Construction and Analysis of a Multimodal Chat-talk Corpus for Dialog Systems Considering Interpersonal Closeness

Yoshihiro Yamazaki, Yuya Chiba, Takashi Nose, Akinori Ito

Graduate School of Engineering, Tohoku University

6-6-05, Aoba, Aramaki-Aza, Aoba-ku, Sendai, Miyagi, Japan

{yoshihiro.yamazaki.t2@dc, yuya@spcom.ecei, tnose@m, aito@spcom.ecei}.tohoku.ac.jp

Abstract

There are high expectations for multimodal dialog systems that can make natural small talk with facial expressions, gestures, and gaze actions as next-generation dialog-based systems. Two important roles of the chat-talk system are keeping the user engaged and establishing rapport. Many studies have conducted user evaluations of such systems, some of which reported that considering the relationship with the user is an effective way to improve the subjective evaluation. To facilitate research of such dialog systems, we are currently constructing a large-scale multimodal dialog corpus focusing on the relationship between speakers. In this paper, we describe the data collection and annotation process, and analysis of the corpus collected in the early stage of the project. This corpus contains 19,303 utterances (10 hours) from 19 pairs of participants. A dialog act tag is annotated to each utterance by two annotators. We compare the frequency and the transition probability of the tags between different closeness levels to help construct a dialog system for establishing a relationship with the user.

Keywords: Interpersonal closeness, Multimodal interaction, Chat-talk corpus, Dialog act

1. Introduction

In recent years, spoken dialog systems are widely used in smart speakers and communication robots. These systems have a function for non-task-oriented conversation such as chat-talk. In a field of the study on the chatbot, more and more researchers are committing to the development of the response generation technique (e.g., (Serban et al., 2016)). Progress in this area of research relies heavily on a large-scale text-based conversational corpus, such as chats on Twitter (Ritter et al., 2010). Thus, a large-scale multimodal dialog corpus is needed to accelerate research on the dialog systems that can handle social signals as well as verbal information, but well-prepared multimodal dialog data has not yet been sufficiently accumulated compared with the linguistic resources.

On the other hand, it is important for the chat-talk system to keep the user engaged and to establish rapport. Thus, many studies focused on dialog strategies for improving users' evaluation. Specifically, several studies reported that an effective way to improve the subjective evaluation is to consider the relationship with the user (Kageyama et al., 2018; Kim et al., 2012). These studies decided the behavior of the system empirically. However, many factors, including the choice of speech intention or non-verbal information are related to establishing rapport (Altman and Taylor, 1973; Vinciarelli et al., 2009), so a corpus-based analysis is needed in order to discover more effective dialog strategies.

To facilitate studies on multimodal chat-talk dialog systems, we are currently constructing a large-scale multimodal dialog corpus with information about the closeness between speakers. The corpus contains clear speech waveforms and video clips that capture the speaker from the second-person point of view. These are useful for constructing the components of the dialog system. In this paper, we describe the data collection and annotation process,

and analysis of the dialog data collected in the early stage of the project. Then, frequency and transition probability of the dialog acts are compared between different closeness levels as the first step of the analysis. Finally, we discuss the analysis results and its implications for the dialog strategy to establish a friendly relationship with the user.

2. Conventional corpora

To record the spontaneous conversational behavior of speakers, the dialog should be a human-human conversation. One example of a large scale human-human conversation corpus is the Corpus of Everyday Japanese Conversation (CEJC) being constructed by Koiso et al. (2018). The corpus contains various kinds of naturally occurring conversations in daily situations. For the same purpose, Oertel et al. (2010) collected as much spontaneous conversation as possible to construct the D64 Multimodal Conversational Corpus (D64). These corpora contain valuable information for analyzing natural human behavior. However, it is difficult to directly apply these corpora to the construction of dialog systems because spontaneous conversation between humans is greatly different from that with current dialog systems. One of the differences from a conversation between humans is that most conventional dialog systems assume one-to-one conversations. Although some studies collected multi-party conversations, such as the Multi-Party Robot corpus (MPR2016) (Funakoshi, 2018), one-to-one conversation is a reasonable target because recent successful applications such as, smart speakers also assume this style of conversation.

