The Dynamics of Action Corrections in Situated Interaction

Antoine Raux
Honda Research Institute USA
Mountain View, CA, USA.
araux@honda-ri.com

Mikio Nakano
Honda Research Institute Japan
Wako, Japan
nakano@jp.honda-ri.com

Abstract

In spoken communications, correction utterances, which are utterances correcting other participants utterances and behaviors, play crucial roles, and detecting them is one of the key issues. Previously, much work has been done on automatic detection of correction utterances in human-human and human-computer dialogs, but they mostly dealt with the correction of erroneous utterances. However, in many real situations, especially in communications between humans and mobile robots, the misunderstandings manifest themselves not only through utterances but also through physical actions performed by the participants. In this paper, we focus on action corrections and propose a classification of such utterances into Omission, Commission, and Degree corrections. We present the results of our analysis of correction utterances in dialogs between two humans who were engaging in a kind of on-line computer game, where one participant plays the role of the remote manager of a convenience store, and the other plays the role of a robot store clerk. We analyze the linguistic content, prosody as well as the timing of correction utterances and found that all features were significantly correlated with action corrections.

1 Introduction

Recent progress in robot technology made it a reality to have robots work in offices and homes, and spoken dialog is considered to be one of the most desired interface for such robots. Our goal is to build a spoken dialog interface for robots that can move around in an office or a house and execute tasks according to humans' requests.

Building such spoken dialog interface for robots raises new problems different from those of traditional spoken/multimodal dialog systems. The intentions behind human utterances may vary depending on the situation where the robot is and the situation changes continuously not only because the robot moves but also because humans and objects move, and human requests change. In this sense human-robot interaction is *situated*.

Of the many aspects of situated interaction, we focus on the timing structure of interaction. Although traditional spoken dialog systems deal with some timing issues such as turn-taking and handling barge-ins, timing structure in human-robot interaction is far more complex because the robot can execute physical actions and those actions can occur in parallel with utterances.

In this work we are concerned specifically with corrections in situated interaction. In joint physical tasks, human corrective behavior, which allows to repair discrepancies in participants' mutual understanding, is tightly tied to actions.

While past work on non-situated spoken dialog systems has shown the necessity and feasibility of detecting and handling corrections (Kitaoka et al., 2003; Litman et al., 2006; Gieselman and Ostendorf, 2007; Cevik et al., 2008), most of these models assume that corrections target past utterances and rely on a strict turn-based structure which is frequently violated in situated interaction. When dialog is interleaved with physical actions, the specific timing of an utterance relative to other utterances and actions is more relevant than the turn sequence.

In this paper, we propose a classification of errors and corrections in physical tasks and analyze the properties of different types of corrections in the context of human-human task-oriented interactions in a virtual environment. The next section gives some characteristics of corrections in situated interaction. Section 3 describes our experi-

Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 165–174,
The University of Tokyo, September 24-25, 2010.

```
Alice : Put it right above    (1)
        the lamp stand
Bob   : Here?                 (2)
Alice : A little bit more     (3)
        to the right.
(Bob Moves the frame left)    (4)
Alice : No the right!         (5)
(Bob Moves the frame right)   (6)
Alice : More...               (7)
        Alright, that's ...   (8)
(A bee flies next to Bob)     (9)
Alice : Watch out! That bee is
        going to sting you!  (10)
```

Figure 1: *Example dialog from a situated task.*

mental set up and data collection effort. Section 4 presents the results of our analysis of corrections in terms of timing, prosodic, and lexical features. These results are discussed in Section 5.

2 Corrections in Situated Tasks

2.1 Situated Tasks

We define a situated task as one for which two or more agents interact in order to perform physical actions in the (real or virtual) world. Physical actions involve moving from one place to another, as well as manipulating objects. In many cases, interaction happens simultaneously with physical actions and can be affected by them, or by other external events happening in the world. For example, Figure 1 shows an extract of a (constructed) dialog where one person (Alice) assists another (Bob) while he hangs a picture frame on a wall.

This interaction presents some similarities and differences with unimodal, non-situated dialogs. In addition to standard back-and-forth turn-taking as in turns 1-3, this example features utterances by Alice which are not motivated by Bob's utterances, but rather by (her perception of) his actions (e.g. utterance 5 is a reaction to action 4), as well as external events such as 9, which triggered response 10 from Alice. Therefore the **content** of Alice's utterances is dependent not only on Bob's, but also on events happening in the world. Similarly, the **timing** of Alice's utterances is not only conditioned on Bob's speech, prosody, etc, but also on asynchronous world events.

Robots and other agents that interact with people in real world situations need to be able to account for the impact of physical actions and world events on dialog. In the next section and the rest of this paper, we focus on correction utterances and how situational context affects how and when speakers produce them.

