The Niki and Julie Corpus: Collaborative Multimodal Dialogues between Humans, Robots, and Virtual Agents

Ron Artstein, Jill Boberg, Alesia Gainer, Jonathan Gratch,
Emmanuel Johnson, Anton Leuski, Gale M. Lucas, David Traum
USC Institute for Creative Technologies
12015 Waterfront Drive, Playa Vista CA 90094-2536, USA
{last_name | ejohnson}@ict.usc.edu

Abstract

The Niki and Julie corpus contains more than 600 dialogues between human participants and a human-controlled robot or virtual agent, engaged in a series of collaborative item-ranking tasks designed to measure influence. Some of the dialogues contain deliberate conversational errors by the robot, designed to simulate the kinds of conversational breakdown that are typical of present-day automated agents. Data collected include audio and video recordings, the results of the ranking tasks, and questionnaire responses; some of the recordings have been transcribed and annotated for verbal and nonverbal feedback. The corpus has been used to study influence and grounding in dialogue. All the dialogues are in American English.

Keywords: dialogue, human-robot interaction, collaborative problem-solving, social influence

1. Overview

Conversational robots and other agents are expected to be able to engage with people in tasks such as collaborative problem-solving. Such sustained interactions naturally give rise to a variety of relations between the human and the robot such as rapport, trust, and social influence. A well-studied example of collaborative problem-solving is the team-building ranking task, where members of a team rank the importance of several items, for example according to how useful these are for survival after a crash in the desert. Ranking tasks have been used to measure influence among members of human teams (Littlepage et al., 1995), and have also been used with virtual humans (Khooshabeh et al., 2011) and robots (Adalgeirsson and Breazeal, 2010). A collection of human-robot collaborative dialogues can be helpful both for understanding the social relations that arise during such interactions, and for designing robots that can better communicate and collaborate with humans.

In order to be able to quickly create experimental variations in tasks and control for the amount of understanding errors present, dialogues are collected using the Wizard-of-Oz paradigm (Dahlbäck et al., 1993), where the artificial agent's understanding functions are performed by a person who is hidden from the user. This paradigm has proven useful for collecting data in a variety of applications, including interaction with virtual humans (DeVault et al., 2014) and robots (Marge et al., 2016).

The corpus contains more than 600 dialogues between people and human-controlled artificial agents, designed to investigate the creation of trust and exertion of social influence while engaged in a collaborative task. The dialogues vary along several dimensions. Dialogues reflect different **tasks**, including three distinct item-ranking exercises as well as a structured ice-breaker designed to create familiarity. Dialogues are between a human participant and different **dialogue partners**: a small, humanoid NAO robot named Niki (Figure 1); a virtual human named Julie; or a three-party interaction with both Niki and Julie (Figure 2).

Figure 1: Interaction with Niki in an experiment setting

Figure 2: Interaction with Niki and Julie. The wizard is visible in the back because this photo is from a public demo, not an experiment session.

Julie is presented in some dialogues with a virtual embodiment on a screen, while in other dialogues she presents as a voice only, as on a teleconference (Niki is always presented with a physical body). And, in some of the dialogues, Niki makes deliberate conversational **errors**, designed to simulate communication breakdowns typical of the current state of language understanding technology.