Regarding one-to-one conversation, Spontal is a large scale human-human conversation corpus (Edlund et al., 2010). This corpus contains dialog between two speakers sitting face to face. The video clips were recorded from the side and the speech data were recorded by close-talking microphones and omni-directional microphones. However, actual dialog systems capture the user from the

Table 1: Comparison of property of multimodal dialog corpora

Name	Language	Dialog participants	Participation	Dialog style	Audio channels	Video perspective	Annotation
IEMOCAP	English	Human-Human	One-to-one	Acted	Mixed	Third-person	TR, EM
CCDb	English	Human-Human	One-to-one	Natural	Mixed	Second-person	TR, DA, FE, SS
Spontal	Swedish	Human-Human	One-to-one	Natural	Mixed	Third-person	TR
D64	English	Human-Human	Multi-party	Highly natural	Mixed	Third-person	TR, AR, SD
SEMAINE corpus	English	Human-Machine	One-to-one	Natural	Separated	Second-person	TR, EM
CEJC	Japanese	Human-Human	Multi-party	Highly natural	Mixed	Third-person	TR, POS, DS, DA, IL
MPR2016	Multi-language	Human-Machine	Multi-party	Natural	Mixed	Third-person	TR, PS, A
Our corpus	Japanese	Human-Human	One-to-one	Natural	Separated	Second-person	TR, DA, Closeness

TR: transcript, DA: dialog act, POS: part of speech tag, DS: dependency structure, IL: intonation label, EM: emotion, AR: arousal, FE: facial expression, SS: social signal, PS: participation status, SD: social distance, A: addressee

front and listen to only the user's utterances. Therefore, for the dialog systems, the video clips in the dialog corpus should be recorded from the second-person perspective, and the speech of the speakers should be completely separated. One corpora that satisfies these specifications is the Cardiff Conversation Database (CCDb) (Aubrey et al., 2013). CCDb contains rich annotations for social signal processing but does not have information about the speakers' closeness, which is the target of the present study. In the field of emotion recognition, the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset (Busso et al., 2008) or SEMAINE corpus (McKeown et al., 2010) are famous datasets. These corpora have labels of emotion but no appended information related to dialog.

In view of the above, we collected one-to-one conversations between humans to facilitate studies of multimodal dialog systems. The contained data should be as natural as possible, but we gave priority to collecting clear audio-visual dialog data. Thus, we separated the participants into different sound-proof chambers connected by audio-visual communication to record the video clips from the second-person perspective and speech in individual channels. One of the remarkable aspects of our corpus is that it contains information about the closeness between speakers. The closeness label is useful for determining the dialog strategy for constructing a friendly chat-talk dialog system. Table 1 summarizes the differences among the multimodal corpora.

3. Construction of Multimodal Chat-talk Corpus

The purpose of this project is to construct a multimodal dialog corpus that is useful for developing the dialog systems. Thus, we collected clear audio signals and video clips that capture speakers' behavior from the second-person perspective. In addition to information on the closeness between speakers, this corpus contains the transcript and the dialog acts of the utterances. This section describes the dialog collection and annotation process.

3.1. Collection of Multimodal Dialog Data

The recording environment is shown in Figure 1. Two participants entered two individual sound-proof chambers to collect speech sounds without overlap. The interaction of the participants was captured by dynamic microphones (AT4055) placed near the participants' mouth and video cameras (GoPro HERO7) on the monitors. The captured video and sound were presented on the monitor and headphones of the partner in the other chamber in real time.
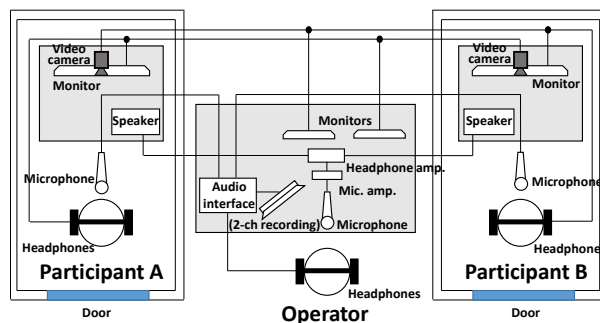

Figure 1: Recording environment

With this setup, there was almost no delay in the video or sound, and the interlocutors could converse naturally.