2.2 Corrections

Generally speaking, a correction is an utterance which aims at signaling or remedying a misunderstanding between the participants of a conversation. In other word, corrections help (re)establish common ground (Clark, 1996).

2.2.1 Previous Work

There are many dimensions along which corrections can be analyzed and many researchers have addressed this issue. Conversational Analysis (CA) has, from its early days, 1 concerned itself with corrections (usually called repairs in CA work) (Schegloff et al., 1977).

More recently, spoken dialog systems researchers have investigated ways to automatically recognize corrections. For instance, Litman et al. (2006) exploited features to automatically detect correction utterances. In addition, several attempts have been made to exploit the similarity in speech sounds and speech recognition results of a correction and the previous user utterance to detect corrections (Kitaoka et al., 2003; Cevik et al., 2008).

Going beyond a binary correction/non-correction classification scheme, Levow (1998) distinguished corrections of misrecognition errors from corrections of rejection errors and found them to have different prosodic features. Rogdriguez and Schlangen (2004) and Rieser and Moore (2005) classify corrections according to their form (e.g. Repetition, Paraphrase, Addition of information...) and function. The latter aspect is mostly characterized in terms of the source of the problem that is being corrected using models of communication such as that of Clark (1996).

In all of this very rich literature, corrections are assumed to target utterances from another participant (or even oneself, in the case of self-repair) that conflict with the hearer's expectations. While some work on embodied conversational agents (Cassell et al., 2001; Traum and Rickel, 2002) does consider physical actions as possible cues to errors and corrections, the actions are typically communicative in nature (e.g. nods, glances, gestures). Comparatively, there is extremely little work on corrections that target task actions.

A couple of exceptions are Traum et al. (1999), who discuss the type of context representation needed to handle action corrections, and Funakoshi and Tokunaga (2006), who present a model for identifying repair targets in human-robot command-and-control dialogs. While important, these papers focus on theoretical planning aspects of corrections whereas this paper focuses on an empirical analysis of human conversational behavior.

2.2.2 Action Corrections in Situated Interaction

As seen above, the vast majority of prior work on corrections concerned corrections of previous (erroneous) utterances (i.e. utterance corrections). In contrast, in this paper, we focus exclusively on corrections that target previous physical actions (i.e. action corrections).

While some classification schemes of utterance corrections are applicable to task corrections (e.g. those based on the form of the correction itself), we focus on differences that are specific to action corrections.

Namely, we distinguish three types of action errors and their related action corrections:

Commission errors occur when Bob performs an action that conflicts with Alice's expectation. Action 4 of Figure 1 is a commission error, which is corrected in turn 5.

Omission errors occur when Bob fails to react to one of Alice's utterances. A typical way for Alice to correct an omission error is to repeat the utterance to which Bob did not react.

Degree errors occur when Bob reacts with an appropriate action to Alice's utterance but fails to completely fulfill Alice's goal. This is illustrated by Alice's use of "More" in turn 7 in response to Bob's insufficient action 6.

Figure 3 illustrates the three error categories based on extracts from the corpus.

In some ways, the dichotomy Commission errors/Omission errors parallels that of Misrecognitions/Rejections by Levow (1998). This type of classification is also commonly used to analyze human errors in human factors research (Wickens et al., 1998). In addition to these two categories, we added the Degree category based on our observation of the data we collected. This aspect is somewhat specific to certain kinds of physical actions (those that can be performed to different degrees, as opposed to binary actions such as "opening the door"). However, it seems general enough to be applied to many collaborative tasks relevant to robots such as guidance, tele-operation, and joint construction.

For an automated agent, being able to classify a user utterance into one of these four categories (including non-action-correction utterances) could be very useful to make fast, appropriate decisions such as canceling the current action, or asking a clarification question to the user. This is important because in human-robot interaction, responsiveness to a correction can be critical in avoiding physical accidents. For instance, if the robot detects that the user issued a commission error correction, it can stop performing its current action even before understanding the details of the correction.

In the rest of the paper, we analyze some lexical, prosodic and temporal characteristics of action corrections in the context of human-human conversations in a virtual world.

3 The Konbini Domain and System

3.1 Simulated Environments for Human-Robot Interaction Research

One obstacle to the empirical study of situated interaction is that it requires a fully functional sophisticated robot to collect data and conduct experiments. Most such complex robots are still fragile and thus it is typically challenging to run user studies with naive subjects without severely limiting the tasks or the scope of the interaction. Another issue which comes with real world interaction is that it is difficult for the experimenter to control or monitor the events that affect the interaction. Most of the time, an expensive manual annotation of events and actions is required (see Okita et al. (2009) for an example of such an experimental setup).

