A Multimodal Corpus of Rapid Dialogue Games

Maike Paetzel, David Nicolas Racca, David DeVault

University of Southern California,
Institute for Creative Technologies,
Playa Vista, CA, USA
8paetzel@informatik.uni-hamburg.de, dracca@computing.dcu.ie, devault@ict.usc.edu

Abstract

This paper presents a multimodal corpus of spoken human-human dialogues collected as participants played a series of Rapid Dialogue Games (RDGs). The corpus consists of a collection of about 11 hours of spoken audio, video, and Microsoft Kinect data taken from 384 game interactions (dialogues). The games used for collecting the corpus required participants to give verbal descriptions of linguistic expressions or visual images and were specifically designed to engage players in a fast-paced conversation under time pressure. As a result, the corpus contains many examples of participants attempting to communicate quickly in specific game situations, and it also includes a variety of spontaneous conversational phenomena such as hesitations, filled pauses, overlapping speech, and low-latency responses. The corpus has been created to facilitate research in incremental speech processing for spoken dialogue systems. Potentially, the corpus could be used in several areas of speech and language research, including speech recognition, natural language understanding, natural language generation, and dialogue management.

Keywords: Multimodal Corpus, Conversational Games, Dialogue Games, Incremental Speech Processing, Spoken Dialogue Systems

1. Introduction

This paper presents the Rapid Dialogue Game (RDG) corpus, a collection of audio and video recordings of human conversations during a series of fast-paced two-person games. The use of games is well attested in natural language processing and dialogue systems research. One recent, high profile example is Watson (Ferrucci et al., 2010), a question-answering system developed by researchers at IBM that competes in the game of *Jeopardy!* against human players. The game was used to focus on research challenges posed by rapidly answering open-domain English language questions in a competitive setting. More broadly, a range of different corpora of participant interactions in games and tasks has facilitated dialogue system research, including work on natural language understanding and generation, turn-taking, dialogue management, and other areas; for a few recent examples see (Fernández et al., 2007; Gravano and Hirschberg, 2009; Koller et al., 2010; Campana et al., 2012).

The corpus presented here is designed to serve as a testbed for future research in incremental speech processing techniques for spoken dialogue systems (SDSs). Traditionally, SDSs have mostly understood and generated utterances with an utterance-level or turn-level granularity, and have relied on a strict turn-taking regime in which they only begin to process and respond to user utterances after the user finishes speaking. This "non-incremental" processing means that systems generally lack many of the highly interactive and low-latency response behaviors that are common in face-to-face human-human conversation. These include providing verbal backchannels (e.g. *yeah, hmm, uh-huh, right, etc.*) and non-verbal backchannels (e.g. head nods) while listening to user speech, using interruptions or overlapping speech in some situations, and having rapid and fluid turn transitions between speakers.

Recent research on incremental speech processing techniques has begun to address this issue by enabling SDSs to incrementally understand, predict, and respond to speech in real-time, as it happens; see e.g. (DeVault et al., 2011; Skantze and Schlangen, 2009; Sagae et al., 2009; Heintze et al., 2010; Skantze and Hjalmarsson, 2010). Incorporating such techniques into SDSs has been shown to be beneficial, including user preference over non-incremental systems, increases in responsiveness (Skantze and Schlangen, 2009; Skantze and Hjalmarsson, 2010), and increased fluency of user speech (Gratch et al., 2006).

The game corpus presented here is designed to support this area of research by enabling a detailed analysis of the real-time speech and dialogue behavior of interlocutors who are trying to communicate in a way that will maximize their game score under substantial time pressure. The two games we have chosen were carefully designed to provide: substantial time-pressure; potential for fun gameplay; naturally occurring overlapping and low-latency responses; and quantifiable game performance.

In the rest of the paper, we present the two RDGs, RDG-Phrase and RDG-Image, that we have designed and then summarize the corpus of human-human data we have collected.

