

Developing a Contextualized Multimodal Corpus for Human-Robot Interaction

Anders Green, Helge Hüttenrauch, Elin Anna Topp, Kerstin Severinson Eklundh

Royal Institute of Technology
School of Computer Science and Communication
100 44 Stockholm
Sweden
{green, hehu, topp, kse}@csc.kth.se

Abstract

This paper describes the development process of a contextualized corpus for research on Human-Robot Communication. The data have been collected in two Wizard-of-Oz user studies performed with 22 and 5 users respectively in a scenario that is called the Home Tour. In this scenario the users show the environment (a single room, or a whole floor) to a robot using a combination of speech and gestures. The corpus has been transcribed and annotated with respect to gestures and conversational acts, thus forming a *core annotation*. We have also annotated or linked other types of data, e.g., laser range finder readings, positioning analysis, questionnaire data and task descriptions, that form the annotated *context* of the scenario. By providing a rich set of different annotated data, the corpus is thus an important resource both for research on natural language speech interfaces for robots and for research on human-robot communication in general.

1. Introduction

The purpose of this paper is to describe a corpus which is used in the research on cognitive robots in the European project Cogniron¹. We will also describe the development process and challenges involved when collecting and annotating the corpus, and the way we are able to contextualize the different types of data. One important aim with the corpus is to support the development of natural language user interfaces for a robot with cognitive capabilities.

We are striving to collect data from many different sources in order to be able to provide a rich context for the modalities that are used for interaction in order to provide means of analyzing the data from different perspectives. Thus we have annotated communicative actions: speech and gesture and other actions related to the task and spatiality: data on positioning of users, objects and locations. This should be seen in contrast to the corpus developed by Maas and Wrede (2006) which focuses on capturing higher dialogue structures (i.e., topics) that emerge during human-robot interaction. Both efforts are in the long run aimed to enable users to train the robot to perform a wide range of tasks that are not preprogrammed – using a multimodal style of interaction. So far we used our corpus in the design process to evaluate the system from a usability perspective (Green et al., 2004), to analyze miscommunication (Green et al., 2006) and to analyze users’ positioning (Hüttenrauch et al., 2006) and task strategies.

1.1. Related research

There are initiatives to collect corpora for multimodal interfaces (Knudsen et al., 2001; Schiel et al., 2002) but few that are targeted for robotics (Bugmann et al., 2001; Bugmann et al., 2004; Wolf and Bugmann, 2005). Koide et al. (2004) have collected and analyzed interaction statistics to investigate human reactions to specific robot behaviors (Koide et al., 2004). Other uses of corpus data include

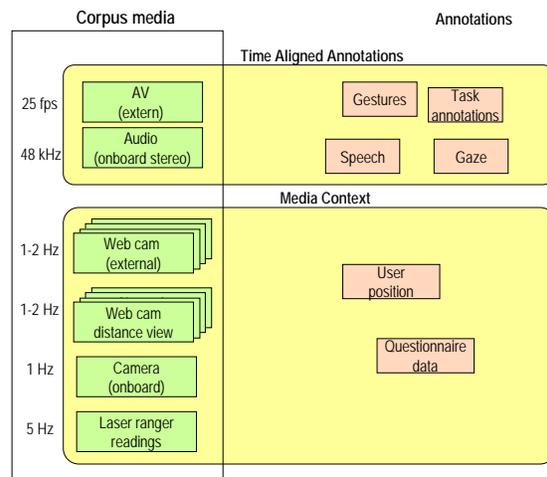


Figure 1: An overview of the corpus

observations of user behavior, e.g., gaze behavior, to evaluate human engagement in interaction (Sidner et al., 2004).

2. General scenario for data collection

The current version of the corpus contains data collected from two user studies (Green et al., 2004; Topp et al., 2006) that have been set up in order to explore user behavior in a scenario that can be characterized as a “Home Tour”. In the scenario the user and robot move around in a home-like environment and the user names objects and locations using a combination of speech and gestures.

This scenario can be characterized as kind of Co-operative Service Discovery and Configuration (Green et al., 2004), stressing the way the user and robot are intended to engage in a joint effort to inform each other of relevant knowledge about the environment. This means that the user is able to *discover* what the robot can do, and is able to *configure* it by actively providing information about the environment. In the studies described in this paper the user can specify

¹www.cogniron.org

names of objects and locations. But in a longer perspective, the user should also be able to interactively specify the tasks the robot can perform related to these locations and objects. In the home tour scenario the user is to guide the robot in an environment containing different objects that could potentially be recognized by the system. Thus the main task for the user is to *introduce herself* to the robot and to *show it objects and locations*. To move the robot around a follow-behavior is used to position the robot during the task. To collect data and explore the character of such an interaction we designed and set up a tele-operated robot system that could be used to perform a Wizard-of-Oz simulation, a technology that has been described in more depth by, e.g., Dahlbäck et al. (1993).