To avoid these issues, robot simulators have been used. Koulouri and Lauria (2009) developed a simulator to collect dialogs between human and simulated robot using a Wizard-of-Oz method. The human can see a map of a town and teaches the robot a route and the operator operates the robot but he/she can see only a small area around the robot in the map. However, the dialog

is keyboard-based, and the situation does not dynamically change in this setting, making this approach unsuitable to the study of timing aspects. Byron and Fosler-Lussier (2006) describe a corpus of spoken dialogs collected in a setting very similar to the one we are using but, again, the environment appears to be static, thus limiting the importance of the timing of actions and utterances.

In this section, we describe a realistic, PC-based virtual world that we used to collect human-human situated dialogs.

3.2 Experimental Setup

In our experiment, two human participants collaborate in order to perform certain tasks pertaining to the management of a small convenience store in a virtual world. The two participants sit in different rooms, both facing a computer that presents a view of the virtual store. One of the participants, the Operator (O) controls a (simulated) humanoid robot whose role is to answer all customer requests. The other participant plays the role of a remote Manager (M) who sees the whole store but can only interact with O through speech.

Figure 2 shows the Operator and Manager views. M can see the whole store at any time, including how many customers there are and where they are. In addition, M knows when a particular customer has a request because the customer's character starts blinking (initially green, then yellow, then red, as time passes). M's role is then to guide O towards the customers needing attention.

On the other hand, O sees the world through the "eyes" of the robot, whose vision is limited both in terms of field of view (90 degrees) and depth (degradation of vision with depth is produced by adding a virtual "fog" to the view). When approaching a customer who has a pending request, O's view display the customer's request in the form of a caption.[1] O can act upon the virtual world by clicking on certain object such as items on the counter (to check them out), machines in the store (to repair them when needed), and various objects littering the floor (to clean them up). Each action takes a certain amount of time to perform (between 3 and 45 seconds), indicated by a progress bar that decreases as O keeps the pointer on the target object and the left mouse button down. Once the counter goes to zero the action is completed and the participants receive 50 (for partially fulfilling a customer request) or 100 points (for completely fulfilling a request).

When the session begins, customers start entering the store at random intervals, with a maximum of 4 customers in the store at any time. Each customer follows one of 14 predefined scenarios, each involving between 1 and 5 requests. Scenarios represent the customer's moves in terms of fixed way points. As a simplification, we did not implement any reactive path planning. Rather, the experimenter, sitting in a different room than either subject has the ability to temporarily take control of any customer to make them avoid obstacles.

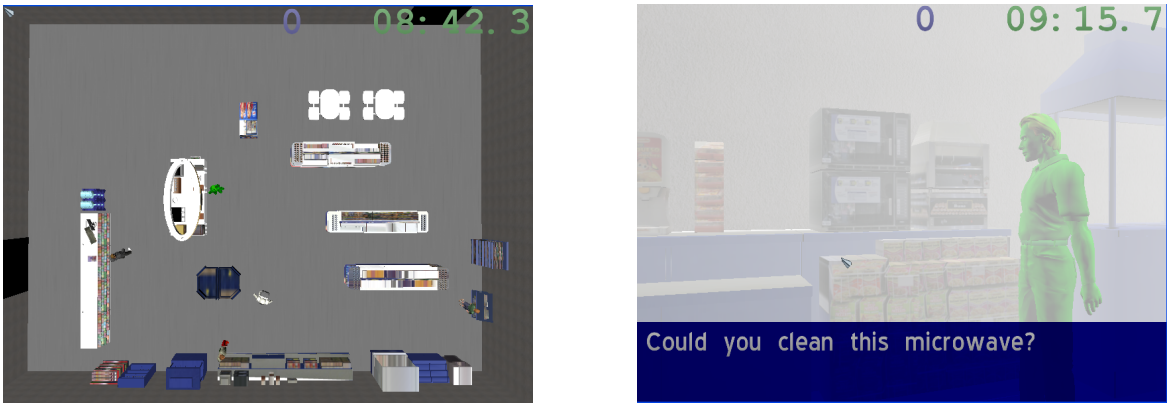
3.3 System Implementation: the Siros architecture

The experimental system described above was implemented using Siros (SItuated RObot Simulator) a new client/server architecture developed at Honda Research Institute USA to conduct human-robot interaction research in virtual worlds. Siros is similar to the architectures used by certain online video games. The server's role is to manage the virtual world and broadcast world updates to all clients so that they can be rendered to the human participants. The server receives commands from the Operator client (robot moves), runs the simulated customers according to the scenarios, and maintains the timer and the score. Anytime the trajectory of an entity (robot, customer, object) changes, the server broadcasts the related information, including entity location, orientation, and speed, to all clients.