2. Rapid Dialogue Games

2.1. RDG-Phrase

In the RDG-Phrase game, one of the participants takes the role of the clue-giver (giver) and the other the role of the clue-receiver (receiver). The team starts with 2 training rounds, followed by 4 main game rounds. Each round is composed of two parts: a betting phase and a guessing phase. In the betting phase, the giver is shown 10 target expressions as a numbered list. See the top left of Figure 1 for an illustration of the betting phase. Here, target expressions

Start Time	End Time	Target expression	Giver's transcription	Receiver's transcription	Score	Time remaining
1.963	5.215	Manicure	okay so this is what you do when you go and get your nails done		0	68 sec
4.968	5.484			manicure	2	65 sec
5.588	6.97	Pedicure	okay uh			64 sec
7.406	8.278		same thing for feet			62 sec
8.378	8.892			pedicure	4	61 sec
9.724	11.255	Babysitting	um this is when you're			60 sec
11.613	13.502		taking care of someone else's child			58 sec
13.881	14.961			uh babysitting	6	56 sec
14.914	15.17		yeah			55 sec
15.563	18.429	Pet Sitting	same thing but when you're taking care of someone's household animal			54 sec
18.686	19.456			uh dog sitting		51 sec
19.788	20.967			uh cat sitting		50 sec
				pet sitting	8	
21.005	21.312		yeah			47 sec

Figure 1: An excerpt of an RDG-Phrase dialogue with corresponding game status. The target expressions shown in the third column of the table belong to the word category "Care". The screenshots show the user interface visible to the giver during the betting phase (left image) and at the moment the target "Babysitting" was shown during the guessing phase (right image).

are generally nouns or phrases that are chosen to be semantically related to a word category. Semantic associations between word categories and target expressions were created in three steps. First, we used a list of commonly used English nouns as a seed to retrieve semantically related words from WordNet (Miller, 1995). Second, we filtered out results that contained less than 10 expressions related to each category. Third, we manually selected those sets of target expressions which seemed semantically coherent and more suitable for the task. Some examples of word categories used are "care", "cars", or "television" which include target expressions like "maternalism" and "babysitting" for "care", "ambulance" and "bus" for "cars", and "video" and "replay" for "television" respectively. When the numbered list of 10 target expressions is shown, the giver has 120 seconds to redefine an order for these and to place "bets" on which ones the receiver will guess correctly. The bets are optional and modify the score associated with selected target expressions.

In the guessing phase, the giver is privately shown the sequence of 10 target expressions one at a time in the order he has defined. Figure 1 includes an example excerpt from the guessing phase. The goal of the giver is to get the receiver to correctly guess the current active expression (AE) using verbal and non-verbal clues. When the receiver guesses the AE correctly, the team scores points and moves on to the next expression. While in the guessing phase, players can skip the AE to continue with the next target expression from the sequence. Skipped expressions are placed at the end of the sequence and teams have another chance to guess them later. The team is penalized and loses the opportunity to guess the AE if the giver explicitly mentions any part of the word or any part of the expressions that come later in the sequence. The receiver is allowed to guess multiple times without penalty with the only disadvantage of potential time loss. A third person acts as a judge who enforces

that the game rules are obeyed and decides whether a guess is correct or not. During the guessing phase, the team has 70 seconds to guess as many expressions as possible from the sequence before the time expires. This time limit was chosen to make the task challenging but not impossible for the participants.

2.2. RDG-Image

In RDG-Image, one person acts as a giver and the other as a receiver. Figure 2 shows a dialogue excerpt of a game interaction in RDG-Image. Players are presented a set of eight images on separate screens. This set of images is exactly the same for both players except that the images are arranged in a different order on the screen. One of the images is randomly selected as a target image (TI) and it is highlighted on the giver's screen with a thick red border as shown in Figure 2. The goal of the giver is to describe the TI so that the receiver is able to uniquely identify it from the whole set of distractors. Different categories were used for the image sets including images of pets, fruits, people wearing make-up (as shown in Figure 2), and castles, among others. When the receiver believes he has correctly identified the TI, he clicks on the image and communicates this to the giver who has to press a button to continue with the next TI. The team scores one point for each correct guess. There is no direct penalty for a wrong guess besides losing the opportunity to score the point. For RDG-Image, the giver is instructed to only provide clues verbally. Participants are told that non-verbal communication, such as gesturing, is not allowed at any time. Otherwise, there are no forbidden words and players are encouraged to converse freely.[1]

During the data collection, teams were asked to play six different game rounds of RDG-Image. The first 2 were training rounds, followed by 4 main game rounds. In each main game round, participants were given 140 seconds to complete up to 24 target images. This time limit was selected after analysis of pilot testing sessions. As in RDG-Phrase, the time limit was chosen in order to make the task challenging but not impossible for the participants.

