Working with Troubles and Failures in Conversation between Humans and Robots: Workshop Report

Frank Förster [1,*]**, Marta Romeo** [2,3]**, Patrick Holthaus** [1]**, Luke Wood** [1]**, Christian Dondrup** [3]**, Joel E. Fischer** [10]**, Farhana Ferdousi Liza** [4]**, Sara Kaszuba** [5]**, Julian Hough** [6]**, Birthe Nesset** [3]**, Daniel Hernández García** [3]**, Dimosthenis Kontogiorgos** [7,8]**, Jennifer Williams** [9]**, Elif Ecem Özkan** [10]**, Pepita Barnard** [11]**, Gustavo Berumen** [12]**, Dominic Price** [11]**, Sue Cobb** [11]**, Martina Wiltschko** [12]**, Lucien Tisserand** [13]**, Martin Porcheron** [11,14]**, Manuel Giuliani** [16]**, Gabriel Skantze** [14]**, Patrick G.T. Healey** [10]**, Ioannis Papaioannou** [15]**, Dimitra Gkatzia** [17]**, Saul Albert** [18]**, Guanyu Huang,** [19]**, Vladislav Maraev** [20]**, Epaminondas Kapetanios,** [1]**.**

[1] *Department of Computer Science, School of Physics, Engineering and Computer Science, University of Hertfordshire, Hatfield, UK*
[2] *Department of Computer Science, The University of Manchester, Manchester, UK*
[3] *School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, UK*
[4] *School of Computing Sciences, University of East Anglia, Norwich, UK*
[5] *Department of Computer, Control and Management Engineering "Antonio Ruberti", Sapienza University of Rome, Rome, Italy*
[6] *School of Mathematics and Computer Science, Swansea University, Swansea, UK*
[7] *Department of Computer Science, Humboldt University of Berlin, Berlin, Germany*
[8] *Science of Intelligence, Research Cluster of Excellence, Berlin, Germany*
[9] *School of Electronics and Computer Science, University of Southampton, Southampton, UK*
[10] *School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK*
[11] *School of Computer Science, University of Nottingham, Nottingham, UK*
[12] *ICREA, Universitat Pompeu Fabra, Barclona, Spain*
[13] *UMR 5191 ICAR, CNRS, Labex ASLAN, ENS de Lyon, Lyon, FR*
[14] *KTH Speech Music and Hearing, Stockholm, SE*
[15] *Alana AI, London, UK*
[16] *University of the West of England Bristol*
[17] *Edinburgh Napier University, UK*
[18] *School of Social Sciences and Humanities, Loughborough University, UK*
[19] *Department of Computer Science, The University of Sheffield, Sheffield, UK*
[20] *Department of Applied IT, Univeristy of Gothenburg, Göteborg, Sweden*

Correspondence*:
Frank Förster
f.foerster@herts.ac.uk

2 ABSTRACT

This paper summarizes the structure and findings from the first *Workshop on Troubles and Failures in Conversations between Humans and Robots*. The workshop was organized to bring together a small, interdisciplinary group of researchers working on miscommunication from two complementary perspectives. One group of technology-oriented researchers was made up of roboticists, Human-Robot Interaction (HRI) researchers and dialogue system experts. The second group involved experts from conversation analysis, cognitive science, and linguistics. Uniting both groups of researchers is the belief that communication failures between humans and machines need to be taken seriously and that a systematic analysis of such failures may open fruitful avenues in research beyond current practices to improve such systems, including both speech-centric and multimodal interfaces. This workshop represents a starting point for this endeavour. The aim of the workshop was threefold: Firstly, to establish an interdisciplinary network of researchers that share a common interest in investigating communicative failures with a particular view towards robotic speech interfaces; secondly, to gain a partial overview of the "failure landscape" as experienced by roboticists and HRI researchers; and thirdly, to determine the potential for creating a robotic benchmark scenario for testing future speech interfaces with respect to the identified failures. The present article summarizes both the "failure landscape" surveyed during the workshop as well as the outcomes of the attempt to define a benchmark scenario.