Clients are in charge of rendering a given view of the virtual world. Rendering itself is performed by the open source Ogre 3D engine (Open Source 3D Graphics Engine, 2010). In addition, clients handle all required user interaction such as robot control and mouse-based object selection. All network messages and user actions are logged into a text file for further processing. Finally, clients have the ability to log incoming audio to a wave file, allowing synchronization between the audio signal, the user actions, and virtual world events. Spoken communication itself is handled by an external VoIP client.[2]

[1] No actual speech interaction happens between the Operator and the simulated customers.

[2] We used the open source Mumble/Murmur (Mumble/Murmur Project, 2010) system.



(a) Manager View

(b) Robot View

Figure 2: Screenshots of the Konbini data collection system.

3.4 The Konbini Corpus

3.4.1 Data Collection

Using the system described above, we collected data from 18 participants. There were 15 male and 3 female participants. All were adults living in the United States, fluent in English. All were regular users of computers but their experience with online games was diverse (from none at all to regular player). All were co-workers (associates or interns) at Honda Research Institute USA, and thus they knew each other fairly well.

Participants were randomly paired into teams. After being given the chance to read a brief introduction to the experiment's design and goals, the participants did two two-minute practice sessions to familiarize themselves with the task and control. To avoid providing too much information about the layout of the store from the start, the practice sessions used a different virtual world than the experimental sessions. The participants switched roles between the two practice sessions to get a sense of what both roles entailed (Manager and Operator). After these sessions, the team did one 10-minute experimental session, then switched roles once again and did another 10-minute session. Because the layout of the store was kept the same between the two experimental sessions, the first session represents a condition in which the Operator learns the store layout as they are performing the tasks, whereas the second session corresponds to a case where the Operator already has knowledge of the layout. Overall, 18 10-minute sessions were collected, including audio recordings as well as timestamped logs of world updates and operator actions.

3.4.2 Annotation

All recordings were orthographically transcribed and checked. The first author then segmented the transcripts into dialog acts (DAs). A DA label was attached to each act, though this information is not used in the present paper.

Subsequently, the first author annotated each semantic unit with the action correction labels described in section 2: Non-correction, Omission Correction, Commission Correction, Degree Correction. This annotation was done using the Anvil video annotation tool,[3] which presented audio recordings, transcripts, a timeline of operator actions, as well as videos of the computer screens of the participants. Only Manager DAs were annotated for corrections. The second author also annotated a subset of the data in the same way to evaluate inter-annotator agreement. Cohen's kappa between the two annotators was 0.67 for the 4-class task, and 0.76 for the binary task of any-action-correction vs non-action-correction, which is reasonable, though not very high, indicating that correction annotation on this type of dialogs is a non-trivial task, even for human annotators.

4 Analysis of Action Corrections

4.1 Overview

The total number of DAs in the corpus is 6170. Of those, 826 are corrections and 5303 are non-corrections. Overall, corrections account thus for 13.4% of the dialog acts. The split among the different correction classes is roughly equal as shown in Table 5 given in appendix. We found however significant differences across participants, in terms

[3] http://www.anvil-software.de

of total number of DAs (form 162 to 516), proportion of corrections among those DAs (from 6.8% to 30.6%), as well as distribution among the three types of action corrections.

In this section, we present the results of our statistical analysis of the correlation between a number of features and correction type. To evaluate statistical significance, we performed a one-way ANOVA using each feature as dependent variable and the correction type as independent variable. All features described here were found to significantly correlate with correction type.

4.2 Features Affecting Corrections

4.2.1 Timing

For each manager DA, we computed the time since the beginning/end of the previous manager and operator DAs, as well as of operator's actions (walk/turn). To account for reaction time, and based on our observations we ignored events happening less than 1 second before a DA.

Table 1 shows the mean durations between these events and a Manager DA, depending on the act's correction class. All corrections happen closer to Manager dialog acts than non-corrections, which reflects the fact that corrections typically occur in phases when the Manager gives instructions, as well as the fact that the Manager often repeats corrections. Commission and Degree corrections are produced closer to Operator actions than either non-corrections or Omission corrections. This reflects the fact that both Commission and Degree corrections are a reaction to an event that occurred (the Operator moved or stopped moving unexpectedly), whereas Omission corrections address a *lack* of action from the Operator, and act therefore as a "time-out" mechanism.

To better understand the relationship between moves and the timing of corrections, we computed the probability of a given DA to be an Omission, Commission and Degree correction as a function of the time elapsed since the Operator last started to move. Figure 4 shows the results.