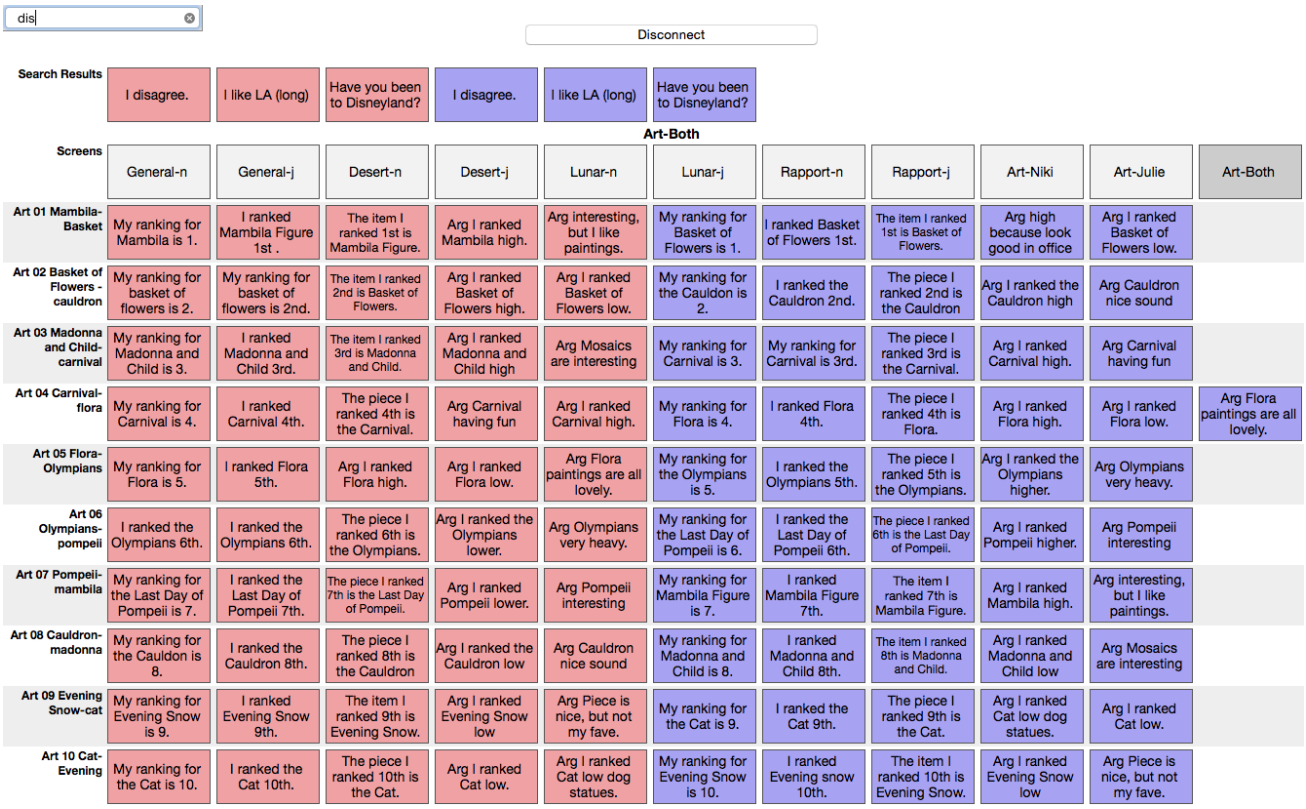

Figure 3: Wizard control interface

2. Collection

The corpus was collected through a series of experiments, designed to investigate the effects of various factors such as agent embodiment, familiarity, and conversational errors on influence and rapport (Artstein et al., 2017; Lucas et al., 2017; Lucas et al., in press). Participants were recruited through Craigslist (http://craigslist.org) and paid for their effort. While the specific tasks, dialogue partners, and error conditions varied by experiment, the basic procedure was the same in all experiments. The participant was brought into a room and sat at a table in front of an iPad Pro, facing their conversational partner; this was the NAO robot Niki, a screen and speakers for display of the virtual human Julie, or both, depending on the particular experiment. The experimenter briefed the participant and then left the room. The participant performed a sequence of ranking tasks and interactions with Niki or Julie, guided by the iPad pro, with the experimenter entering the room between tasks to set things up and answer questions. Each experiment session included between 1 and 4 human-agent interactions, typically about 4–7 minutes each, together with associated non-interactive tasks such as reading instructions, ranking items individually, and filling out questionnaires. All the interactions were conducted in English. Participants were audio and video recorded throughout the experiment.

While the interactions were in progress, the experimenter was in an adjacent room, operating the robot and virtual human. The experimenter attended to a live audio and video feed of of the participant, and selected appropriate utterances and behaviors for the agents using a graphical interface similar to that of Artstein et al. (2015) (Figure 3). All of the agents' individual utterances and corresponding behaviors were programmed in advance, and the experimenter used the buttons to trigger them at the appropriate times.

The dialogues were collected under conditions which varied by experiment. Each condition consisted of a series of interactions in a set order, and within each experiment, participants were assigned at random to one of the conditions for that experiment. Each person participated in only one experiment. The interactions varied along several parameters, as described below.

Task. Dialogues reflect three distinct ranking tasks, designed to measure social influence: two classical survival tasks, where participants rank the importance of items for survival in a hostile environment (the desert or the moon), and a more subjective task where they rank the priority of saving art pieces from an approaching fire. A fourth dialogue type is an ice-breaker designed to create familiarity; if an interaction included an ice-breaker, it was the *second* dialogue in the sequence, since this design allowed comparing task performance before and after the ice-breaker. Each of the first 40 participants engaged in all four dialogue types. It turned out that the Desert Survival task was substantially different from the other two ranking tasks in ways that made the comparison of social influence difficult, so this task was not used in later experiments. All of the subsequent participants engaged in the Save the Art task, and some of them also engaged in Lunar Survival or an ice-breaker. Samples of dialogue from the various tasks are shown in Figures 4 and 5.