Keywords: human-robot interaction, speech interfaces, dialogue systems, multi-modal interaction, communicative failure, repair

1 INTRODUCTION

Speech interfaces, user interfaces that allow interaction with technology through spoken commands or queries, are commonplace in many types of robots and robotic applications. Despite the progress in speech recognition and many other areas of natural language processing in recent years, failures of speech interfaces in robotic scenarios are numerous, especially in real-world situations (Porcheron et al., 2018; Fischer et al., 2019). In contrast to the common experience of failure of speech interfaces in robotics, the literature is positively skewed towards the success and good performance of these. While Marge et al. (2022) identified key scientific and engineering advances needed to enable effective spoken language interaction with robotics; little attention was given to communicative failures. To our knowledge, the documentation of failure in speech interfaces and systematic studies of such failures and their causes is exceedingly rare. Honig and Oron-Gilad (2018) provides the most in-depth literature review of prior failure-related HRI studies. The authors found that research in HRI has focused mostly on technical failures, with few studies focusing on human errors, many of which are likely to fall under the umbrella of conversational failures. In addition to this focus on technical errors, the majority of failure-related studies in HRI take place in controlled experimental

conditions, where 'failures' are explicitly designed and occur only at specific moments (Ragni et al., 2016; Washburn et al., 2020a; Cuadra et al., 2021; Green et al., 2022), instead of a natural occurrence of the interactions between humans and robots. Closer to the topic of the workshop is the recently proposed taxonomy of Tian and Oviatt (2021) that focuses on social errors in HRI and their relationship with the perceived socio-affective competence of a robot. However, while there is significant overlap between social errors, as categorized by Tian and Oviatt, and the workshop topic of conversational failure, the perspective on the role of these errors and failures in interaction as well as the view as to whether these could be overcome eventually differs significantly. While social errors should ultimately be reduced by increasing a robot's perceived socio-affective competence, it appears unlikely that conversational failure could be totally extinguished by means of technological progress. Too frequent is their occurrence in human-human conversation and too deeply ingrained are the related repair mechanisms in the fabric of human communication.

To the best of our knowledge, there are currently no survey papers specifically on conversational failures in human-robot interaction, a fact that illustrates an important gap in the research landscape. To address this gap, we conducted a two-phase workshop with experts in adjacent fields. This paper presents the findings from this workshop series that brought together a multidisciplinary group of researchers from fields such as robotics, human-robot interaction (HRI), natural language processing (NLP), conversation analysis, linguistics and pragmatics. The workshop provided a platform to discuss the multitude of failures of speech interfaces openly and to point out fruitful directions for overcoming these failures systematically. The workshop focused mainly on human-robot joint action scenarios involving multimodal coordination between humans and robots, as these are the norm in scenarios where robotic speech interfaces are deployed. The identified types of failures range from failures of speech recognition to pragmatic failures and infelicities.

We begin by describing the aims, structure, and materials used in the workshop in Sect. 2. We then present findings that result from the workshop, including participant contributions and outcomes of the structured discussion in Sect. 3. This leads to Sect. 4, where we reflect on problems and identify themes that emerged from the workshop's discussions before concluding the paper.

2 MATERIALS AND METHODS

The *Working with Troubles and Failures (WTF) in Conversations between Humans and Robots* workshop included a virtual gathering over two consecutive days in June 2022 and an in-person full-day meeting at the University of Hertfordshire in September 2022. Here, we sketch the structure and summarize the findings for each of these parts.

2.1 Before the Workshop

In order to attract workshop participants interested in an open discussion of their experience and investigations of failing speech interfaces, we directly contacted some of the potentially interested research groups within the United Kingdom. Additionally, the workshop was advertised via mailing lists relevant to the HRI (e.g. *hri-announcement*, *robotics-worldwide*, *euRobotics-dist*), natural language processing (NLP, e.g. *ACM sigsem*), and artificial intelligence communities (e.g. ACM *sigai-announce*). To verify participants' genuine interest in the topic and to collate information on the different types of conversational failures experienced by them, they were asked to submit the following pieces of information:

1. the number of years of experience using or developing speech interfaces,
2. an indication of what they perceive to be the most pressing issue or the biggest source of failure for speech interfaces,
3. their most memorable WTF moment, that is, which of their experiences of failure with a speech interface they remembered most vividly,
4. a summary of their motivation to attend the workshop,
5. a suggestion for a future benchmark scenario that would expose the kind of failure described in their WTF moment.

Applicants that stated a meaningful entry for item 4, and made some attempt to answer the other questions, were admitted to the workshop. As a result, 15 participants were admitted and initially attended the virtual part. Of these fifteen participants, eight would go on to attend the face-to-face part of the workshop. The face-to-face workshop was re-advertised via the above-mentioned mailing lists and the same set of questions and answers was used to filter out additional prospective participants. Ultimately, six new participants joined the face-to-face part of the workshop, resulting in fourteen non-speaker, non-organiser participants. Two of these attended the face-to-face workshop virtually, as we decided to go for a hybrid format in order not to exclude anyone who was not able or willing to travel on site.