The robot we used for the trials was a ActivMedia PeopleBot (Figure 2). The pan-tilt camera mounted on the robot was moved by the wizard during the sessions so that it appeared as looking at the things that were specified by the user.

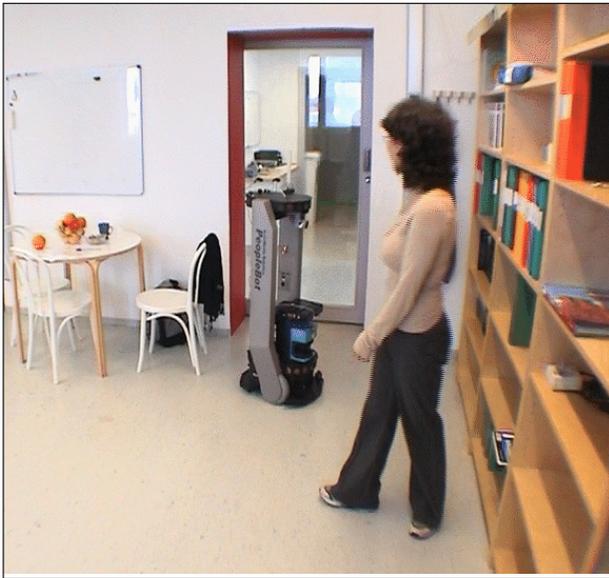


Figure 2: The modified ActivMedia PeopleBot used for the corpus collection and a user engaged in interaction.

2.1. User study 1: Single room – constrained dialogue model

For the first user study we used a room in our robot lab (see Figure 2 and 3) which had been equipped with a set of furniture: a sofa, a dinner table with some chairs, a tv set, book shelf, objects on the table (fruit bowl, mobile phone) etc.

We recruited 22 test persons among students on the KTH campus. This means that there is a bias towards well-educated young people in the study (9 female, 13 male, ~24 years old), but since the aim of the study is primarily explorative we have accepted this circumstance.

2.1.1. Instructions to users

When a user arrived, the test leader informed the subject of the purpose of the study, without revealing that the wizards were controlling the system. Instead the wizards were described as “technicians” with the purpose of controlling the

technical setup and making “online annotations”. During the trial there were three researchers present; one acting as test leader/navigator; one acting as communicator; and one acting as observer. During the setup the observer was positioned in one of the sofas taking notes.

After the introduction the subject signed an agreement giving consent to storing of personal information. The instruction to the user was first of all provided as demonstration, where the test leader addressed the robot, made it follow (i.e. by saying “follow me”); showing it an object by pointing and saying “this the green book”). Then the test leader commanded the robot to get back to the starting position by saying: “go to the recharge station”. The user was also given a written instruction describing the task and principal services the robot could perform:

Task: The user was instructed to use the available dialogue capabilities to teach it objects and locations that were depicted on the back of sheet of paper they were holding.

Following: The follow behavior was described to the user by providing an explicit example of what to say, i.e., “Say ‘Follow me’ to make the robot follow you”.

Showing objects and locations: The Show task was described in an indirect way, i.e., not providing any explicit phrases to the user, with the aim to avoid priming of lexical choice:

You may use your hands to show a single object to the robot. Objects that the robot should know can be indicated if they lie on a flat surface like a coffee table. The surface need to be free from other objects – the robot will use its vision system to collect information about the objects. Say the name of the object that the robot should learn to the robot and use your hand to point where it is.

These descriptions should work as a both an aid for the wizard and a constraining factor for the scenario. The underlying assumption for introducing the user to a simulation of a natural language user interface is to provide the freedom to interact in a way that seems natural to the user – without actually implementing the system for real. However, it is important to provide a set of constraints that bring some realism into the situation of use. This is what Maulsby et al. (1993) refer to being “true to the algorithm”.

2.2. User study 2: multiple rooms – less constrained dialogue

The home tour scenario described earlier is also relevant to the concept of Human Augmented Mapping (HAM) introduced by Topp and Christensen (2005) the aim of which is to provide a link between human-robot interaction and robotic mapping in a way that is compatible with human cognitive representations.