The purpose of the dialog was to build a relationship, and each participant talked with their partner about five specific topics in Japanese to become more friendly. We consider that chatting about one's preferences and tastes is an appropriate topic for chat-talk with the system, and so we prepared 10 topics based on the "Work (or studies)" and "Tastes and interests" categories of "The self-disclosure questionnaire" (Jourard and Lasakow, 1958). Table 2 shows the prepared topics. When recording a dialog, the operator selected five preferred topics from among these 10 topics.

In this study, we analyzed part of the collected dialog corpus. This initial corpus contained 95 sessions (about 10 hours) of 19 Japanese university students (15 males and 4 females). The participants were recruited randomly, but half of the participants paired with the acquaintances, and the other half of the participants paired with the strangers. Here, one "session" was a dialog about one topic. Figure 2 is an example of the collected dialog data. Each session has image sequences, speech waveforms, and transcripts of the utterances. The transcripts are translations of the Japanese.

3.2. Labels of Closeness between Dialog Participants

Before recording, the interlocutors were asked to answer three questions:

Q1: whether the participants known their partner or not

Q2: how long the participants have known their partner

Q3: how close the participants feel to their partner (i.e., subjective closeness)
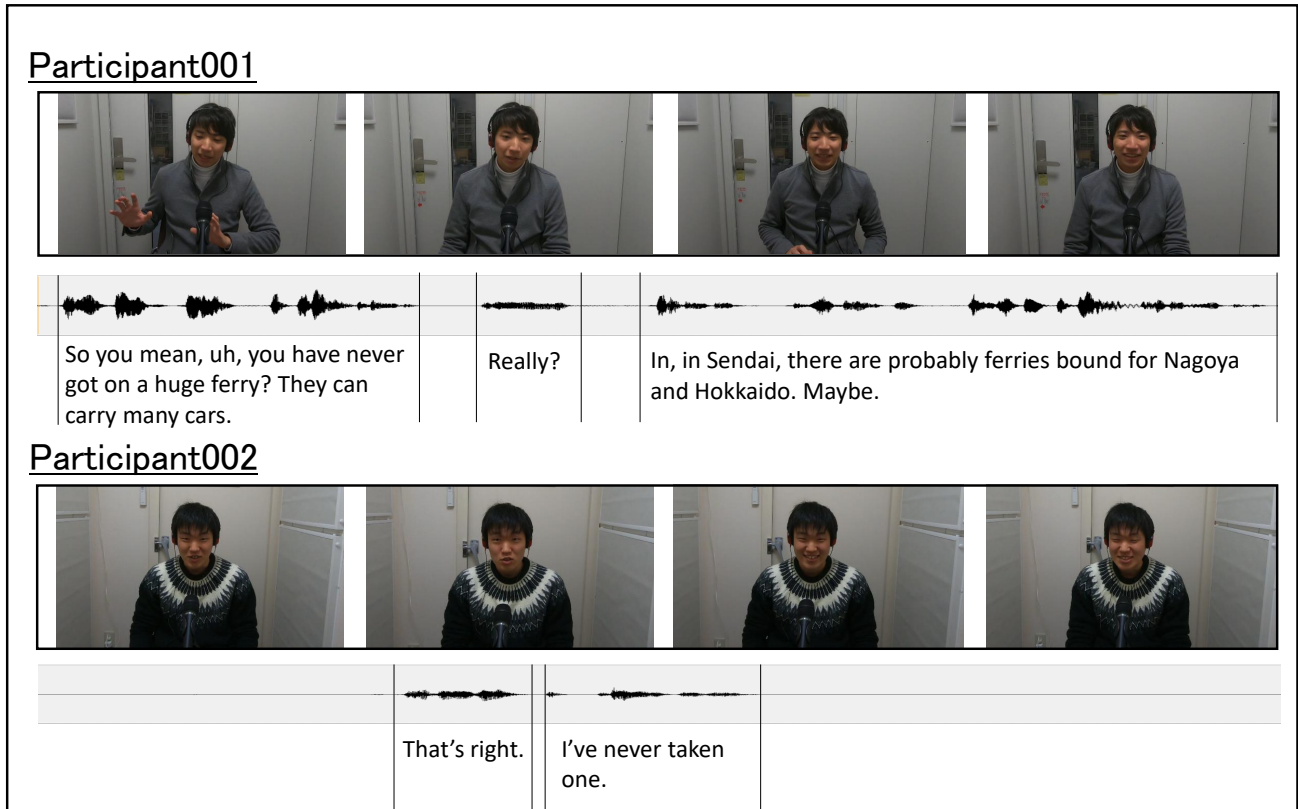
Figure 2: Example of collected dialog data (image sequences, speech waveforms, and transcripts of utterances). The utterances are translated from the Japanese.

