Tackling Interactional Challenges in Social Robots: A Multimodal Conversation Analytic Approach

Hannah R. M. Pelikan
hannah.pelikan@liu.se
Linköping University
Linköping, Sweden

ABSTRACT

The field of social robotics has grown considerably in recent years and social and collaborative robots have entered the consumer market. However, communicative aspects such as timing of utterances and correct interpretation of actions remain a major challenge for social robots. In this position paper I argue that to build collaborative robotic systems that act in socially and interactionally appropriate ways, we need to focus on humans as "the other" in robot-human interaction, whom robotic utterances should be designed for. I present multimodal conversation analysis (CA), a video-based approach that focuses on how actions are interpreted by participants in the context of the ongoing interaction. Identifying three different scales at which CA can be applied, I demonstrate how this approach can support various stages of robot interaction design, making social robots easier to collaborate with from a human perspective.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing design and evaluation methods**; *HCI design and evaluation methods.*

CHI '19 Workshop, May 04, 2019, Glasgow, UK

KEYWORDS

social robots, conversation analysis, human-robot collaboration, robot interaction design

ACM Reference Format:
Hannah R. M. Pelikan. 2019. Tackling Interactional Challenges in Social Robots: A Multimodal Conversation Analytic Approach. In *CHI' 19 Workshop: The Challenges of Working on Social Robots that Collaborate with People.* ACM, New York, NY, USA, 6 pages.

INTRODUCTION

Social interaction between robots and humans can be challenging: Robots start speaking at what feels like the wrong moment, rudely interrupting an ongoing conversation, or take forever to respond. They may say things that seem off, not fitting the conversational context, which may result in funny moments the first time around but quickly becomes dissatisfying when it occurs repeatedly. When engaging in larger activities, robots often fail to reach closure and humans are left with no other option than reformulating until the robot finally does what it should do. While robots have pretty much learned to see, speak and listen, they still struggle with basic social interaction and often fail to communicate with humans in socially appropriate ways. Humans usually adjust to whatever a robot does, but robots still have a long way to go for natural interaction with people, in which they respect interactional rules and collaborate with people in engaging and relevant ways over longer stretches of time.

For social robots to succeed in face-to-face encounters with humans, we need to acknowledge that communication is dialogical [3] and happens between two or more people that are interacting with one another. The qualitative micro-sociological method of conversation analysis (CA) that I am presenting here has shown that humans do not process one sentence at a time but build an understanding in relation to the continuously changing situational context. We also modify our utterances online while producing them, and may design different utterance parts for specific people in our audience [1]. Humans explicitly design their turns for this other, the recipient of their talk, even if this other is a robot [6]. Humans adapt, but so should robots. Interaction is a cooperative and joint endeavor between the interacting parties, so for social and collaborative robots it is not enough to make people learn how they work, but robots should also adjust to the human interaction partner by following basic human interaction principles.

This position paper presents multimodal conversation analysis as a helpful approach for the design and study of robot-human interaction. CA is a data-driven empirical approach that uses video data and transcripts thereof (see Excerpts 1-5) to scrutinize ordinary and mundane interaction. Conversation analysis finds human social interaction to be orderly [9], with people orienting to tacit interactional rules. Scrutinizing how interaction evolves turn-by-turn, action-by-action in natural encounters,

multimodal CA has built a detailed understanding of the interactional principles that humans orient to over the last decades. We can draw on this research in robot interaction design and teach robots basic human interaction rules.

CONVERSATION ANALYSIS AT DIFFERENT SCALES AND DESIGN STAGES

As I demonstrate in the following paragraphs, CA can be applied at different analytic scales: focusing on larger activities, action sequences as well as millisecond timing. Each of these scales may inform different stages of robot interaction design. My arguments are illustrated with transcripts of human interaction from the CA literature and transcripts of robot-human interaction from my own research.