Keynote speakers for both parts of the workshop were chosen based on their expertise in the subject area. The subject areas considered most relevant to the workshop were robotics-centred NLP on the one hand and Conversation Analysis (CA) on the other. The emphasis on CA was based on the fact that the documentation and analysis of conversational failure have been an integral part of this discipline since its very inception. Moreover, it was hoped that having keynote speakers and participants from both areas would soften discipline-specific boundaries and limitations and potentially open up new directions for future research.

101 2.1.1 Motivations for Attending the Workshop

102 The following is a summary of the participants' motivation for attending the workshop as extracted
103 from the application forms:
104 Several PhD students were hoping to connect and network with other researchers working in speech
105 interaction technologies. Multiple other researchers working on the CA-HRI interface wanted to
106 learn more about how conversational trouble emerges, while others occupied with developing speech
107 interfaces, or with integrating these into robots were interested in gaining a deeper understanding of
108 current issues. Many of them were also interested in sharing their experiences with peers.
109 One researcher working in animal communication hoped to learn something from a different domain
110 of "inter-being communication", while yet another researcher working on speech privacy wanted
111 to connect to other researchers working on speech interfaces. One participant saw value in the aim
112 of identifying or creating a benchmark scenario that would be able to tease out the most common
113 failures, if they occurred - an aim explicitly set out by the workshop.
114 Another motivation of multiple participants to attend the workshop was their shared belief that a
115 deeper analysis of communicative failures would not only help to improve future speech interfaces
116 but also gain a deeper understanding of (human) conversations themselves.
117 Finally, a researcher interested in explainable AI was interested to see what other types of failures,
118 apart from faulty explanations, there are and how these may connect to research in explainable AI.

119 **2.2 Virtual Workshop**

120 To facilitate participation in the virtual session of the workshop, it was divided into two half-day
121 events. On the first day, the workshop opened with a keynote talk by Prof. Patrick Healey, Professor
122 of Human Interaction and Head of the Cognitive Science Research Group in the School of Electronic
123 Engineering and Computer Science at Queen Mary University of London, on "Running repairs:
124 Coordinating meaning in dialogue" (Section 3.1.1). This was followed by participants' lightning
125 talks on their most memorable WTF moments when working with communication between humans
126 and robots (Section 3.2). Following the lightning talks, and based on the underlying themes identified
127 by the organisers, participants were divided between 4 breakout rooms to continue discussing the
128 issues they brought to the workshop. The four identified themes were: (i) Context Understanding,
129 (ii) Handling Miscommunication, (iii) Interaction Problems, and (iv) General Failures.

130 The second day of the virtual workshop saw Dr. Saul Albert, Lecturer in Social Science (Social
131 Psychology) in Communication and Media at Loughborough University, give a keynote talk on
132 "Repair, recruitment, and (virtual) agency in a smart homecare setting" (Section 3.1.2). Following
133 the talk, each group from the breakout rooms of the first day reported what was discussed and each
134 debate was opened to all participants. The workshop ended with a short summary of the day.

2.3 Face-to-Face Workshop

The in-person part of the workshop was held at the University of Hertfordshire three months after the virtual event. During this full-day meeting, keynote talks were given by Prof. Gabriel Skantze, Professor in Speech Technology at KTH Royal Institute of Technology on "Building Common Ground in Human-Robot Interaction" (Section 3.1.3) and by Dr. Ioannis Papaioannou, Chief Technology Officer & Co-Founder of Alana [1] on "Tackling the Challenges of Open-Domain Conversational AI Systems" (Section 3.1.4).

Since the registration to the face-to-face workshop was also opened to participants who did not take part in the virtual workshop, new attendees were given the opportunity to present their own lightning talks on their WTF moments (Section 3.2).

A central part of the face-to-face workshop was the World Café session[2], which provided participants an opportunity to freely discuss troubles and failures in small groups across several table topics. Based on the participants' submitted WTF moments, and the themes from the breakout rooms of the virtual part, four themes were chosen for this session: (i) Context Understanding, (ii) Interaction Problems, (iii) Handling Miscommunication, and (iv) Suggested Benchmark Scenarios. Each theme was allocated to one table, and each table had one designated organizer. Participants and speakers were split into four different groups and moved between the tables within time slots of approximately 15 minutes per theme. The tasks of a table's organizer were to summarize the findings and discussions from previous groups to a newly arriving group, to encourage discussions around the table topic, and to either encourage note taking or take notes themselves on a large flip chart that was allocated to each table.

3 RESULTS

In this section, we present findings from both the virtual and the face-to-face parts of the workshop, describing how the keynotes shaped the discussion and how the participant lightning talks contributed to identify some of the most pressing problems in conversations between humans and robots. Most importantly, we will present the outcomes of the structured discussion, summarising the workshop findings.