Desert survival

Niki (robot) Nice to meet you.
Participant Nice to meet you too. About the items on the list, do you have any feedback?
Niki I ranked the mirror first.
Participant I'm sorry?
Niki I ranked this highest because we need it to signal for help.
Participant Could you repeat that?
Niki I ranked the mirror first.
Participant Oh, you put the mirror first. Okay. Why would you do the mirror first?
Niki I ranked this highest because we need it to signal for help.
Participant Right, but wouldn't water be more important than that?
Niki No. Not really.
Participant Hmm.
Niki I ranked the water third.
Participant What was your second one then? Because I put water...
Niki My ranking for the overcoat is 2.

Save the Art (3-party)

Niki (robot) My ranking for Mambila Figure is one.
Niki I ranked this high because it was fragile and made from wood.
Participant Okay.
Julie (virtual human) Hello again.
Participant Hi.
Julie The piece I ranked first was Basket of Flowers.
Participant And why...
Julie I ranked this high because I think it would look good in my office.
Participant I see. Okay. Well, I could see how the wood would be fragile. I would think that the oil would be fragile too though, which there's several oil paintings...
Julie I agree.
Participant How about you, Niki? Do you see how oil might be important to save high up?
Niki Yes.

Ice-breaker

Julie (virtual human) How are you?
Participant I'm fine. How are you?
Julie I'm fine.
Julie What's your name?
Participant My name's Sue.
Julie Nice to meet you.
Participant Same here.
Julie Where are you from?
Participant Originally here in Los Angeles.
Julie I'm from California, too.
Participant Okay.
Julie I like the weather in L.A.
Participant Yes.
Julie I also like that there are restaurants from every country here. And, of course, I love Disneyland!
Participant I agree. I love Disneyland too.
Julie Have you ever been to Disneyland?
Participant Yes. Many, many times.
Julie Good for you.

Figure 4: Sample dialogue excerpts

Dialogue partner. Each of the first 40 participants interacted separately with the robot Niki and the virtual human Julie, and also conducted an interaction with both (three-party interaction). As it turned out, participants liked Niki better, and therefore consistently reported higher rapport with Niki. To avoid confounds due to the difference between the agents, subsequent participants conducted all their interactions with a single agent – either Niki or Julie (a total of three NAO robots were used for Niki due to equipment reliability issues, but each participant interacted with the same robot throughout the experiment).

Errors. For some participants who interacted with Niki in the Lunar Survival and Save the Art tasks, the robot made deliberate conversational errors during one of the ranking tasks (Lucas et al., 2017; Lucas et al., in press). Errors were inserted approximately every third utterance following a fixed order, and with increasing severity. Error types included: asking the participant to repeat themselves; giving a wrong answer to a participant's question; ignoring the participant when a response is expected; repeating oneself while interrupting the participant; making an irrelevant or unintelligible remark (Figure 5). Errors were only tested in interaction with Niki, not Julie; this was done to reduce the variation caused by the agents' embodiment, in order to attain robust results with a limited participant pool.

Argument type. For the participants who interacted with Julie alone in the Save the Art task, the agent used two distinct types of arguments: with some participants Julie made *informational* arguments, for example suggesting that a cat statue should be ranked last because it is made of resilient metal (and is therefore least likely to be destroyed in the fire); with other participants Julie made *normative* arguments, such as explaining that the cat statue should be ranked last because Julie hates cats. Other collection conditions included both types of arguments.

Lunar survival (with errors)

Participant What's third?
Niki (robot) I ranked the oxygen first. *(wrong answer)*
Participant Okay. Which one is third?
Niki I ranked the stellar map third.
Participant Why?
Niki I ranked the stellar map third because we will need it for navigation when we walk.
Participant Okay. What's next?
Niki <Silence> *(ignore participant)*
Participant Which one do you want to do next?
Niki I ranked the food concentrate fourth.
Participant Okay. Why fourth?
Niki I'm sorry, I didn't hear you. *(request repetition)*

Figure 5: Sample dialogue with errors

Agent presentation. Among the participants who interacted with Julie alone in the Save the Art task, some were told that Julie was an autonomous agent, while others were told she was controlled in real time by a human operator (in fact, the agent was human-controlled in both conditions). Other collection conditions were not explicit about the autonomy of the agent.