Naturally Occurring Activities and Collaborative Practices

Conversation analysis studies participants' understanding of actions in an interaction, which they display in their own subsequent actions [8]. So rather than asking what someone is intending to do, CA studies how people understand each other at every step in the ongoing interaction. Focus is on the recipient, the one who interprets the current speaker's action, as the subsequent course of interaction very much depends on how the action is taken up by this other. Studying the understanding of actions that participants display to each other in the ongoing interaction, CA lends itself to *in the wild* user studies that look at how people actually do things (rather than what they say they do).

Applied at the scale of larger activities, CA can help to answer questions like *How do people actually interact in a particular context?* that may arise early in the design process. Especially in collaborative settings where robots enter existing teams, CA can yield insights into team practices by unpacking interaction patterns for specific activities, which may then be oriented to in design of a collaborative robot for a particular setting. Similarly, conversation analysis serves as a useful tool when evaluating robots (and other voice agents) that are already deployed in a particular context (e.g. in surgical teams [5], or the home [7]), providing answers to questions such as *How does a robot influence collaborative practices in this particular setting? What kind of practices do people develop in interaction with the robot?* In such contexts, conversation analytic studies may constitute an initial step for a redesign. Studying the interactional patterns during a particular activity, the conversation analytic approach helps to unpack causes of interactional trouble that may remain obscure when e.g. asking users about their experience with a robotic system.

Building Action Sequences

Conversation analysis focuses on the actions that speakers accomplish with their talk. A "Hello" at the beginning of an interaction for instance, is usually understood as a greeting. The same phrase may be used in a different way later in the interaction (e.g. to get the robot's attention, or to display anger/indignation in some languages [11]). How a current action is understood is determined by

Excerpt 1: Greeting and return greeting. Adapted from [10, p. 22].

01		((phone rings))
02	Ava	H'llo:?
03	Bee	(h)Hi:,
04	Ava	Hi:?
05	Bee	(h)Howuh you:?

Excerpt 2: Charade game: Nao and Gary.

01	Nao	+(0.6) hello:
	nao	+waving –>
02		(0.4)
03	Gary	>hi<
04	Nao	(0.5) i'm nao.
05		(0.8)
06	Gary	i'm+ gary
	nao	–>+
((lines 07-09 skipped))		
10	Nao	what's your name?
11	Nao	(0.4) da ↑dup
12	Gary	(0.7) >gary<

Excerpt 3: Cozmo at the dinner table.

01	Cozmo	((idle while scanning D's face))
02	Derrick	let's try something different
	derrick	((lifts beer glass))
03	Derrick	do you like-
04	Cozmo	[derrick]
05	Derrick	[do you like] ((local brand name)) beer?
06	Cozmo	((happy sound))=
07	Derrick	=ah (([laughing]))
08	Cozmo	[°derrick°]
09	Cozmo	((happy sound + movement))
10	Derrick	ah yes ((laughing))
11	Derrick	full approval ((laughing))
((continued jokes about cozmo and beer))		

[1]https://tla.mpi.nl/tools/tla-tools/elan/

previous actions, and the current action shapes and constrains next actions. An interactionally appropriate response to a greeting is a return greeting (see Excerpt 1), so if A says "Hello", B should also say "Hello" or "Hi" (rather than "The sky is blue" or "Please have some coffee").

Studying sequential patterns of interaction, CA helps to address crucial questions regarding dialogue design, such as *What are socially and interactionally appropriate responses to an action? What should the robot say or do in response to a particular human action?* Conversation analysis uses the concept of "conditional relevance" [10] to describe that a current action makes particular next actions relevant. For instance, a question requires an answer and if it is not produced, it is noticably missing (which potentially causes interactional trouble). If a robot produces a greeting, it should leave time for the human interactional partner to also produce a greeting, respecting that humans typically respond to greetings with a return greeting (see Excerpt 2, lines 01-03). Similarly, if a human asks a question, the robot should produce a response to that question (note that humans will hear anything that follows a question as an answer to that question, potentially resulting in misunderstandings such as in Excerpt 3, where the robot produces a "happy sound" (l. 06 and 09) upon finishing with scanning the participant's face, which the participant interprets as positive response to his question).