3.1 Keynotes

To frame the discussion on troubles and failures with experiences from different perspectives, we invited four keynote speakers from scientific areas that are concerned with research problems around

---

[1] https://alanaai.com/

[2] https://theworldcafe.com/key-concepts-resources/world-cafe-method/

164 conversations between humans and robots. This section summarises their presentations in the context
165 of the workshop goals to scope and identify common troubles and failures in conversation between
166 humans and robots. In the virtual part of the workshop, the first keynote (Sect. 3.1.1) provided a
167 conversation analytical perspective on repairs and meaning in dialogue, while the second one looked
168 at repairs but from a more applied perspective in a user's home (Sect. 3.1.2). The in-person workshop
169 provided insights considering human-robot interactions (Sect. 3.1.3) and an industry viewpoint
170 (Sect. 3.1.4).

171 ### 3.1.1 Running Repairs: Coordinating Meaning in Dialogue

172   Healey presented the Running Repairs Hypothesis (Healey et al., 2018b), which captures the idea
173 that successful communication depends on being able to detect and adjust to misunderstandings on
174 the fly. The basic assumption is that no two people ever understand exactly the same thing by the
175 same word or gesture and, as a result, misunderstandings are ubiquitous. Data from conversations
176 support this assumption. For example, the utterance "huh?" occurs around once every 84 seconds in
177 conversation and appears to be universal across human languages (Enfield, 2017; Dingemanse et al.,
178 2015). Around a third of turns in ordinary conversation involve some sort of real-time adjustments
179 in language use (Colman and Healey, 2011).

180   The processes for detecting and resolving problems with understanding have conventionally been
181 regarded as 'noise in the signal' by the cognitive sciences (Healey et al., 2018a). However, there
182 is evidence that they are fundamental to our ability to adapt, in real-time, to new people, new
183 situations and new tasks. Conversation analysts have described a set of systematic turn-based *repair*
184 processes that structure how people identify and respond to misunderstandings (Schegloff et al.,
185 1977a; Schegloff, 1992a, 1997). Experimental evidence shows these repair processes have a critical
186 role in building up shared understanding and shared languages on the fly (Healey et al., 2018b;
187 Healey, 2008, 1997).

188   The Running Repairs Hypothesis characterises human communication as a fundamentally error-
189 prone, effortful, active, collaborative process but also highlights how these processes are structured
190 and how they make human communication flexible and adaptable to new people and new situations.
191 This can liberate human-robot interaction from the fantasy of perfect competence (Park et al., 2021).
192 Instead, robots could, in principle, take advantage of the resources of interaction by engaging in
193 repairs. This requires developing the ability to recognise critical verbal and non-verbal signals of
194 misunderstanding and the use of incremental online learning processes that build on the sequential
195 structure of interaction to make real-time revisions to language models (see e.g. Howes and Eshghi
196 2021; Purver et al. 2011).

3.1.2 Repair, Recruitment, and (virtual) Agency in a Smart Homecare Setting

Albert argued that moments of trouble and failure can provide researchers with ideal empirical material for observing the structure of the participation frameworks we use to get things done in everyday life (Goodwin, 2007; Albert and Ruiter, 2018). His presentation used multimodal video analysis to show how a disabled man and his (human) carer leveraged troubles and failures in their interactions with an Amazon Echo with voice-controlled lights, plugs, and other devices to co-design an effective smart homecare participation framework.

Instances in this case study highlighted how the human carer used troubles and failures to prioritise the independent role and agency of the disabled person within a joint activity. For example, the carer would stop and wait for the disabled person to resolve the trouble in their interactions with the virtual agent and complete their task even when it would have been faster for the carer to complete the disabled person's task manually. In other examples, trouble in the interactions between the carer and the virtual assistant provided an opportunity for the disabled person to intervene and assist the carer by correcting and completing their vocal instruction to the device. The disabled person was also able to tacitly 'recruit' (Kendrick and Drew, 2016) assistance from the human carer by repeatedly re-doing failed commands to the virtual assistant within earshot of the carer, soliciting support without having to ask for help directly.

These episodes show how people can harness trouble and failures in interaction with a virtual assistant to enable subtle shifts of agency and task-ownership between human participants. This kind of hybrid smart homecare setting can support and extend the independence of a disabled person within an interdependent, collaborative participation framework (Bennett et al., 2018). More broadly, the communicative utility of trouble and failure in interactions with machines highlights the shortcomings of our idealized–often ableist–models of the 'standard' user, and medicalized models of assistive technology (Goodwin, 2004; Albert and Hamann, 2021).