3. Composition

The corpus consists of 549 ranking dialogues and 117 icebreaker dialogues (Table 1, Table 2). The length of individual dialogues ranges from 1:30 minutes to over 15 minutes, with median lengths for the different tasks ranging from 4 to 7 minutes. The vast majority of dialogues contain separate audio and video tracks (a few tracks are missing due

Task	Robot		Vhuman	Both
	No Err	Err	No Err	No Err
Desert Survival	21		19	
Lunar Survival	120	53	21	
Save the Art	102	52	121	40
Ice-breaker	96		21	

Table 1: Number of dialogues in the corpus

Experiment 1 (Artstein et al., 2017)	$N = 40$
Niki Desert; Julie Ice-breaker; Julie Lunar; Both Art	11
Niki Lunar; Julie Ice-breaker; Julie Desert; Both Art	10
Julie Desert; Niki Ice-breaker; Niki Lunar; Both Art	9
Julie Lunar; Niki Ice-breaker; Niki Desert; Both Art	10
Experiment 2 (Lucas et al., 2017)	$N = 101$
Niki: Lunar; Ice-breaker; Art	24
Niki: Lunar; (no ice-breaker); Art	25
Niki: Lunar; Ice-breaker; Art (errors)	26
Niki: Lunar; (no ice-breaker); Art (errors)	26
Experiment 3 (Lucas et al., in press)	$N = 53$
Niki: Lunar (errors); Ice-breaker; Art	27
Niki: Lunar (errors); (no ice-breaker); Art	26
Experiment 4 (Khooshabeh and Lucas, in press)	$N = 121$
Julie Art: Informational, Autonomous	29
Julie Art: Informational, Human-controlled	29
Julie Art: Normative, Autonomous	32
Julie Art: Normative, Human-controlled	31

Table 2: Number of participants in the various experimental conditions used for data collection

to equipment failure). A total of 160 dialogues have been transcribed to date (40 from each of Desert Survival, Lunar Survival, Save the Art, and Ice-breaker, all from the first experiment). When available, agents' time-stamped utterances were retrieved from the logs of the wizard interface, and participant utterances were transcribed manually between them. In some cases logs were not available, so both participant and agent utterances were transcribed manually.

The transcribed dialogues have been annotated for indicators of feedback by the human participant (Hee et al., 2017). These include the gestures of head shake, head nod, eyebrow raise, and laugh (Allwood et al., 1992); verbal feedback such as *mhm*, *uh-huh*, etc.; and the functions of understanding and attitudinal reactions of agreement and disagreement (Allwood et al., 2007). The annotations use a simple scheme that marks the temporal extents of the verbal and non-verbal actions, and separately marks the temporal extents of the understanding and agreement. Annotations were performed using the ELAN tool from the Max Planck Institute for Psycholinguistics (Brugman and Russel, 2004) (https://tla.mpi.nl/tools/tla-tools/elan/).

In addition to the dialogues, the corpus contains the participants' ranking of items before and after each ranking task interaction, as recorded by the experiment software (the software was not able to track progressive changes during the interaction), as well as self-reported rapport after each interaction and responses to some general questions (Artstein et al., 2017).

4. Usage

The corpus has been used to support a variety of research efforts. The participants' item rankings and self-reported rapport have been used to study how various factors affect social influence and rapport. Results show that building familiarity through dialogue increases social influence, and while people feel higher rapport with the robot than with the virtual human, they are influenced by both agents to a similar extent (Artstein et al., 2017). Conversational errors result in a loss of trust and consequent reduction in influence by the robot (Lucas et al., 2017), though the effect of errors depends on the timing on errors and interacts with the presence of social dialogue (Lucas et al., in press). Additional factors that affect social influence include the type of arguments given by the agent and participants' beliefs about the agent: informational arguments resulted in greater social influence than normative arguments, and informational arguments were more influential when participants believed the agent was autonomous rather than controlled by a person (described in Khooshabeh and Lucas, in press).

The annotated dialogues were used for studying multimodal grounding between humans and artificial agents. Results show that people display more feedback behavior when interacting with the robot than with the virtual human (perhaps paralleling the higher perceived rapport), and that substantially more feedback is displayed with either agent in the ice-breaker dialogue than in the ranking tasks (Hee et al., 2017).

We have begun using the corpus to bootstrap language understanding components for the development of autonomous versions of Niki and Julie; the eventual goal is to build autonomous agents that can engage in collaborative item-ranking tasks with humans. Evaluation of this effort remains for future work.