Time since last...	NC	O	C	D
Mgr. DA	3.4 s	2.4 s	2.8 s	2.6 s
Ope. DA	5.8 s	6.7 s	6.5 s	7.5 s
Ope. move start	3.8 s	3.1 s	2.3 s	2.5 s
Ope. move end	3.9 s	3.3 s	2.7 s	2.3 s

Table 1: *Mean duration between dialog acts and Operator movements and the beginning of different corrections.*

Feature	Non-Corr	Om	Com	Deg
Perc. voiced	0.48	0.46	0.55	0.53
Min F0	-0.61	-0.41	-0.40	-0.56
Max F0	0.81	0.68	1.02	0.46
Mean F0	-0.03	0.12	0.28	-0.05
Min Power	-1.35	-1.24	-1.18	-1.55
Max Power	0.85	0.89	1.14	0.62
Mean Power	-0.03	0.09	0.24	-0.2

Table 2: *Mean Z-score of prosodic features for different correction classes.*

The probability of a DA being an Omission correction is relatively stable over time. This is consistent with the fact that Omission corrections are related to lack of action rather than to a specific action to which the Manager reacts. On the other hand, the probability of a Commission, and to lesser extent, Degree correction sharply decreases with time after an action.

4.2.2 Prosody

We extracted F0 and intensity from all manager audio files using the Praat software (Boersma and Weenink, 2010). We then normalized pitch and intensity for each speaker using a Z-transform in order to account for individual differences in mean and standard deviation. For each DA, we computed the minimum, maximum, and mean pitch and intensity, using values from voiced frames.

Table 2 shows the mean Z-score of the prosodic features for the different correction classes. Commission corrections feature higher pitch and intensity than all other classes. This is due to the fact that such corrections typically involve a higher emotional level, when the Manager is surprised or even frustrated by the behavior of the Operator. In contrast, Degree corrections, which represent a smaller mismatch between the Operator's action and the Manager's expectations are more subdued, with mean power and intensity values lower than even those of non-corrections.

4.2.3 Lexical Features

In order to identify potential lexical characteristics of correction utterances, we created binary variables indicating that a specific word from the vocabulary (804 distinct words in total) appears in a given DA based on the manual transcripts. We computed the mutual information of those binary

variables with DA's correction label.

Figure 3 shows the 10 words with highest mutual information. Not surprisingly, negative words ("NO", "DON'T"), continuation words ("MORE", "KEEP") are correlated with respectively commission and degree corrections. On the other hand, positive words ("OKAY", "YEAH") are strong indicators that a DA is not a correction.

Another lexical feature we computed was a flag indicating that a certain Manager DA is an exact repetition of the immediately preceding Manager DA. The intuition behind this feature is that corrections often involve repetitions (e.g. "Turn left [Operator turns right] Turn left!"). Overall, 10.6% of the DAs are repetitions. This number is only 6.4% on non-corrections but jumps to 45.6%, 22.5%, and 43.4% on, respectively, Omission, Commission, and Degree corrections. This confirms that, as for utterance corrections, detecting exact repetitions could prove useful for correction classification.

4.2.4 ASR Features

Since our goal is to build artificial agents, we investigated features related to automatic speech recognition. We used the Nuance Speech Recognition System v8.5. Using cross-validation, we trained a statistical language model for each correction category on the transcripts of the training portion of the data. We then ran the recognizer sequentially with all 4 language models, which generated a confidence score for each category.

Table 4 shows the mean confidence scores obtained on DAs of each class using a language model trained on specific classes. While the matching LM gives the highest score for any given class, some classes have consistently higher scores than others. In particular, Commission corrections receive low confidence scores, which might hurt the effectiveness of these features. Indeed, lexical content alone might not be not enough to distinguish non-corrections and various categories of corrections since the same expression (e.g. "Turn left") can express a simple instruction, or any kind of correction, depending on context.

LM \ Corr.	NC	O	C	D
Non-Correction	32.3	28.5	25.0	29.5
Omission	24.0	30.0	23.3	27.2
Commission	26.6	29.8	25.7	27.9
Degree	24.2	28.7	23.9	32.6

Table 4: *Mean ASR confidence score using class-specific LMs.*

5 Discussion

The results provide support for the correction classification scheme we proposed. Not only do corrections differ in many respects from non-correction utterances, but there are also significant differences between Omission, Commission, and Degree corrections. Timing features seem most useful to distinguish Commission and Degree corrections from Omission corrections and non-corrections. Emphasized prosody (high pitch and energy) is a particularly strong indicator of Commission, as well as Omission corrections. Lexical cues could be useful to all categories, provided the speech recognizer is accurate enough to recognize them, which is particularly challenging on this data given the very conversational nature of the speech. Finally, ASR scores are also potentially useful features, particularly for Omission and Degree corrections.