3.1.3 Building Common Ground in Human-robot Interaction

Skantze highlighted two aspects of miscommunication and error handling in human-machine interaction. First, he discussed how language is ultimately used as part of a joint activity. For communication to be meaningful and successful, the interlocutors need to have a mutual understanding of this activity, and of their common ground (Clark, 1996). From this perspective, language processing is not a bottom-up process, where we first figure out what is being said before interpreting and putting it in context. Rather, we use the joint activity to steer the interpretation process and possibly ignore irrelevant signals. Skantze exemplified this with an early experiment, where a noisy channel (including a speech recognizer) was used in a human-human communication task, where one person had to guide another person on a virtual campus (Skantze, 2005). Although much of what was said did not get through (due to the error prone speech recognition), the humans

232 very seldom said things like "sorry, I didn't understand", which are frequent responses in human-
233 machine interactions. Instead, they relied on the joint activity to ask task-related questions that
234 contributed to task progression. Another implication of this view on communication is that the
235 idea of "open-domain dialogue", where there is no clear joint activity, is not meaningful to pursue
236 (Skantze and Doğruöz, 2023).

237   The second aspect that was discussed was the need to incorporate user feedback when the system
238 is speaking, and use that feedback to model what can be regarded as common ground between the
239 user and the system. Skantze exemplified this issue with a research project at KTH (Axelsson and
240 Skantze, 2023), where an adaptive robot presenter is being developed (in the current demonstrator
241 it is talking about classic works of art in front of a human listener). The robot presenter uses a
242 knowledge graph to model the knowledge it is about to present, and then uses that same graph to
243 keep track of the "grounding status" of the different pieces of information (Axelsson and Skantze,
244 2020). Multimodal feedback from the user (e.g., gaze, facial expressions, nods and backchannels)
245 are interpreted as negative or positive, and the graph is updated accordingly, so that the presentation
246 can be adapted to the user's level of knowledge and understanding (Axelsson and Skantze, 2022).

3.1.4 Addressing the Challenges of Open-Domain Conversational AI Systems

248   Papaioannou's presentation showed how designing conversational AI systems able to engage in
249 open-domain conversation is extremely challenging and a frontier of current research. Such systems
250 are required to have extensive awareness of the dialogue context and world knowledge, the user
251 intents and interests, requiring more complicated language understanding, dialogue management,
252 and state and topic tracking mechanisms compared to traditional task-oriented dialogue systems.

253   In particular, some of these challenges include: (a) keeping the user engaged and interested over
254 long conversations; (b) interpretation and generation of complex context-dependency phenomena
255 such as ellipsis and anaphora; (c) mid-utterance disfluencies, false starts, and self-corrections
256 which are ever-present in spoken conversation (Schegloff et al., 1977b; Shriberg, 1994) (d) various
257 miscommunication and repair phenomena such as Clarification Requests (Purver, 2004) and Third
258 Position Repair (Schegloff, 1992b) whereby either the user or system does not understand the other
259 sufficiently or misunderstands, and later repairs the misunderstanding. (b-d) are all crucial to robust
260 Natural Language Understanding in dialogue.

261   A modular conversational AI system, (called *Alana*), tackling some of the aforementioned
262 challenges (i.e. user engagement over long conversations, ellipsis and anaphora resolution, and
263 clarification requests) was developed between 2017-2019 (Papaioannou et al., 2017; Curry et al.,
264 2018) and deployed to thousands of users in the United States as part of the Amazon Alexa Challenge
265 (Ram et al., 2018). The Alana system was also evaluated in a multimodal environment and was used
266 as the overall user conversational interaction module in a multi-task and social entertainment robotic

267 system as part of the MuMMER project (Foster et al., 2019). The integrated system was deployed in
268 a shopping mall in Finland and was able to help the user with specific tasks around the mall (e.g.
269 finding a particular shop or where they could buy a certain product, finding the nearest accessible
270 toilet, or asking general questions about the mall) while at the same time engaging in social dialogue
271 and being entertaining.

272 The output of that research was fed to the implementation of the 'Conversational NLU' pipeline by
273 Alana AI, a modular neuro-symbolic approach further enhancing the language understanding of the
274 system. The Conversational NLU module is able to detect and tag a number of linguistic phenomena
275 (e.g. disfluencies, end-of-turn, anaphora, ellipsis, pronoun resolution, etc) as well as detect and
276 repair misunderstandings or lack of sufficient understanding, such as self-repairs, third-position
277 corrections, and clarifications. The system is currently being evaluated by blind and partially sighted
278 testers in the context of multi-modal dialogue allowing the users to find mislocated objects in their
279 environment via a mobile application.