Table 2: Prepared topics

No.	Topic
1	My favorite foods and beverages, and the ones I don't like.
2	My favorite music, and the ones I don't like.
3	My favorite reading matter.
4	My favorite movies and animations.
5	The best and the worst places that I have ever been to.
6	My tastes in clothing.
7	My favorite ways of spending spare time.
8	What I enjoy most, and get the most satisfaction from in my present school.
9	How I feel about my friends.
10	My strong and weak points for my work.

Table 3: Concordance rate and Cohen's κ of dialog act annotation (mean $\pm$ standard error).

Concordance rate	0.827±0.004
Cohen's κ	0.790±0.004

For Q3, the participants rated the score on a 5-grade scale, from one (not at all) to five (very much). Q2 and Q3 were asked to only acquainted pairs. Here, the mean and standard error of Q2 were 0.88±0.32 years. The mean and standard error of the scores of Q3 were 4.00±0.161, which reflects the fact that many of the acquainted pairs of the initial corpus had a close relationship.

3.3. Transcript and Dialog Act Annotation

Five crowd-workers transcribed the collected dialog data, and the first author revised orthographical variants and errant punctuation. These transcripts were used for the forced alignments to append the onset and offset times of utterances to the audio signals. Finally, we obtained 19,303 utterances.

Then, two annotators annotated a dialog act tag for each utterance. We prepared 21 dialog act tags based on the tag sets of SWBD-DAMSL (Jurafsky et al., 1997), JAIST annotated corpus (Shirai and Fukuoka, 2018), and listening-oriented dialog corpus (Meguro et al., 2009). The selected 21 tags are shown in Table 4. The annotation criteria were unified by a discussion between annotators. Table 3 shows the mean and the standard error of the inter-rater agreement and Cohen's κ of each dialog. These scores show that the reliability of the annotation is adequate, and the prepared tag set covers the kinds of utterance intention in the collected dialog to some extent.

3.4. Division of Data for Analysis

In this paper, we investigate the conversational phenomena that depends on the level of closeness to determine a dialog strategy to establish a friendly relationship with the user. The utterances of stranger pairs and those of acquaintance pairs are categorized into *low* and *high* levels of closeness, respectively. Table 5 summarizes the dialog data for the analysis. The analysis was conducted session by session.