One advantage of timing over all other features discussed here is that timing information is available *before* the correction is actually uttered. This means that such information could be used to allow fast reaction, or to prime the speech recognizer based on the instantaneous probability of the different classes of correction.

6 Conclusion

In this paper, we analyzed correction utterances in the context of situated spoken interaction within a virtual world. We proposed a classification of action correction utterances into Omission, Commission, and Degree corrections. Our analysis of human-human data collected using a PC-based simulated environment shows that the three types of corrections have unique characteristics in terms of prosody, lexical features, as well as timing with regards to physical actions. These results can serve as the basis for further investigations into automatic detection and understanding of correction utterances in situated interaction.

References

Paul Boersma and David Weenink. 2010. Praat: doing phonetics by computer, http://www.fon.hum.uva.nl/praat.

Word (W)	$P(Non-Corr\|W)$	$P(Om\|W)$	$P(Com\|W)$	$P(Deg\|W)$
MORE	0.41	0.01	0.02	0.56
NO	0.55	0.04	0.33	0.07
RIGHT	0.67	0.15	0.04	0.14
TURN	0.69	0.17	0.06	0.08
LEFT	0.65	0.18	0.07	0.10
OKAY	0.99	0.00	0.00	0.00
YEAH	0.99	0.00	0.00	0.00
DON'T	0.59	0.01	0.33	0.07
WAY	0.49	0.05	0.42	0.04
KEEP	0.79	0.03	0.03	0.15

Table 3: *Keywords with highest mutual information with correction category.*

Donna K. Byron and Eric Fosler-Lussier. 2006. The OSU Quake 2004 corpus of two-party situated problem-solving dialogs. In *Proc. 15th Language Resource and Evaluation Conference (LREC'06)*.

Justine Cassell, Timothy Bickmore, Hannes Hgni Vilhjámsson, and Hao Yan. 2001. More Than Just a Pretty Face: Conversational Protocols and the Affordances of Embodiment. *Knowledge-Based Systems*, 14:55–64.

Mert Cevik, Fuliang Weng, and Chin hui Lee. 2008. Detection of repetitions in spontaneous speech dialogue sessions. In *Proc. Interspeech 2008*, pages 471–474.

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

Kotaro Funakoshi and Takenobu Tokunaga. 2006. Identifying repair targets in action control dialogue. In *Proc. EACL 2006*, pages 177–184.

Petra Gieselman and Mari Ostendorf. 2007. Problem-Sensitive Response Generation in Human-Robot Dialogs. In *Proc. SIGDIAL 2002*.

Norihide Kitaoka, Naoko Kakutani, and Seiichi Nakagawa. 2003. Detection and Recognition of Correction Utterance in Spontaneously Spoken Dialog. In *Proc. Eurospeech 2003*, pages 625–628.

Theodora Koulouri and Stanislao Lauria. 2009. Exploring miscommunication and collaborative behaviour in human-robot interaction. In *Proc. SIGDIAL 2009*, pages 111–119.

Gina-Anne Levow. 1998. Characterizing and recognizing spoken corrections in human-computer dialogue. In *Proc. COLING-ACL '98*, pages 736–742.

Diane Litman, Julia Hirschberg, and Marc Swerts. 2006. Characterizing and predicting corrections in spoken dialogue systems. *Computational Linguistics*, 32(3):417–438.

The Mumble/Murmur Project. 2010. http://mumble.sourceforge.net.

Sandra Y Okita, Victor Ng-Thow-Hing, and Ravi K Sarvadevabhatla. 2009. Learning Together: ASIMO Developing an Interactive Learning Partnership with Children. In *Proc. RO-MAN 2009*.

OGRE Open Source 3D Graphics Engine. 2010. http://www.ogre3d.org.

Verena Rieser and Johanna Moore. 2005. Implications for generating clarification requests in task-oriented dialogues. In *Proc. 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*, pages 239–246.