280 ## 3.2 Lightning Talks

281 The following section contains short summaries of the lightning talks of both the virtual and the
282 face-to-face part of the workshop. From the presentations, three themes were identified: *Description*
283 *and Analysis of Failures and Troubles* (Sect. 3.2.1) grouping presentations that have a descriptive
284 or analytical focus; *Technical Aspects of Conversational Failure* (Sect. 3.2.2) for presentations
285 that have a more technical focus; and *Adjacent Topics in Speech Interfaces* (Sect. 3.2.3), grouping
286 presentations on topics that, while not focusing strictly on conversational failures, covering other
287 forms of errors and issues that fall into the wider topic of speech-centric human-machine interactions.
288 Note that many of the talks falling into the second, technical category still contain a substantial
289 element of analysis that enabled or inspired the technical solutions described therein.

290 ### 3.2.1 Description and Analysis of Failures and Troubles

291 The following ten of the contributions took a more analytical approach to the failure they reported
292 in their lightning talks. They describe possible reasons or implications of the failure they present.

293 #### 3.2.1.1 *Laundrobot: Learning from Human-Human Collaboration*

294 Barnard and Berumen presented their work on *Laundrobot*, a human acting as a collaborative robot
295 designed to assist people in sorting clothing into baskets. The study focused on participants' ability
296 to collaborate through verbal instructions and body movements with a robot that was sometimes
297 erroneous when completing the task. The team analysed social signals, including speech and gestures,
298 and presented three cases demonstrating human-human collaboration when things do not go as
299 expected. In one of the cases, a participant gave clear instructions to an erroneous Laundrobot, which

led to frustration on the participant's part, with statements such as "Okay, I'm doing this wrong". The presenters described how the participant appeared to take responsibility for the errors made by the robot. They examined the use of language and expression of intent in different instances for pieces of clothing that were either correctly or incorrectly identified by Laundrobot. During this analysis, Barnard, Berumen, and colleagues came across an interesting case regarding the use of the word "right", which was frequently used in both erroneous and non-erroneous instances. The group explored how that word had different meanings depending on the success or failure of Laundrobot. For instance, for one participant (P119), the word had a single meaning of indicating a direction in erroneous instances, whereas, on other occasions, it had alternative purposes. It was sometimes used to refer to directions and, at other times, used for confirmation, immediacy ("right in front of you"), or purpose ("Right, OK").

3.2.1.2   Sequential Structure as a Matter of Design and Analysis of Trouble

As part of the *Peppermint project*[3] corpus, Tisserand presented a transcript fragment, reproduced below. They designed a Pepper robot as an autonomous reception desk agent that would answer basic requests asked by library users. They captured *naturally-occurring interactions*: the robot was placed in the library, and users were free to interact and leave whenever they wanted.

```
01 Hum: where can I find books of maths?           | Sequence A – Part 1
02 Rob: ((provides the direction for books of maths)) | Sequence A – Part 2
03 Rob: is it clear to you?                         | Sequence B – Part 1
04 Hum: yes thanks                                  | Seq B-2 && Seq A-3
05 Rob: okay, I will repeat ((repeats turn line 2)) | Sequence C – Part 1
```

The failure here is the fact that the robot recognized "no thanks" instead of two separate actions: "yes" + "thanks" (l.4); the robot thus repeats the answer to the user's question. Reflecting on this WTF moment, Tisserand highlighted how this failure occurred due to decisions made during the scenario design phase. Firstly, poor speech recognition differentiation between the words "yes" and "no" had led the scenario design team to add "no thanks" to a word list provided for recognising an *offer rejection*:(a *dispreferred turn design* for this type of action (Schegloff, 2007, Chap.5)) in another scenario in which the robot makes an offer. Secondly, because the state machine was based on isolated so-called "contexts", it was designed only to make one decision when processing a spate of talk. Here, therefore, the clarification check turn in line 3 was treated as independent from the question response in line 2. Because the speech recognition system struggled to differentiate "yes" and "no", and was using the word list that labelled "no thanks" as a case of *offer rejection*, here it erroneously recognized "yes thanks" in line 4 as a negation (a *clarification denial*), and proceeded to repeat the turn.

---

[3] `https://peppermint.projet.liris.cnrs.fr/`

---

334   What should have happened is that when the robot asks the user to confirm (l.3), it should recognize
335   that this sequence is embedded in the previous question/answer sequence (l.1-2). In this case, the
336   human's "yes" (l.3) is a response to the just-prior confirmation request while the "thanks" responds
337   (in the first structurally provided sequential slot) to the Robot's answer as a 'sequence closing third'
338   (l.3). This is why the team is now *sequentially* annotating training datasets to show what utterances
339   correspond not only to questions and answers, but also the cement in-between: how the user might
340   delay, suspend, abandon, renew or insert actions (e.g. repair). Here interaction is seen as a temporally
341   continuous and incremental process and not a purely logical and serial one. In other words, context
342   is seen as an organized resource more than an adaptability constraint.