To explore this scenario we designed another user study where the environment where the user and robot interacted was larger. In this case we extended the experiment area to a whole floor of the robot laboratory. This was done in order to provide a scenario that is sufficiently complex both from a technical point of view, i.e., where data collection is

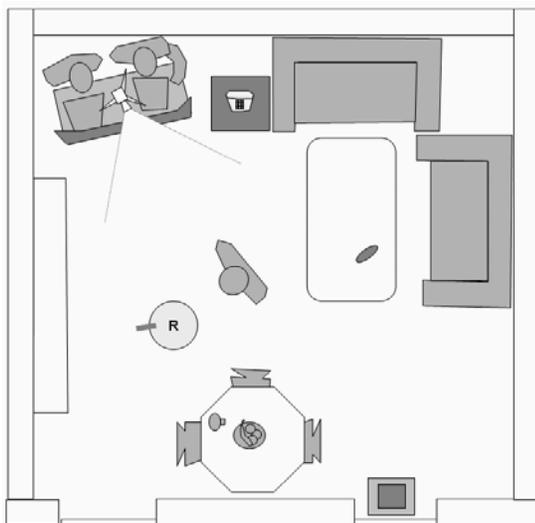


Figure 3: A map of the room used in the data collection. In one corner the position of the wizards is shown. The different objects, like the fruit bowl and the remote control, were always placed on the same initial positions before the study started.

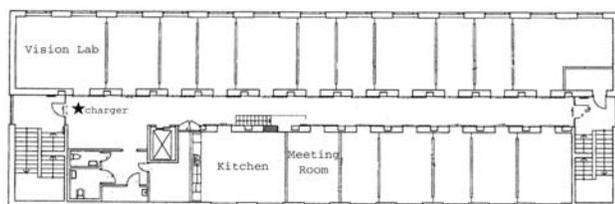


Figure 4: The floor plan of the office environment where the experiment took place. The most prominent places was the kitchen and the robot lab.

used to evaluate algorithms, and provide data on interaction between robot and user.

For this study, which is still ongoing, we have only recruited five users, where knowledge of the environment was a requirement. The users were instructed to use the follow behavior and to present the environment with respect to what locations that the users perceived as important for the robot to know.

The instructions did not include explicit directions on how to name locations to the robot, this was left for the user to find out. The interaction provided by the robot was limited to acknowledging: a) that a location had been received and b) that a follow task has been initiated (“robot is following”). The interaction was recorded using a hand-held video camera and a video camera placed onboard the moving robot, providing a “robot perspective”. After the experiment the subject was interviewed by the experiment leader.

3. Corpus annotation

The recorded video material from both studies is being digitized, transcribed and annotated along several dimensions to support usability research and development of

cognitive modules. The annotations fall into two broad categories: annotation of *communicative acts* and *supportive* or *context-providing annotations*. This is dependent on how the specific type of annotation can be used to perform analyses without using other data. We should note that what is regarded as context is of course dependent on what the analysis is focusing on. However, we have noted that annotations that fall in the *core* category, i.e., time-aligned speech, communicative acts and gestures are invaluable for navigating in the material, e.g., when looking at distances (Hüttenrauch et al., 2006).

3.1. Annotation of Communicative acts

The audio and video recordings are annotated up to what we could characterize as a baseline level: speech utterances and gestures have been transcribed and synchronized in order to provide a format that can be used to navigate the recordings. The synchronized transcriptions have been converted into Anvil XML files (Kipp, 2004) allowing the sessions to be displayed in several layers.

We are using a coding taxonomy to capture communicative acts that can be viewed as multimodal extension of the DAMSL coding schema (Allen and Core, 1997). Our extension of the schema currently involves deictic gestures, emblems, and iconic gestures. We are using a multi-layered style of annotation that allows for more detailed analysis. Our approach is similar to Villaseñor et al. (2000), who proposes the extension of DAMSL with the notion of contribution as participatory communicative acts, according to (Clark and Schaefer, 1989).

Report-task: The categories Report-Task and Report-Task-fail were primarily asserted to utterances where the robot provides a report concerning the task, much like a comment, e.g., “Robot is following” or “Cannot do that”. Allen (Allen and Core, 1997) classifies utterances related to the task on the Information-level, using the categories Task and Task-Management. We have chosen not to annotate the Information-level in our corpus, since the style of interaction used by the user contains very little communication management. Instead we have annotated utterances as Action-Directive or Report-Task (fail) when they are task-related. Allen and Core (1997) also annotated Communication management, but since we are also interested in utterances related to miscommunication (Green et al., 2006) we have annotated repairs using the categories Repair-Action and Request-repair. Utterances that are aimed at self-repair or manage the speakers contributions have been annotated as Own Communication Management (Allwood et al., 1991).