We can also exploit the fact that humans strongly orient to the relevance of particular next actions by deliberately designing a robot's utterances to narrow the range of possible human responses. As humans will usually produce the relevant next action, we can naturally keep users within the robot's interaction capabilities by producing the corresponding first action. CA thus helps to address questions like *How should we design robot utterances to prompt particular human responses?* For instance in human interaction, the way a mediator formulates their utterances affects how mediation clients will react, with formulations focusing on whether the client would be "willing" to commit to mediation leading to stronger agreement [12]. If a robot introduces itself by saying its name for instance, this may be sufficient to make the user add his or her name as well, making an explicit question for the user's name redundant (see Excerpt 2, l. 06-12). Similarly, a robot's questions may be designed to prompt a response that is easy to process for the robot (e.g. "yes"/"no", rather than "well, actually ..."). While this may be depending on the context that a robot is deployed in, the CA approach provides a systematic way for studying this both in the lab and in the wild and can thereby support design of robotic dialogue engines.

Fine-Tuning Interaction at the Millisecond Level

By collecting video data of interaction and annotating it in software such as ELAN[1], CA allows to study interaction on a millisecond level. This yields crucial insights into the timing of utterances, helping to answer questions such as *How should a robot time its responses with respect to human talk?* Humans try to minimize silence when speaker change occurs and do so by projecting the end of the current speaker's turn [8]. There are cultural differences in how long speakers typically take for

Excerpt 4: Timing of human agreement. Adapted from [2, p. 166]

01	Dianne	Jeff made en asparagus pie
02	Dianne	it was s::so[:goo:d.
03	Clacia	[+I love+ + it. +
	clacia	+nod−+ +nod+

Excerpt 5: Reformulating of a turn in the face of silence

01	Nao	are you ready?
02	Nao	(0.2) da ↑d[up::]
03	Jess	[yeas]:
04	Jess	(0.3) °(h)hh°
05	Jess	(1.4) ye:s, i'm ready
06	Jess	(.) e(h)hh
07	Nao	(0.8) da↓ dap::
08	Nao	(0.7) goo:d.

a speaker change, but in most languages a response produced after 200 ms of silence with respect to a preceding question is perceived as delayed [13], which potentially causes interactional trouble. Positive actions such as agreements even tend to be produced early, in overlap with the end of the current speaker's turn [2, 13] (see Excerpt 4). From the human perspective that pays attention to fine timing, delays in the robot's speech processing are therefore often perceived as indicating interactional trouble and people tend to reformulate their turns when facing long silences (see Excerpt 5).

Social robots can exploit the fine timing of human interaction in several ways. Rather than saying "Sorry, I did not understand that" a robot could employ silence to indicate to participants that repair (e.g. a reformulation of their utterance) is needed. Similarly, if the robot needs more time to retrieve a certain piece of information, or to process information about the environment, it should not stay idle but take the floor by producing some action, such as for example an "um" or another non-lexical utterance to indicate that it is going to respond but needs more time.

Finally, body and voice can be finely coordinated at the millisecond level. *Multimodal* CA focuses on visual behavior that accompanies talk and shows that gestures, gaze and other movements are carefully timed with respect to the ongoing interaction (see e.g. Excerpt 4, l. 03). This can help in addressing questions regarding the fine-tuning of the robot systems, such as *How should the robot's bodily actions and speech be coordinated?* Multimodal CA can provide detailed recommendations regarding the timing of robot utterances and how they should be coordinated with movement and facial expressions. Insights on human gaze patterns for instance have been successfully included in the design of robotic turn-taking behavior [4].

CONCLUSION

The conversation analytic approach scales to studying interaction at the level of milliseconds, action sequences and larger activities. Focusing on how actions are understood and oriented to by the participants in the interaction, CA uncovers the tacit rules that humans attend to when interacting. CA can inform robot-human interaction design by providing concrete answers to questions like ***How*** *do people actually interact with each other and robots?* ***What*** *should a robot say or do next?* ***When*** *should a robot produce a certain action?* Integrating well with other approaches, it is a helpful tool at various design stages, ranging from initial user studies to dialogue design to fine-tuning of motors and speech engine.