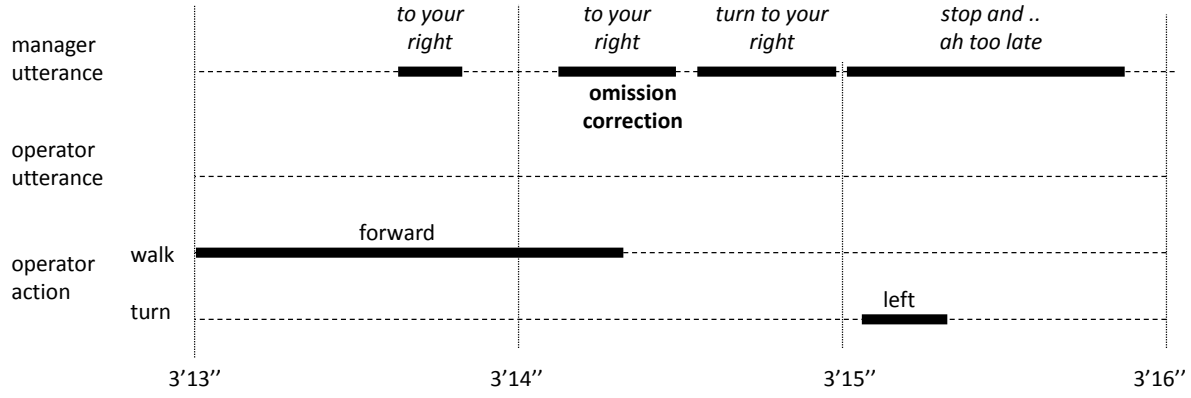
Kepa Josepa Rogdriguez and David Schlangen. 2004. Form, intonation and function of clarification requests in german task-oriented spoken dialogues. In *Proc. 8th Workshop on the Semantics and Pragmatics of Dialogue (CATALOG'04)*.

Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.

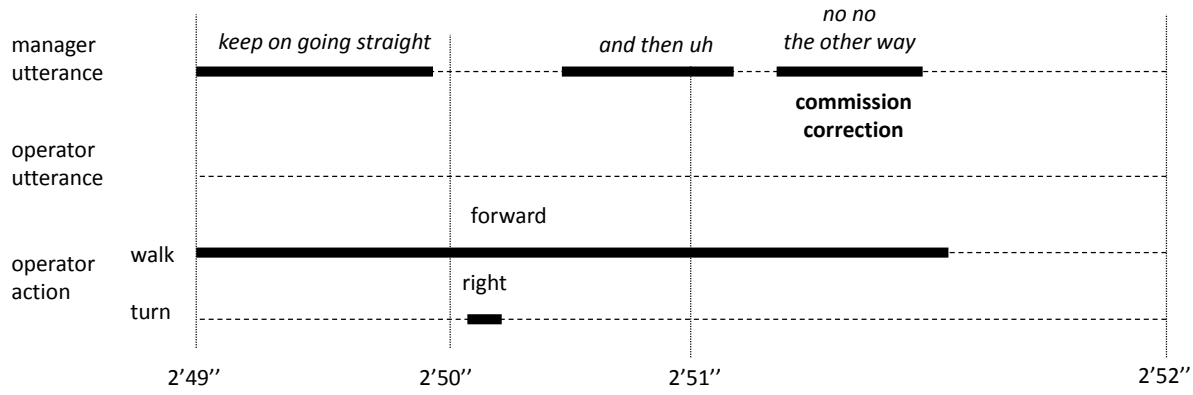
David Traum and Jeff Rickel. 2002. Embodied agents for multiparty dialogue in immersive virtual worlds. In *Proc. International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS 2002)*, pages 766–773.

David R. Traum, Carl F. Andersen, Waiyian Chong, Darsana P. Josyula, Yoshi Okamoto, Khemdut Purang, Michael O'Donovan-Anderson, and Donald Perlis. 1999. Representations of Dialogue State for Domain and Task Independent Meta-Dialogue. *Electron. Trans. Artif. Intell.*, 3(D):125–152.

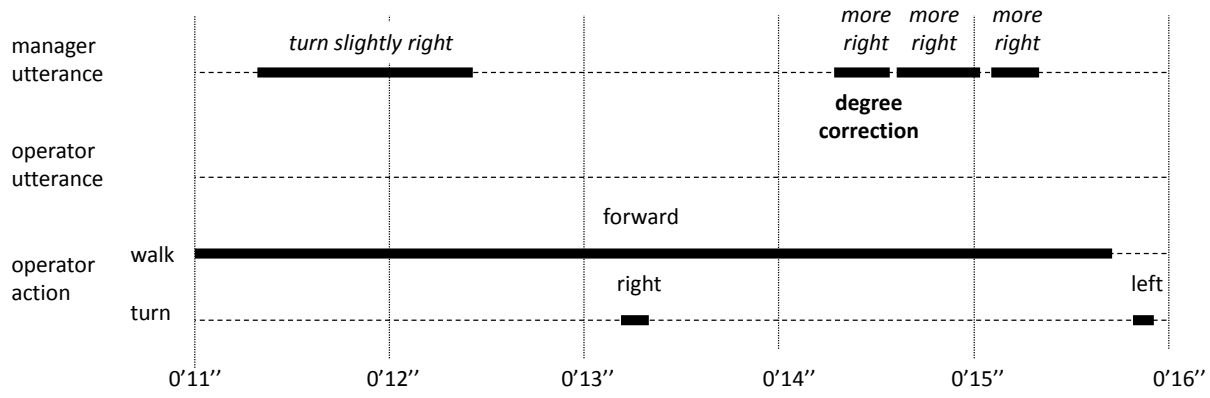
Christopher D. Wickens, Sallie E. Gordon, and Yili Liu. 1998. *An Introduction to Human Factors Engineering*. Addison-Wesley Educational Publishers Inc.