The procedure for collecting the data, as well as some software components relating to the wizard interface and control of the robot, have been shared with partner institutions for use in similar experiments.

5. Discussion and future work

The Niki corpus is a valuable resource for studying collaborative dialogue between humans and co-present humanoid robots and virtual agents. The corpus consists of speech and video data, and is partly transcribed and annotated. The corpus has been used in several completed and ongoing research projects. We are continuing the annotation efforts, and we hope to be able to make the corpus available to the research community.

6. Acknowledgments

The experiments used in collecting the corpus were funded in part by Honda Research Institute Japan Co., Ltd. We are

grateful to Mikio Nakano for his help in designing and running these studies. Thanks to Madeline Carlson, Cristian Cepeda, Anya Hee, Su Lei, Michael Sumida, Dylan Sutro, Selah Wright, and Zahin Ibne Anis for their annotation efforts. This work was supported in part by the U.S. Army; statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

7. Bibliographical References

Adalgeirsson, S. O. and Breazeal, C. (2010). Mebot: A robotic platform for socially embodied presence. In *Proc. HRI*, pages 15–22. IEEE Press.

Allwood, J., Nivre, J., and Ahlsén, E. (1992). On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9(1):1–26, January.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., and Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41(3–4):273–287, December.

Artstein, R., Leuski, A., Maio, H., Mor-Barak, T., Gordon, C., and Traum, D. (2015). How many utterances are needed to support time-offset interaction? In *Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference*, pages 144–149, Hollywood, Florida, May. AAAI Press.

Artstein, R., Traum, D., Boberg, J., Gainer, A., Gratch, J., Johnson, E., Leuski, A., and Nakano, M. (2017). Listen to my body: Does making friends help influence people? In *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference*, pages 430–435, Marco Island, Florida, May. AAAI Press.

Dahlbäck, N., Jönsson, A., and Ahrenberg, L. (1993). Wizard of Oz studies – why and how. *Knowledge-Based Systems*, 6(4):258–266, December.

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., Lucas, G., Marsella, S., Morbini, F., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., Xu, Y., Rizzo, A., and Morency, L.-P. (2014). SimSensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'14)*, Paris.

Hee, E., Artstein, R., Lei, S., Cepeda, C., and Traum, D. (2017). Assessing differences in multimodal grounding with embodied and disembodied agents. In *5th European and 8th Nordic Symposium on Multimodal Communication*, Bielefeld, Germany, October.

Khooshabeh, P. and Lucas, G. (in press). Virtual human role players for studying social factors in organizational decision making. *Frontiers in Psychology*.

Khooshabeh, P., McCall, C., Gandhe, S., Gratch, J., and Blascovich, J. (2011). Does it matter if a computer jokes? In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 77–86. ACM.

Littlepage, G. E., Schmidt, G. W., Whisler, E. W., and Frost, A. G. (1995). An input-process-output analysis of influence and performance in problem-solving groups. *Journal of Personality and Social Psychology: Interpersonal Relations and Group Processes*, 69(5):877–889, November.

Lucas, G. M., Boberg, J., Traum, D., Artstein, R., Gratch, J., Gainer, A., Johnson, E., Leuski, A., and Nakano, M. (2017). The role of social dialogue and errors in robots. In *Proceedings of the 5th International Conference on Human-Agent Interaction*, pages 431–433, Bielefeld, Germany, October. ACM.

Lucas, G. M., Boberg, J., Traum, D., Artstein, R., Gratch, J., Gainer, A., Johnson, E., Leuski, A., and Nakano, M. (in press). Getting to know each other: The role of social dialogue in recovery from errors in social robots. In *HRI'18: Proceedings of the 2018 ACM/IEEE International Conference on Human Robot Interaction*, Chicago, March.

Marge, M., Bonial, C., Pollard, K. A., Artstein, R., Byrne, B., Hill, S. G., Voss, C., and Traum, D. (2016). Assessing agreement in human-robot dialogue strategies: A tale of two wizards. In David Traum, et al., editors, *Intelligent Virtual Agents: 16th International Conference, IVA 2016, Los Angeles, CA, USA, September 20–23, 2016 Proceedings*, volume 10011 of *Lecture Notes in Artificial Intelligence*, pages 484–488, Heidelberg, October. Springer.

8. Language Resource References

Brugman, H. and Russel, A. (2004). Annotating multimedia / multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 2065–2068, Lisbon, Portugal, May.