2.3. Related Games and Corpora

Our two games have been designed with specific features to support research in incremental speech processing for dialogue systems. These features include substantial time-pressure, naturally occurring overlapping and low-latency responses, and quantifiable game performance. Together, these features highlight the value that rapid, real-time understanding of user speech and incremental response capabilities can provide to spoken dialogue systems. Another design goal for these games has been the potential for fun gameplay, which can make recruitment of participants easier.

RDG-Phrase has some similarity to other spoken word guessing games such as Taboo[2] and The Pyramid Game[3]. It differs in the presence of a betting phase and in its less restrictive approach to prohibited words. A related web-based word guessing game which uses typing rather than speech is Verbosity (von Ahn et al., 2007). We share with Verbosity the motivation to create a fun game that people will enjoy playing and that can serve to create useful corpora for research purposes.

RDG-Image is an extension of an object identification game played by the COREF agent (DeVault, 2008; DeVault and Stone, 2009). In previous versions of this game, participants described simple colored geometric shapes, rather than the more diverse object sets represented in this RDG-Image corpus. Additionally, previous versions used a teletype-based, chat style interaction, rather than speech, and did not include time pressure. The new object sets, together with spoken interaction under time pressure, are designed to create a game that is both more fun and more challenging for human players.

There are many spoken dialogue corpora, including some that share some of the research goals and features that motivate this work. For example, the Fruit Carts domain (Campana et al., 2012; Aist et al., 2006) has been used to explore incremental understanding in a scenario where one participant gives instructions to another while they view a shared visual context. The Pentomino puzzle domain has also supported research on incremental dialogue processing in an instruction following task (Fernández et al., 2007). Research using the Columbia Games Corpus, which again involves a game where objects are identified and moved around on the screen, has looked at dialogue phenomena such as turn-yielding cues (Gravano and Hirschberg, 2009). Generally, the RDG-Image corpus uses a simpler game based only on object identification rather than instruction following, while emphasizing a much wider array of object types and time limited interactions.

3. Set-up and Recordings

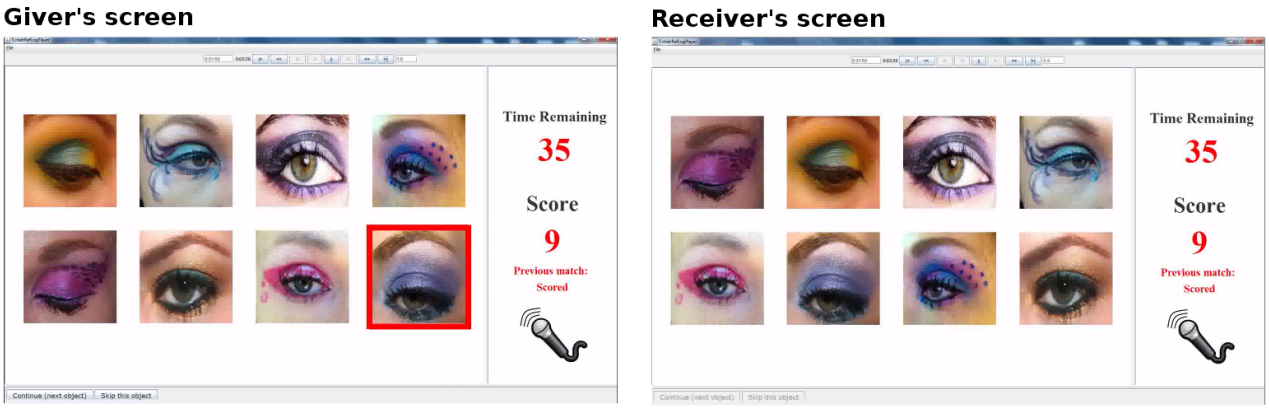
Figure 3 depicts the experiment and hardware set-up used for the data collection. Participants were seated on opposite sides of a table, side A and side B, facing each other. Two displays were used to show the game interface, positioned so that participants could see each other but not their partner's display.