Perception, Attention and Contact: We have classified contributions related to the management of attention and willingness to interact with the categories Request-and Provide-Attention, and Request- and Provide Contact. Management of contact and attention can be performed using different modalities (Allwood et al., 1991). This draws on findings by (Allwood et al., 1991) and extends the schema adopted by (Gill et al., 2000) who annotated the body move category Attempt-Contact.

The category Signal non-perception (SNP) is similar to the DAMSL category Signal non-understanding (SNU), but fo-

	2.56	2.57	2.58	2.59	3.00
user		robot		this is a book	
com-act		req-att		ref, assert	
robot					
com-act					
u-gest			point		
com-act			ref		
u-gaze		robot	object	robot	



Figure 5: Different corpus data visualized as a score annotation similar to what it may look like in the Anvil tool developed by Kipp (2004). Here we have simplified the image to make it appear better in print.

cuses on the (reported) perceptual status of the participant. A typical example found in the corpus is "I cannot see you" uttered by the robot whenever it lost track of the user.

We have annotated sequences when the user is paying close attention by looking at the robot with the category Monitor. By paying attention to the robot, the user displays a basic positive level of willingness to interact.

Reference: We have annotated events of reference using the category Reference, knowing that there probably is a need to refine this further, for instance, (Gill et al., 2000) uses the more restricted type Demonstrative reference (Dem-Ref). But as this only is used for non-verbal referencing we have chosen the less specific category Reference which we aim to analyze in-depth to arrive at more precise scheme at a later stage.

Emotional display: There are very few occurrences of emotional displays in the corpus. We have annotated obvious examples of emotional display when we have deemed them as being relevant for the communication, e.g., user laughing when the robot speaks something that appears as ill-phrased or out of context. Another category that is related to emotions is Emphasis (Emph), i.e., where a gesture is stressing some aspect of a contribution (e.g., protruding finger during pointing at an object). Furthermore, instances of self-touch, e.g., touching the face or lips have been observed and annotated because it may signal the emotional state of the user or be seen as a sign of invasion of personal space (Sommer, 1969).

3.2. Supportive annotations

The *supportive* or *context-providing* annotations form a heterogeneous set of resources that can be used for different purposes during analysis. For instance, our interest related to the spatiality dimension of embodiment make data on positioning and spatial distance important to analyze users movement patterns. Another interest lies in the relation between dialogue acts and physical acts (Traum, 2000) and how we may use them to analyze the possible goals that the user and robot can possess respectively. For this we need to have a scene overview, and be able to determine the intentional attention of users by analyzing their gaze patterns. We have also annotated the general task that is going on at a specific time to be used as general background information

and organization of tasks at a higher level.

Gaze: We have also annotated the general direction of the user's gaze in terms of domain related concepts, i.e., the robot itself, object ("tv", "telephone") and locations ("corner"). We have also noted that the user looks around in the room when looking for something or while thinking. In the gaze-track this has simply been annotated as "room".

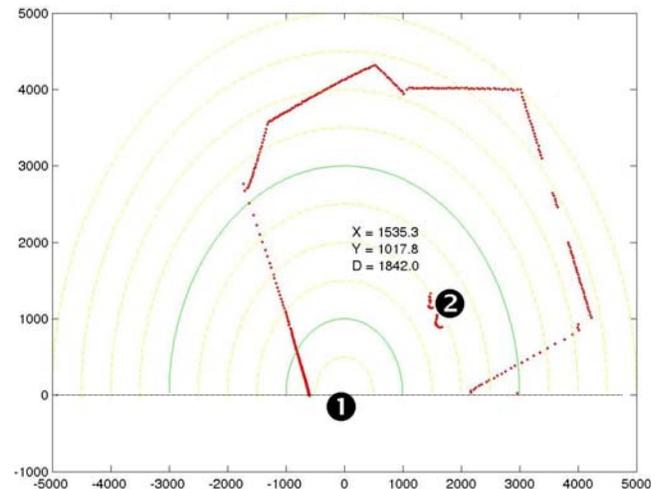


Figure 6: A visualization of laser data using an algorithm for tracking the user (Topp and Christensen, 2005), showing the robot center (1) and the tracked user (2). The walls sensed by the robot are displayed as a (red) dotted line. What appears as two holes in the wall on the right hand side of the image is the shadow of the user's legs (2).