4. Analysis of Dialog Acts based on Level of Closeness

4.1. Relative Frequency of Dialog Acts

First, we compared the frequency of the dialog acts between different closeness levels. We only used those tags

Table 4: Dialog acts tag set and examples of utterances

Index	Tag (abbreviation)	Description	Example
1	Question (QUES)	Expects a response from partner	Did you watch that movie?
2	Affirmative answer (AFM)	Affirmative answer to partner's question	Yes. / She is.
3	Negative answer (NG)	Negative answer to partner's question	No. / I didn't.
4	Self-disclosure (SD)	Disclosing one's preference, experience, habit, plan, attribute, or inner feelings	I'm into K-pop lately.
5	Opinion (OPI)	Conveying one's opinion of something	I don't think there is anybody who likes it.
6	Information (INFO)	Conveying objective information or hearsay	Well, that is the club for climbing.
7	Commit (CMT)	Suggesting something to partner	You should watch that.
8	Request (RQ)	Requesting something to partner	Stop it!
9	Clarification request (CR)	Confirming content of partner's utterance	Is that so?
10	Confirmation (CFM)	Confirming known information about the partner	You don't like oily and heavy food, right?
11	Complement (CMP)	Complementing one's or their partner's utterance content	Do you know "light novels" as a genre?
12	Repeat-phrase (RPT)	Repeating the phrase included in partner's utterance	France and Italy.
13	Agree (AGR)	Agreeing with partner's information	Exactly.
14	Disagree (DAG)	Disagreeing with partner's information	No, no.
15	Sympathy (SYM)	Sympathizing with partner's opinion	Indeed.
16	Disapproval (DAP)	Opposing partner's opinion	I don't think it suits me.
17	Backchannel or filler (BCF)	Encouraging partners utterance, or filler words	Yeah, yeah.
18	Admiration (AM)	Expressing surprise or admiration for partner	That's amazing!
19	Self-talk (ST)	Talking to oneself	I wonder why.
20	Communication management (COM)	Expressing objective information about the conversation	What were we talking about again?
21	Other (OTR)	Utterance not included in above categories	A sea squirt is (stop speech)

Table 5: Summary of analytical dialog data

Level of closeness	# pairs	# sessions	# utterances
Low	8	40	7218
High	11	55	12085

that agreed between two annotators. The relative frequency of the dialog acts is calculated as follows:

$$P(a_i) = \frac{C(a_i)}{N} \quad (1)$$

where, a_i is the dialog act tag of the i-th utterance, $C(a_i)$ is the frequency of dialog act a_i in the session, and N is the number of utterances in the session.

Table 6 shows the mean and standard error of the relative frequency of each dialog act. The table also shows the results of the unpaired t-test between two closeness levels. We observed significant differences for five dialog acts (QUES, SD, OPI, CFM, and CMP) and marginal differences for three dialog acts (RQ, DAP, and ST). The speakers of the low closeness group made the utterances of questions (QUES) and self-disclosure (SD) more frequently. These results indicate that the speakers tried to get to know each other in the initial stage of a relationship. Conversely, the speakers of the high closeness group did not need to exchange information about the partner's preferences or tastes because they already knew each other well. On the other hand, the speakers of the high closeness group stated their opinions (OPI) and made request utterances (RQ) more frequently. It is considered that the speakers of the high closeness group did not hesitate to convey their intention more directly. In addition, the frequency of the complement (CMP) in the high closeness group was more than the low closeness group. Complementing one's own preceding utterance is a typical example of a casual way of talking and frequently occurred in the acquaintance pairs.

Table 6: Relative frequency of dialog acts of two different closeness levels (mean $\pm$ standard error). The name of the dialog act tag is denoted by its abbreviation (see Table 4).

Dialog act	Level of closeness		p-value
	Low	High	
QUES	0.104±0.006	0.084±0.006	0.016*
AFM	0.034±0.003	0.033±0.003	0.689
NG	0.005±0.001	0.005±0.001	0.916
SD	0.158±0.009	0.118±0.006	<0.001**
OPI	0.109±0.009	0.143±0.007	<0.001**
INFO	0.117±0.010	0.130±0.007	0.318
CMT	0.001±0.001	0.003±0.001	0.117
RQ	0.000±0.000	0.002±0.001	0.054†
CR	0.022±0.002	0.018±0.002	0.223
CFM	0.000±0.000	0.002±0.000	<0.001**
CMP	0.010±0.002	0.017±0.001	<0.001**
RPT	0.014±0.002	0.018±0.002	0.125
AGR	0.022±0.002	0.021±0.002	0.685
DAG	0.000±0.000	0.000±0.000	0.428
SYM	0.040±0.005	0.041±0.003	0.785
DAP	0.001±0.000	0.002±0.001	0.067†
BCF	0.338±0.014	0.344±0.016	0.756
AM	0.010±0.002	0.008±0.001	0.220
ST	0.007±0.002	0.004±0.001	0.092†
COM	0.001±0.001	0.003±0.001	0.175
OTR	0.007±0.001	0.008±0.001	0.530