Gestures and movements of each participant were recorded individually with Microsoft Kinect cameras and Logitech webcams. In addition, a webcam installed on the ceiling captured video and audio of both subjects during the game interactions. Figure 4 shows a screen capture taken from this overhead camera during a game interaction. Audio was also recorded for each subject individually using wired Sennheiser microphones and stored in 16 kHz mono WAV files with 16 bit samples. Two additional audio streams were captured with the Microsoft Kinects' internal micro-

[1] In the RDG-Image game, in practice, the two players are primarily looking at their respective computer screens while using a mouse with one hand, so the role of gesture is somewhat reduced in comparison with many face-to-face dialogue contexts. Since we intended to use the RDG-Image corpus to build an automated dialogue agent that will lack gesture recognition and generation capabilities, we instructed participants not to rely on gesture in their object descriptions.

[2] http://en.wikipedia.org/wiki/Taboo_(game)

[3] http://en.wikipedia.org/wiki/Pyramid_(game_show)



Start Time	End Time	Giver's transcription	Receiver's transcription	Score	Time remaining
104.574	106.148	now this is another purple		9	56 sec
107.174	108.252	with a grey eye			53 sec
109.735	112.072	uh greenish eye uh it's like a			51 sec
112.451	113.065		purple		48 sec
112.478	113.317	lavender			
113.904	116.795		is it like blond hair oh that's all blond hair		47 sec
113.985	114.438	purple <pur >			
116.203	118.64	no she has like a brownish eyebrow			44 sec
119.132	119.991		uh		41 sec
119.403	120.572	kinda thick eyebrow			
120.678	121.322		you said it's purple		40 sec
121.49	122.039	yes			39 sec
122.323	123.277		does it have stars		38 sec
123.66	124.877	no stars has nothing			37 sec
124.398	125.495		ok I think I got it		36 sec
126.329	126.637		yeah		34 sec

Figure 2: An excerpt of an RDG-Image dialogue and corresponding screen captures. The image at the bottom right of the Giver's screen is the target.

