*. Accepted for SEMdial

Harvey Sacks, Emanuel A. Schegloff, Gail Jefferson 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. In: *Language, 50*: 696-735.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 184–185.
Los Angeles, California, 21–23 September 2011.

Cognitive Models of Failure and Recovery in Natural Language Interactions - A Joint Actions Approach

Arthi Murugesan
NRC/NRL Postdoctoral fellow, 4555 Overlook Ave, Washington D.C.
Arthi.Murugesan.ctr@nrl.navy.mil

Derek Brock
Naval Research Lab, 4555 Overlook Ave, Washington D.C., 20375, USA.
Derek.Brock@nrl.navy.mil

Wende K. Frost
Naval Research Lab, 4555 Overlook Ave, Washington D.C., 20375, USA.
Wende.Frost@nrl.navy.mil

Dennis Perzanowski
Naval Research Lab, 4555 Overlook Ave, Washington D.C., 20375, USA.
Dennis.Perzanowski@nrl.navy.mil

Abstract

Natural language interaction, like any other joint action, is a coordination problem involving agents who work together to convey and thus coordinate their interaction goals. Joint actions frequently fail, as agents act on their best guesses of what is intended by the other person. The ability of agents to correct each other, and recover from failures, makes it possible for joint actions to succeed even in highly error prone situations. In the modeling work presented here, a sequence of interrelated modules, originally developed in the Polyscheme cognitive architecture to understand simple commands to video application, is modified to implement error discovery and accommodate possible user-initiated repairs.

1 Introduction

Natural language can be viewed as a collaborative means for expressing and understanding intentions using a body of widely shared conventions. The challenge of conveying an intention from one agent to another, for example, from a speaker to an addressee, can be characterized as a coordination problem that participants must work together to solve. People rely on a procedural convention for collaborating with each other (Clark 1996) that can be summarized as follows: 1) make the focus of the coordination problem explicit or salient; 2) pose a problem one expects the addressee will be able to solve; and 3) frame the problem in a manner that makes it easy for the addressee to solve it.

Previous modeling work by Murugesan et al. (2011) demonstrates how a sequence of interrelated cognitive models can simulate the stages of reasoning involved in understanding simple commands issued to a video monitoring system. This paper builds on the previous work and describes how agents can initiate repairs to recover from failures in each of these stages of reasoning.

2 Natural Language Interactions as Joint Actions

All agents that perform joint actions must rely on certain heuristic presumptions regarding the set of actions they expect to carry out together. In the case of conversation, this includes posing and understanding the problem, working out the intention and acting upon the expected intention. The heuristic presumptions of *salience* and *solvability* are modeled in the Polyscheme cognitive architecture developed by Cassimatis (2006). New modeling work related to initiating repairs is discussed in the following two sections.

3 Repairs in Salience

Clark's principle of joint salience suggests, roughly, that the ideal solution to a coordination problem is one that is most prominent between the agents with respect to their common ground. Thus, for example, when the model's user enters "…the red car…," it is expected that these words are intended to make objects that correspond to this phrase more prominent than other objects in the knowledge and experiences the user shares with the interactive system that is being addressed, which in our case is an interactive video monitoring application.

However, when the same user enters a word the application does not know, for e.g. "… the ted car …" due to a typo 't' instead of 'r', the model recognizes that it is unable to identify the user's intention because the word "ted" is not in the common ground shared by the user and the application(see figure 1). The model responds by showing the user a message saying "I do not recognize the word ted."

```
<constraint>
IsA(?word, WordUtteranceEvent, E, ?w) ^
Orthography(?word, ?orth, E, ?w) ^
-IsA(?orth, LexicalEntry, E, ?w)
==>
EncounteredUnknownWord(?word, E, ?w) ^
-InSharedLexiconWithUser(?orth, E, ?w)
</constraint>
```

Figure 1. A sample constraint from the model that identifies an unknown word.

The user now has the option of recovering by either rephrasing the utterance with words known to the system, or in the case of advanced users, adding the specific unknown word and its syntactic, semantic and common sense implications to the common ground.

4 Repairs in Solvability

The first stage in solving the coordination problem posed by a natural language utterance involves parsing it, forming its semantic interpretation and combining the semantic knowledge with relevant world knowledge in the common ground. In the second stage, the listener reasons further to identify the intention or goal behind the speaker's actions, the actions in this case being the speaker's words.

4.1 Repairs in Stage 1 – Natural Language Understanding

Sentence processing may terminate abruptly due to any of several causes for failure, the most common being an inability to form a valid parse of the sentence. On failure, the model reports the problem in parsing to the user, and initiates a repair by asking the user to enter a simpler or more grammatically correct sentence.

The process of understanding the semantics or intended meaning of a sentence within the context of domain knowledge may also result in inconsistencies. For example, a contradiction arises when "…the stalled car passed the truck…" (i.e., a car previously referentially identified in this way) is combined with simple common sense knowledge that stalled objects do not move. The model again initiates a repair by identifying the contradiction, reporting that a stalled car cannot be motion. The user can then alter the input (e.g., "the silver car passed the truck") or, more elaborately, make changes to the domain rules associated with this input (e.g., sometimes stalled cars are towed and can thus be in motion).

4.2 Repairs in Stage 2 – Task Recognition

When one agent's intentions must be understood and acted upon by another, addressees ordinarily presume the speaker has a practical outcome or task in mind that they can recognize and help achieve. For example, when a user says, "Show me the red car passing the black car," the monitoring application's model recognizes that the user expects it to find and display a corresponding scene. But coordinating tasks specified in this way can fail in at least two ways: 1) the intended task may not be correctly recognized — when the user says "Show me the next stop of the bus", the literal meaning of the bus at a signal light is not intended (converstional implicatures) or 2) the application may not be able to perform the identified task—for example, the application currently set up to display only one scene is incapable of responding to "Show me everywhere the red turns left." The model is able to identify when it is incapable of performing the task and allows the user to revise or repair the command.

Conclusion

This paper presents various stages in which a natural language interaction can fail and introduces the notion that cognitive models can be created to accommodate error recovery initiated by an agent participating in the conversation.

References

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

Nicholas L. Cassimatis. 2006. A Cognitive Substrate for Achieving Human-Level Intelligence. AI Magazin 27(2): 45-56.

Arthi Murugesan, Derek P. Brock, Wende K. Frost and Dennis Perzanowski. 2011. Accessing Previously Shared Interaction States through Natural Language. Springer-Verlang, HCII 2011, CCIS 173: 590-594.