$^\dagger p < 0.10, ^* p < 0.05, ^{**} p < 0.01$

4.2. Transition Probability of Dialog Acts

Next, we analyzed the transition of the dialog acts. Here, we focus on the utterances at turn-taking. The same as in the previous section, we used those tags that agreed between two annotators. The transition probability of the dialog acts is calculated as follows:

$$P(a_{i+1}|a_i) = \frac{C(a_i, a_{i+1})}{C(a_i)} \quad (2)$$

where, $C(a_i)$ and $C(a_i, a_{i+1})$ are the frequencies of dialog act a_i and the dialog act transition from a_i to a_{i+1} in

Table 7: Frequency of pairs of dialog acts (mean $\pm$ standard error) in which a significant or marginal difference was observed. The name of the dialog act tag is denoted by its abbreviation (see Table 4).

Current→Next	Level of closeness		
	Low	High	p-value
OTR→QUES	0.491±0.109	0.239±0.057	0.049*
NG→QUES	0.194±0.112	0.485±0.122	0.091†
QUES→SD	0.368±0.033	0.275±0.029	0.035*
ST→OPI	0.000±0.000	0.375±0.183	0.080†
DAP→SD	0.000±0.000	0.354±0.156	0.057†
ST→INFO	0.000±0.000	0.312±0.162	0.095†
AFM→INFO	0.198±0.044	0.100±0.035	0.084†
NG→CR	0.250±0.118	0.015±0.015	0.070†
SD→RPT	0.060±0.013	0.097±0.018	0.097†
CMP→SYM	0.014±0.014	0.073±0.028	0.060†
AFM→CR	0.000±0.000	0.022±0.011	0.052†
INFO→CMP	0.003±0.003	0.017±0.007	0.073†
SD→CMP	0.000±0.000	0.009±0.005	0.065†
SD→CFM	0.000±0.000	0.006±0.004	0.095†
SD→DAP	0.000±0.000	0.005±0.003	0.091†

$†p < 0.10, *p < 0.05, **p < 0.01$

the session, respectively. Fillers and backchannels were excluded from the calculation.

From the unpaired t-test for the mean of transition probability, we obtained significant and marginal differences for the 15 items shown in Table 7. As shown in the table, the transition from one speaker's question to another speaker's self-disclosure (QUES→SD) occurs more in the low closeness group than the in high closeness group. This result also suggests that the participants at the first stage of the relationship tend to make typical questionins and answers to get to know each other. In addition, the transition from other to question (OTR→QUES) is another typical transition in the stranger group. OTR was frequently assigned to the interrupted utterance. The speakers in the low closeness group could not take the initiative smoothly and tended to start to speak at the same time when asking a question.

5. Discussion

Our analysis suggested that the collected corpus is useful for obtaining a dialog model that considers the closeness to the user because the conversational behavior appears to depend on closeness level in terms of the dialog acts. In particular, the analysis showed that the speaker in the initial stage of a relationship frequently conducted simple questioning and answering to get to know the dialog partner. On the other hand, the speaker of the acquaintance pairs tended to state an opinion, give an impression, and make a request without hesitation. In this group, confirmation of the known information and making utterance in an informal style like inverse expression also occur.

Although more large-scale comprehensive analysis is required to determine the dialog strategy to establish rapport, a dialog system that considers these results would be useful for building a relationship with the user. For example, the chat-talk system should conduct questioning and answering for a novice user, and increase the frequency of opinions

and requests as the user becomes used to the system.