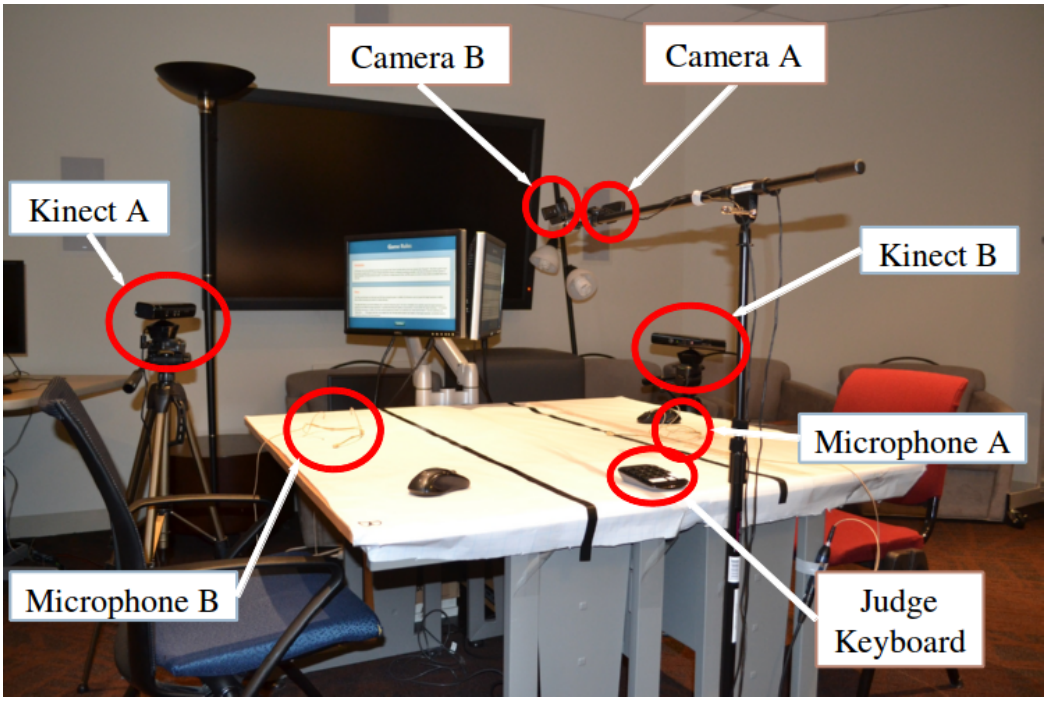


Figure 3: Experiment and hardware set-up for the data collection. In addition to cameras A and B, a third camera was installed on the ceiling to obtain an overhead view of the interactions.

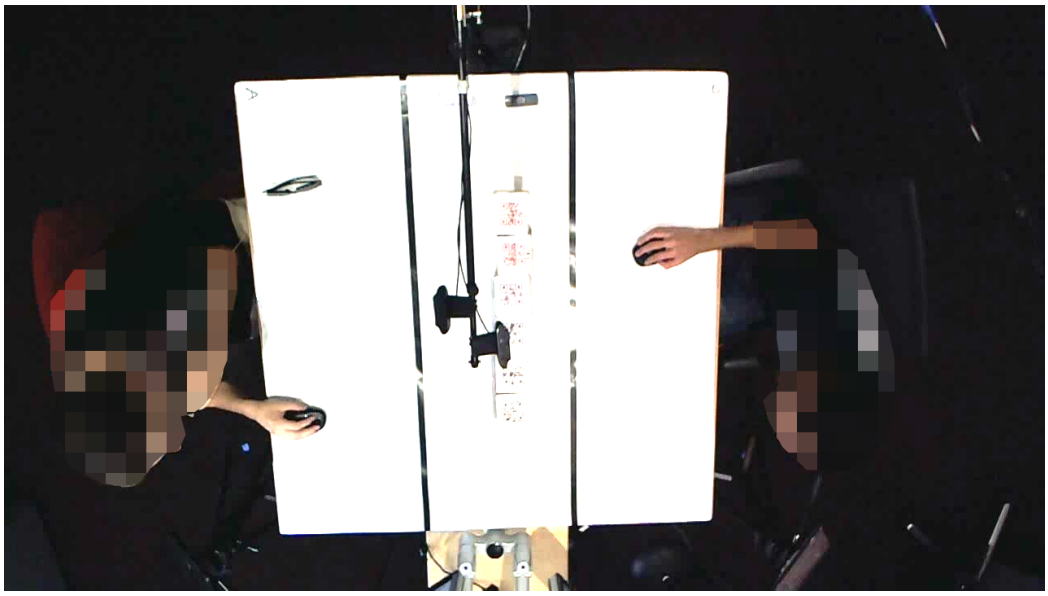


Figure 4: Example of a capture taken from the overhead camera during a game interaction. Participants have been pixelated for privacy reasons.

phones. Video was recorded in 1280x960 and 640x480, at 30fps, and was stored as AVI files.

Besides audio and video recordings, automatic screenshots of both giver and receiver's displays were captured every 200 milliseconds. These were later used to create videos of the game interactions. More importantly, game events and some user interface events, such as mouse-click events, were saved in log files. Game logs include millisecond precision timestamps in order to synchronize audio, video, and screenshots against game events. In this way, the game state throughout each game interaction was logged to permit the analysis of player's language use and performance.