343   *3.2.1.3 Design a Robot's Spoken Behaviours Based on How Interaction Works*

344   Huang pointed out that spoken interaction is complicated. It is grounded in the social need to
345   cooperate (Tomasello, 2009; Holtgraves, 2013) and requires interlocutors to coordinate and build
346   up common ground on a moment-by-moment basis (Krauss and Fussell, 1990, p.112)(Holtgraves,
347   2013).

348   Speech is only one tool in a larger picture. Some errors are caused by failures in natural language
349   understanding (NLU) as illustrated in the following sequence:

```
350   01 User:  Let's talk about me.
351   02 Robot: What do you want to know about 'me'?
```

352   Other issues, however, could be caused by a lack of understanding of common ground. For example,
353   when a naive user asked, "Where to find my Mr Right", the system provided a place named "Mr
354   & Mrs Right" and told the user it was far away. This reply contains several layers of failure: (1)
355   the robot fails to capture the potential semantic inference of the expression *Mr Right*; (2) it fails
356   to consider the social norm that Mr Right belongs typically to one person only; and (3) it makes
357   a subjective judgement about distance. One may argue that this error would not happen if the
358   user knew a question-answer robot could not chat casually. However, the issue is whether a clear
359   boundary of a social robot's capability is set in the system or communicated to the user during the
360   interaction. It is difficult to tell why speech interfaces may fail and how to work around the limits
361   without understanding what makes interaction work and how speech assists in the process.

362   Also, spoken interaction requires interlocutors, including robots, to adjust their behaviours based
363   on the verbal and non-verbal feedback provided by others. A social robot that does not react
364   appropriately could be deemed improperly functional, as illustrated in the following sequence. In
365   the scenario, the robot failed to generate satisfactory answers several times in an open conversation;
366   the user felt frustrated.

```
367   User:  You are generating GPT rubbish.
```

```
368  Robot: (No response, carries on)
```

369   *3.2.1.4   Hey Siri . . . You Don't Know How to Interact, huh?*

370   The WTF moment Wiltschko presented concerned the use of *huh* in interaction with Siri, Apple's
371   voice assistant.

```
372  User: Hey Siri, send an e-mail.
373  Siri: To whom shall I send it?
374  User: huh?
375  Siri: I couldn't find huh in your contacts. To whom shall I send it?
```

376   It is evident from the example that Siri cannot understand *huh*. This is true for *huh* used as an
377   other-initiated repair strategy as in the example above, but it is also true for its use as a sentence-final
378   tag. This is a significant failure as in human-human interaction the use of *huh* is ubiquitous. In fact,
379   *huh* as a repair strategy has been shown to be available across a number of unrelated languages
380   (Dingemanse et al., 2013). Wiltschko speculates that successful language use in machines is restricted
381   to propositional language (i.e., language used to convey content) whereas severe problems arise in
382   the domain of interactional language (i.e., language used to regulate common ground building as
383   well as the conversational interaction itself). The question that arises, however, is whether human
384   users feel the need to use interactional language with machines. After all, this aspect of language
385   presupposes interaction with another mind for the purpose of common ground construction and it
386   is not immediately clear whether humans treat machines as having a mind with which to share a
387   common ground.

388   *3.2.1.5   Utilising Explanations to Mitigate Robot Failures*

389   Kontogiorgos presented current work on failure detection (Kontogiorgos et al., 2020a, 2021)
390   and how robot failures can be used as an opportunity to examine robot explainable behaviours.
391   Typical human-robot interactions suffer from real-world and large-scale experimentation and tend to
392   ignore the 'imperfectness' of the everyday user (Kontogiorgos et al., 2020b). Robot explanations
393   can be used to approach and mitigate robot failures by expressing robot legibility and incapability
394   (Kwon et al., 2018), and within the perspective of common-ground. The presenter discussed
395   how failures display opportunities for robots to convey explainable behaviours in interactive
396   conversational robots according to the view that miscommunication is a common phenomenon
397   in human-human conversation and that failures should be viewed as being an inherent part of
398   human-robot communication. Explanations, in this view, are not only justifications for robot actions,
399   but also embodied demonstrations of mitigating failures by acting through multi-modal behaviours.