(a) Omission correction

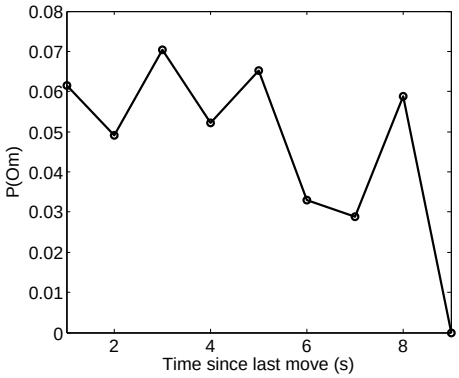


(b) Commission correction

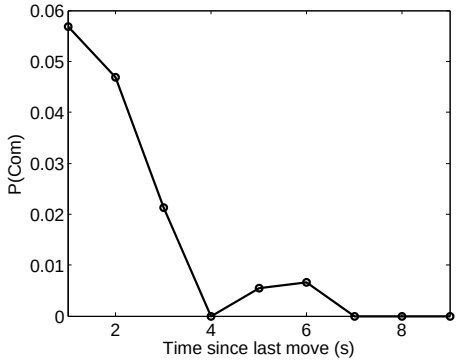


(c) Degree correction

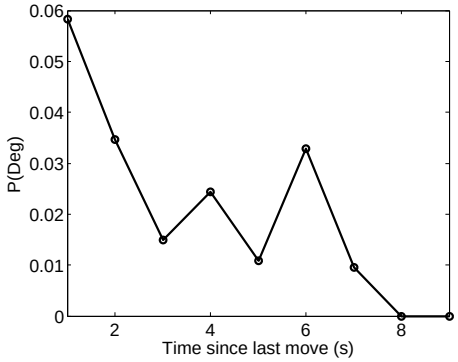
Figure 3: Example omission, commission, and degree errors and corrections. The corresponding videos can be found at `http://sites.google.com/site/antoineraux/konbini`.



(a) Omissions



(b) Commissions



(c) Degree

Figure 4: Evolution of the probability of occurrence of corrections over time after an Operator move.

Participant	Total	Non-Corr	Omission	Commission	Degree
Total	6170	5303 (86.6%)	298 (4.8%)	277 (4.5%)	251 (4.1%)
1	338	299 (88.5%)	19 (5.6%)	15 (4.4%)	5 (1.5%)
2	249	232 (93.1%)	10 (4.0%)	2 (0.8%)	0 (0.0%)
3	440	383 (87.0%)	25 (5.7%)	11 (2.5%)	9 (2.0%)
4	265	247 (93.2%)	4 (1.5%)	8 (3.0%)	0 (0.0%)
5	313	270 (86.3%)	15 (4.8%)	5 (1.6%)	17 (5.4%)
6	238	198 (83.2%)	22 (9.2%)	13 (5.5%)	5 (2.1%)
7	426	361 (84.7%)	30 (7.0%)	10 (2.3%)	23 (5.4%)
8	244	218 (89.3%)	3 (1.2%)	13 (5.3%)	9 (3.7%)
9	162	137 (84.6%)	4 (2.5%)	13 (8.0%)	8 (4.9%)
10	229	202 (88.2%)	6 (2.6%)	3 (1.3%)	12 (5.2%)
11	380	326 (85.8%)	16 (4.2%)	19 (5.0%)	19 (5.0%)
12	427	385 (90.2%)	16 (3.7%)	11 (2.6%)	15 (3.5%)
13	327	281 (85.9%)	5 (1.5%)	14 (4.3%)	27 (8.3%)
14	516	358 (69.4%)	38 (7.4%)	79 (15.3%)	39 (7.6%)
15	362	332 (91.7%)	13 (3.6%)	6 (1.7%)	11 (3.0%)
16	392	321 (81.9%)	34 (8.7%)	27 (6.9%)	10 (2.6%)
17	362	338 (85.4%)	19 (4.8%)	22 (5.6%)	17 (4.3%)
18	466	415 (89.1%)	19 (4.1%)	6 (1.3%)	25 (5.4%)

Table 5: *Frequency of the different types of corrections per participant.*