4. Demographics and Post-Questionnaire

64 participants were recruited using Craigslist[4]. To provide demographic data and subjective information about perceived game performance, all participants filled out demographic questions and a post-game questionnaire. From a total of 32 two-people teams, 17 were of mixed gender and 15 were of the same gender. 55% of the 64 subjects were female. When asked about their age, 56 of the total pool of subjects answered and the mean age was 35 years with a standard deviation of 12. Regarding their experience playing similar guessing games, 89% of participants informed that they had never played similar games before or that they had only played a few times. Therefore, only 11% of participants reported to be frequent players of similar games. All subjects were over 18 years old and native English speakers.

5. Transcriptions and Corpus Overview

Each pair participated in 6 game rounds of each of the two games, for a total of 384 rounds. 256 of these were main game rounds, while the rest were training rounds. 17,804 speech segments including 84,615 words have been manually transcribed from the total of 384 dialogues that we have collected. Participant utterances were transcribed using Transcriber (Barras et al., 1998) with start and end timestamps as shown in Figures 1 and 2. Speech segmentation was done by creating a new segment after any unfilled pause of 300ms or longer.

We report here a few summary statistics drawn from a subset of 239 main game rounds (17 of the RDG-Image rounds are excluded due to technical problems during these rounds). The average speaking rate throughout this subset of dialogues (including both participants and both games) is 143.94 words per minute. This measure, called overall speaking rate (Yuan et al., 2006), was obtained by dividing the total number of words found in the dialogues (63,117 words) by the total elapsed time of the dialogues (about 438 minutes). Overlapping speech occurred in approximately 7.82% of the elapsed dialogue time. Filled pauses (such as "mm", "um", and "uh") were included in the transcriptions; they appear 4,479 times, constituting 7.09% of word tokens in the transcripts. Finally, in relation to latency of responses and overlapping speech, the average inter-segment latency, which is defined as the average temporal delay between the conclusion of segment N and the start of segment $N+1$, was found to be 410ms with a standard deviation of 1,800ms (note that negative values could occur due to overlapping segments).

More detailed analysis and annotation of participant speech during the games is a focus of current work.

6. Conclusion

This paper has presented the Rapid Dialogue Game (RDG) corpus, a collection of multimodal human-human interactions created using two fast-paced conversational games. The RDGs were specifically designed to engage participants in a fast paced dialogue with the objective of capturing overlapping speech, low-latency responses and ut-

[4] http://www.craigslist.org

terance interruptions as well as other spontaneous speech phenomena.

In future work, descriptive strategies used by human participants during the games will be analyzed, and automated spoken dialogue systems will be developed that can play the games. Special emphasis will be on opportunities to use incremental speech processing to create more effective and human-like response policies in automated systems.

7. Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1219253. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Thanks to Ramesh Manuvinakurike for help with data analysis.

The images of eye make-up used in Figure 2 are excerpted from pictures protected by copyright and released under Creative Commons licenses by their original authors. In the following attributions, we will identify the 8 images shown in the Giver's screen capture from left-right and top-down direction, with a number from 1 to 8. Thanks to Angela Ceaser for images 4[5], 7[6] and 2[7], all licensed under CC BY-NC-SA 3.0[8], as well as image 5[9], licensed under CC BY-SA 3.0[10]. Thanks to Courtney Rhodes for images 8[11] and 6[12] and to Deborah Austin for image 1[13], all licensed under CC BY 2.0[14]. Image 3[15] by Georgya10[16] is licensed under CC BY 3.0[17].

8. References

Aist, Gregory, Allen, James, Campana, Ellen, Galescu, Lucian, Gallo, Carlos A. Gomez, Stoness, Scott, Swift, Mary, and Tanenhaus, Michael. (2006). Software architectures for incremental understanding of human speech. In *In Proceedings of Interspeech/ICSLP*.