Positioning and spatial distance: We are annotating spatial formation, i.e., the dynamic aspects of spatial arrangements using a taxonomy based on Kendon's F-formation system (Kendon, 1990). This system is based upon the observation that certain patterns of posture and orientation between participants are maintained during interaction.

We are also coding interpersonal distances according to the classification proposed by Hall (1966). Social interaction is based upon and governed by four interpersonal distances: *intimate* (0–1.5 feet), *personal* (1.5–4 feet), *social* (4–12 feet), and *public* (>12 feet).

Scenario and users		CORE ANNOTATIONS			CONTEXT/TIME ALIGNED DATA	
		Speech	Comm. acts	Gesture	Task	Posture & Positioning
Single room	22 users interacting ~15 minutes, task constraints given by system	×	×	×	Objects and locations	Hall distances and F-formations, data from laser range finder
Multiple room	5 users interacting ~15 minutes, accepting strategy used by the system	×	×	×	Locations stored in system logs	
		MEDIA			BACKGROUND DATA	
		Video	Audio	Webcam	On-board cam	Task-descriptions
Single room	MiniDV (25 fps)	Stereo onboard (16 kHz), MiniDV (48 kHz)	One web cam in each corner (1 fps)		Time-aligned task annotations	As data file or text document
Multiple room	Handheld MiniDV (25 fps)	MiniDV (48 kHz)		MiniDV (25 fps)		Interview/Videotaped

Table 1: Corpus annotations

Both the F-formation system and social distances provide discrete representations for spatiality. Therefore we are also collecting and synchronizing laser data and video recordings to be able to study this topic further.

The data from the laser range finder is stored as raw data files with time stamps. This allows for development of different types of applications, e.g., tools for visualization or tracking algorithms. In figure 6 a tracking algorithm has been applied to the data showing the legs of the user as two half circles close to the point indexed (2) in the image.

Task and scene overview: We have annotated tasks on a high level, e.g., as categories related to the general services provided by the system: FOLLOW, SHOW, FIND, GREET, etc. The aim is to provide background information and to visualize organization related to the users' way of solving the task. Another means of providing a general sense of what is going on in the data are images from four network web-cams, that are time aligned to the video. The web-cams were placed in each corner of the single room scenario. It is thus possible to get several perspectives of the scene, and disambiguate the scene linked to the corpus using the timecode. In the multiple room scenario this coverage was not possible to achieve, since it would require a huge amount of cameras. Instead a handheld video-camera provided another perspective on the interaction.

Text descriptions and questionnaire data: During the analysis of spatiality we also wrote down observations on events in the session. These text descriptions have been

time aligned so they can be used as links to specific points of interest in the data. Answers to questionnaires administered to users concerning their attitudes towards the system are also available as a data file.

3.3. Conclusions and future work

We have described the process of developing a contextualized corpus for human-robot interaction. By providing links to data sources, e.g., laser data and text descriptions and data that is annotated using well established taxonomies we aim to support activities related to the development of a cognitive robot. In the near future we will use this corpus in the development of adaptive models of users' style of communication and to study communicative behavior related to the spatial configuration of the robot and user.

4. Acknowledgments

The work described in this paper was conducted within the EU Integrated Project COGNIRON ('The Cognitive Robot Companion' - www.cogniron.org) and was funded by the European Commission Division FP6-IST Future and Emerging Technologies under Contract FP6-002020.

5. References

James Allen and Mark Core. 1997. Draft of DAMSL: Dialog Act Markup in Several Layers. webpage. <http://www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual/>.