ACKNOWLEDGMENTS

This work is funded by the Swedish Vetenskapsråd, project no. 2016-00827. I would like to thank my supervisors Prof. Leelo Keevallik and Prof. Mathias Broth for their helpful comments on this paper.

Transcription conventions.

(0.2)	Timed pause in seconds
(.)	Pause shorter than 0.2 seconds
=	Latching of utterances
-	Cut off
[a]	Overlapping talk
a:	Lengthening of sound
>a<	Utterance noticeably speeded up
°a°	Utterance noticeably softer
a	Stress through pitch and/or amplitude
?	Rising intonation
,	Continuing intonation
.	Falling intonation
↑	Rise in intonation of next syllable
↓	Drop in intonation of next syllable
h	Outbreath
(h)	Hearable aspiration
+action+	Participant's embodied actions
((comment))	Transcriber's descriptions

REFERENCES

[1] Charles Goodwin. 1979. The interactive construction of a sentence in natural conversation. In *Everyday Language: Studies in Ethnomethodology*, George Psathas (Ed.). Irvington Publishers, New York, 97–121.

[2] Charles Goodwin and Marjorie Harness Goodwin. 1992. Assessments and the Construction of Context. In *Rethinking Context: Language as an Interactive Phenomenon*, Alessandro Duranti and Charles Goodwin (Eds.). Cambridge University Press, Cambridge, Chapter 6, 147–190. http://www.sscnet.ucla.edu/anthro/faculty/goodwin/Assessments_and_Construction_of_Context.pdf

[3] Per Linell. 2009. *Rethinking Language, Mind, and World Dialogically*. Information Age Publishing, Charlotte, NC.

[4] Bilge Mutlu, Takayuki Kanda, Jodi Forlizzi, Jessica Hodgins, and Hiroshi Ishiguro. 2012. Conversational Gaze Mechanisms for Humanlike Robots. *ACM Trans. Interact. Intell. Syst.* 1, 2, Article 12 (Jan. 2012), 33 pages. https://doi.org/10.1145/2070719.2070725

[5] Hannah R. M. Pelikan. 2018. "What's going on there?" Negotiating common ground in robotic vs. open surgery : A comparison of surgeon-initiated requests for action in open and robotic surgery. http://essay.utwente.nl/76805/

[6] Hannah R. M. Pelikan and Mathias Broth. 2016. Why That Nao?: How Humans Adapt to a Conventional Humanoid Robot in Taking Turns-at-Talk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 4921–4932. https://doi.org/10.1145/2858036.2858478

[7] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 640, 12 pages. https://doi.org/10.1145/3173574.3174214

[8] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 50, 4 (1974), 696–735. http://www.jstor.org/stable/412243

[9] Emanuel A. Schegloff. 1987. Analyzing Single Episodes of Interaction: An Exercise in Conversation Analysis. *Social Psychology Quarterly* 50, 2 (1987), 101–114. https://doi.org/10.2307/2786745

[10] Emanuel A. Schegloff. 2007. *Sequence organization in interaction: Volume 1: A primer in conversation analysis*. Cambridge University Press, Cambridge.

[11] Margret Selting. 2010. Affectivity in conversational storytelling: An analysis of displays of anger or indignation in complaint stories. *Pragmatics* 20, 2 (2010), 229–277. https://doi.org/10.1075/prag.20.2.06sel

[12] Rein Sikveland and Elizabeth Stokoe. 2016. Dealing with resistance in initial intake and inquiry calls to mediation: The power of "willing". *Conflict Resolution Quarterly* 33, 3 (2016), 235–254. https://doi.org/10.1002/crq.21157

[13] Tanya Stivers, N. J. Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter de Ruiter, Kyung-Eun Yoon, and Stephen C. Levinson. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences* 106, 26 (2009), 10587–10592. https://doi.org/10.1073/pnas.0903616106