[5]"Cotton Candy Battleship Eye Makeup": http://fav.me/d5hcmm9

[6]"Checkered Valentine Eye Makeup": http://fav.me/d585m29

[7]"Unique Cosmetic Makeup Eye Design": http://fav.me/d54eauv

[8]http://creativecommons.org/licenses/by-nc-sa/3.0/

[9]"Fuschia Leopard Eye Makeup": http://fav.me/d56ed5w

[10]http://creativecommons.org/licenses/by-sa/3.0/

[11]"Blue purple eyeshadow on a green eye": http://www.flickr.com/photos/pumpkincat210/4167122682/

[12]"Copper and blue super macro green eye": http://www.flickr.com/photos/pumpkincat210/4082716621/

[13]"Blue/yellow shadow": http://www.flickr.com/photos/littledebbie11/4434273627/

[14]http://creativecommons.org/licenses/by/2.0/

[15]"Sailor Saturn": http://fav.me/d3dwemd

[16]http://georgya10.deviantart.com/

[17]https://creativecommons.org/licenses/by/3.0/

Barras, Claude, Geoffrois, Edouard, Wu, Zhibiao, and Liberman, Mark. (1998). Transcriber: a free tool for segmenting, labeling and transcribing speech. In *First International Conference on Language Resources and Evaluation (LREC)*, pages 1373–1376.

Campana, Ellen, Allen, James, Swift, Mary, Tanenhaus, Michael K, et al. (2012). Fruit carts: A domain and corpus for research in dialogue systems and psycholinguistics. *Computational Linguistics*, 38(3):469–478.

DeVault, David and Stone, Matthew. (2009). Learning to interpret utterances using dialogue history. In *Proceedings of the 12th Conference of the European Association for Computational Linguistics (EACL)*.

DeVault, David, Sagae, Kenji, and Traum, David. (2011). Incremental interpretation and prediction of utterance meaning for interactive dialogue. *Dialogue & Discourse*, 2(1):143–170.

DeVault, David. (2008). *Contribution tracking: participating in task-oriented dialogue under uncertainty*. Ph.D. thesis, Department of Computer Science, Rutgers, The State University of New Jersey.

Fernández, Raquel, Lucht, Tatjana, and Schlangen, David. (2007). Referring under restricted interactivity conditions. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 136–139.

Ferrucci, David, Brown, Eric, Chu-Carroll, Jennifer, Fan, James, Gondek, David, Kalyanpur, Aditya A., Lally, Adam, Murdock, J. William, Nyberg, Eric, Prager, John, Schlaefer, Nico, and Welty, Chris. (2010). Building Watson: An Overview of the DeepQA Project. *AI Magazine*, 31(3).

Gratch, Jonathan, Okhmatovskaia, Anna, Lamothe, Francois, Marsella, Stacy, Morales, Mathieu, van der Werf, Rick J, and Morency, Louis-Philippe. (2006). Virtual rapport. In *Intelligent virtual agents*, pages 14–27. Springer.

Gravano, Agustín and Hirschberg, Julia. (2009). Turn-yielding cues in task-oriented dialogue. In *Proceedings of the 10th SIGdial Workshop on Discourse and Dialogue*, pages 253–261.

Heintze, Silvan, Baumann, Timo, and Schlangen, David. (2010). Comparing local and sequential models for statistical incremental natural language understanding. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 9–16. Association for Computational Linguistics.

Koller, A., Striegnitz, K., Byron, D., Cassell, J., Dale, R., Moore, J., and Oberlander, J. (2010). The first challenge on generating instructions in virtual environments. In Krahmer, E. and Theune, M., editors, *Empirical Methods in Natural Language Generation*, volume 5790, pages 337–361. Springer.

Miller, George A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Sagae, Kenji, Christian, Gwen, DeVault, David, and Traum, David R. (2009). Towards natural language understanding of partial speech recognition results in dialogue systems. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North Amer-*

ican Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, pages 53–56. Association for Computational Linguistics.

Skantze, Gabriel and Hjalmarsson, Anna. (2010). Towards incremental speech generation in dialogue systems. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–8. Association for Computational Linguistics.

Skantze, Gabriel and Schlangen, David. (2009). Incremental dialogue processing in a micro-domain. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 745–753. Association for Computational Linguistics.

von Ahn, L., Kedia, M., and Blum, M. (2007). Verbosity: A game for collecting common-sense knowledge. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Yuan, Jiahong, Liberman, Mark, and Cieri, Christopher. (2006). Towards an integrated understanding of speaking rate in conversation. In *INTERSPEECH*.