- Jens Allwood, Joakim Nivre, and Elisabet Ahlsén. 1991. On the semantics and pragmatics of linguistic feedback. Technical Report 64, Gothenburg Papers on Theoretical Linguistics.
- G. Bugmann, S. Lauria, T. Kyriacou, E. Klein, J. Bos, and K. Coventry. 2001. Using Verbal Instruction for Route Learning. In *Proceedings of 3rd British Conference on Autonomous Mobile Robots and Autonomous Systems: Towards Intelligent Mobile Robots (TIMR'2001)*, Manchester, April.
- Guido Bugmann, Ewan Klein, Stanislaw Lauria, and T. Kyriacou. 2004. Corpus-based robotics: A route instruction example. In *Proceedings of IAS-8*, pages 96–103, Amsterdam, NL, March 10–13.
- Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13:259–294.
- Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies - why and how. *Knowledge-Based Systems*, 6(4):258–256.
- Satinder P. Gill, Masahito Kawamori, Yasuhiro Katagiri, and Atsushi Shimojima. 2000. Role of body moves in dialogue. *International Journal of Language and Communication, RASK*, 12, April.
- Anders Green, Helge Hüttenrauch, and Kerstin Severinson Eklundh. 2004. Applying the Wizard-of-Oz framework to Cooperative Service Discovery and Configuration. In *13th IEEE International Workshop on Robot and Human Interactive Communication RO-MAN 2004*, pages 575–580, 20–22 Sept.
- Anders Green, Britta Wrede, Kerstin Severinson Eklundh, and Shuyin Li. 2006. Integrating Miscommunication Analysis in the Natural Language Interface Design for a Service Robot. submitted to IROS2006.
- Edward T. Hall. 1966. *The Hidden Dimension: Man's Use of Space in Public and Private*. The Bodley Head Ltd, London, UK.
- Helge Hüttenrauch, Kerstin Severinson Eklundh, Anders Green, and Elin Anna Topp. 2006. Investigating Spatial Relationships in Human-Robot interaction. Submitted to IROS2006.
- Adam Kendon. 1990. *Conducting interaction - Patterns of behavior in focused encounters. Studies in interactional sociolinguistics*. Press syndicate of the University of Cambridge, Cambridge, NY, USA.
- Michael Kipp. 2004. *Gesture Generation by Imitation - From Human Behavior to Computer Character Animation*. Dissertation.com, Boca Raton, Florida.
- M. W. Knudsen, Laila Dykjær, and Niels Ole Bernsen. 2001. Surveys of multimodal data resources, annotation schemes and tools. In *Proceedings of the COCOSDA'2001 Workshop on Language Resources and Technology Evaluation - Technical, Global and Regional Perspectives*, pages 135–146, Aalborg, Denmark, 2 September.
- Y. Koide, T. Kanda, Y. Sumi, K. Kogure, and H. Ishiguro. 2004. An Approach to Integrating an Interactive Guide Robot with Ubiquitous Sensors. In *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004. (IROS 2004)*, volume 3, pages 2500–2505, 28 Sept/ 2 Oct.
- Jan Frederik Maas and Britta Wrede. 2006. BITT: A Corpus for Topic Tracking Evaluation on Multimodal Human-Robot-Interaction. In *Proceedings of the Fifth international conference on Language Resources and Evaluation LREC2006*.
- David Mausby, Saul Greenberg, and Richard Mander. 1993. Prototyping an Intelligent Agent through Wizard of Oz. In *INTERCHI'93*, pages 277 – 282. ACM, April.
- Florian Schiel, Silke Steininger, and Ulrich Türk. 2002. The SmartKom Multimodal Corpus at BAS. In *Proceedings of Second International Conference on Language Resources and Evaluation LREC2000*, pages 200–206.
- Candace L. Sidner, Cory D. Kidd, Christopher Lee, and Neal Lesh. 2004. Where to look: a study of human-robot engagement. In *IUI'04: Proceedings of the 9th international conference on Intelligent User Interfaces*, pages 78–84, New York, NY, USA. ACM Press.
- Robert Sommer. 1969. *Personal Space*. Prentice-Hall.
- Elin A. Topp and Henrik I. Christensen. 2005. Tracking for Following and Passing Persons. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2005)*, Edmonton, Alberta, August.
- Elin Anna Topp, Helge Hüttenrauch, Henrik Christensen, and Kerstin Severinson Eklundh. 2006. Acquiring a shared environment representation. In *In Proceedings of HRI2006 1st annual conference on Human-Robot Interaction*, Salt Lake City, UT, USA, March 2–3. ACM.
- David R. Traum. 2000. 20 Questions for Dialogue Act Taxonomies. *Journal of Semantics*, 17(1):7–30.
- Luis Villaseñor, Antonio Mass, and Luis Pineda. 2000. A multimodal dialog contribution coding scheme. In *The First EAGLES/ISLE Workshop on Meta-Description and Annotation Schemes for Multimodal/Multimedia Language Resources in conjunction with the Second International Conference on Language Resources and Evaluation LREC 2000*, Greece, May.
- Joerg C. Wolf and Guido Bugmann. 2005. Multimodal Corpus Collection for the Design of User-Programmable Robots. In *TAROS 2005 Towards Autonomous Robotic Systems Incorporating the Autumn Biro-Net Symposium*, 12th–14th September.