In future studies, we are planning to evaluate the effectiveness of dialog management based on the closeness with the user by a long-term dialog experiment. In addition, the effect of the closeness on the non-verbal information related to the dialog is also interesting. We are going to analyze the *entrainment* or *mirroring* of the social signals in each stage of the relationship.

6. Conclusion

In this study, we constructed a multimodal Japanese chat-talk corpus considering the closeness between speakers. The collected corpus contains 19,303 utterances (10 hours) from 19 pairs of participants. We analyzed the frequency and transition probability of the dialog acts between the high and low closeness groups to investigate the effect of the closeness between speakers on their behavior. The analysis showed that these factors were different between the closeness levels. In future studies, we will expand the corpus by collecting another 100 participants' chat-talk. We are currently preparing to make the collected corpus publicly available. In addition, we will analyze the non-verbal information between closeness levels and apply the knowledge obtained from the analysis to dialog systems and conduct user evaluations.

7. Acknowledgment

8. Bibliographical References

Altman, I. and Taylor, D. A. (1973). *Social penetration: The development of interpersonal relationships*. New York: Holt, Rinehart & Winston.

Aubrey, A. J., Marshall, D., Rosin, P. L., Vendeventer, J., Cunningham, D. W., and Wallraven, C. (2013). Cardiff Conversation Database (CCDb): A Database of Natural Dyadic Conversations. In *Proc. CVPR*, pages 277–282.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strönbergsson, S., and House, D. (2010). Spontal: A Swedish spontaneous dialogue corpus of audio, video and motion capture. In *Proc. LREC*, pages 2292–2995.

Funakoshi, K. (2018). A Multimodal Multiparty Human-Robot Dialogue Corpus for Real World Interaction. In *Proc. LREC*, pages 35–39.

Jourard, S. M. and Lasakow, P. (1958). Some factors in self-disclosure. *The Journal of Abnormal and Social Psychology*, 56(1):91–98.

Jurafsky, D., Shriberg, E., and Biasca, D. (1997). Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. Technical report, Institute of Cognitive Science.

Kageyama, Y., Chiba, Y., Nose, T., and Ito, A. (2018). Improving user impression in spoken dialog system with gradual speech form control. In *Proc. SIGDIAL*, pages 235–240.

Kim, Y., Kwak, S. S., and Kim, M. (2012). Am I acceptable to you? Effect of a robot's verbal language forms on people's social distance from robots. *Computers in Human Behavior*, 29:1091–1101.

Koiso, H., Den, Y., Iseki, Y., Kashino, W., Kawabata, Y., Nishikawa, K., Tanaka, Y., and Usuda, Y. (2018). Construction of the Corpus of Everyday Japanese Conversation: An interim report. In *Proc. LREC*, pages 4259–4264.

McKeown, G., Valstar, M. F., Cowie, R., and Pantic, M. (2010). The SEMAINE corpus of emotionally coloured character interactions. In *Proc. IEEE International Conference on Multimedia and Expo*, pages 1079–1084.

Meguro, T., Higashinaka, R., Dohsaka, K., Minami, Y., and Isozaki, H. (2009). Analysis of listening-oriented dialogue for building listening agents. In *Proc. SIGDIAL*, pages 124–127.

Oertel, C., Cummins, F., Campbell, N., Edlund, J., and Wagner, P. (2010). D64: A corpus of richly recorded conversational interaction. In *Proc. LREC*, pages 19–28.

Ritter, A., Cherry, C., and Dolan, B. (2010). Unsupervised modeling of twitter conversations. In *Proc. HLT*, pages 172–180.

Serban, I., Sordoni, A., Bengio, Y., Courville, A., and Pineau, J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proc. Thirtieth AAAI Conference on Artificial Intelligence*, pages 3776–3783.

Shirai, K. and Fukuoka, T. (2018). JAIST annotated corpus of free conversation. In *Proc. LREC*, pages 741–748.

Vinciarelli, A., Pantic, M., and Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759.