3.2.1.6 *Challenging Environments for Debugging Voice Interactions*

Porcheron presented the challenge of how we expect users to understand and debug issues with 'eyes-free voice interactions', and of parallelism to the prospects of voice-based robots. A recurrent promise of voice-based technologies is their simplicity: we issue a command to a computer and it can respond accordingly. Of course, not all technology use goes as planned and sometimes errors occur. With graphical user interfaces (GUIs), we have a plethora of well-tested heuristics (e.g., Nielsen (1995)), especially for dealing with 'errors' where users need 'fix' something. However, with voice, in situations where people encounter something going wrong, they have to carry out work to figure out how to resolve the issue (Porcheron et al., 2018; Fischer et al., 2019). One specific example is responses which do not reveal specifics, such as "I had an issue responding to that request". Users are given little purchase with which to debug this issue, and attempt to resolve this. This user challenge is exacerbated in the new settings where voice technologies are appearing: in our cars, on our bikes, and anywhere we take our smartwatch—in these settings, there is often little time to read and respond to a text, little audible information to go on, and plenty of distraction for the user. Porcheron suggested that if we want to consider voice as a modality for controlling robots, we first need to think through how we help users understand and recover from 'errors' in these sorts of environments first.

3.2.1.7 *The Robot [does not] Hear Me Correctly*

Hernández García presented examples of failures of user interactions with a receptionist robot. Gunson et al. (2022) developed a social conversational robot that can act as a receptionist in a hospital waiting room for tasks such as greeting visitors, answering common questions, providing directions and/or scheduling information, etc. During the evaluation with users (student participants in a lab setting), a majority of failed interactions that prevented user task completion were the result of problems with the ASR. These ASR failures predominately took the form of an ASR result that contained only part of the user's utterance, misrecognition of critical words, or a combination of the two (to the extent that the user's original intention could not be discerned), leading the dialogue management into a non-coherent path and frustrating the user.

```
USER: can you help me to find the exit please | ASR: please
ROBOT: Anyway it's a big building we're in, isn't it.
Did you know you can ask me for help if you ever get lost here?
USER: can you direct me to the exit please | ASR: can you direct me
to Netflix please
ROBOT: Such a lot of waiting around isn't there.
USER: thank you  [user left]
```

434   Participants rated the conversation negatively when these errors occurred, as they experienced
435 difficulties in making themselves understood. The user evaluations reported by Gunson et al. (2022)
436 highlighted that users did not feel it was *natural* or that it *flowed* in the way they expected. Participants
437 did not believe that *"the robot heard me correctly most of the time"* or that *"the robot recognised the
438 words I said most of the time"* nor *"felt confident the robot understood the meaning of my words"*.

439   Conversational troubles may start at a *speech recognition* level, but these failures are propagated
440 throughout the whole *speech interface* pipeline, compounding to create WTF moments and leading
441 to poor performance, increasing user frustration, and loss of trust, etc.

3.2.1.8   Hello, It's Nice to "Meat" You

443   Nesset shared examples of WTF moments encountered while interacting with Norwegian chatbots
444 through written text. The first failure presented was users' committing spelling mistakes interacting
445 with a virtual agent through chat. This caused the agent to misunderstand the overall context of the
446 conversation. A good example of this is misspelling meet with meat, and the chatbot then replying
447 with a response about sausages.

448   The second part entailed a user failure that is specifically for multilingual users. In some non-native
449 English-speaking countries, such as Norway, technical terms and newer words are often commonly
450 said in English. This potentially leads users to interact with agents in two languages within the same
451 sentence/conversation. This can lead to the agent struggling to interpret the terms in the second
452 language, and assuming that they mean something else in the original interaction language. These
453 are some examples of how uncertain user output can result in failures from the robot.

3.2.1.9   Speech Misrecognition: A Potential Problem for Collaborative Interaction in
Table-grape Vineyards

456   Kaszuba presented troubles and failures encountered while designing a spoken human-robot
457 interaction system for the *CANOPIES project*[4]. This project aims to develop a collaborative paradigm
458 for human workers and multi-robot teams in precision agriculture, specifically in table-grape
459 vineyards. When comparing some already existing speech recognition modules (both online and
460 offline), the presenter identified communication issues associated with the understanding and
461 interpretation of specific words of the vineyard scenario, such as "grape", "bunch", and "branch".
462 Most of the tested applications could not clearly interpret such terms, leading the user to repeat the
463 same sentence/word multiple times.

464   Hence, the most significant source of failure in speech interfaces that Kaszuba has described is
465 *speech misrecognition*. Such an issue is particularly relevant, since the quality and effectiveness of

---

[4] `https://www.canopies-project.eu/`

the interaction strictly depend on the percentage of words correctly understood and interpreted. For this reason, the choice of the application scenario has a crucial role in the spoken interaction, and preliminary analysis should be taken into consideration when developing such systems, as the type and position of the acquisition device, the ambient noise and the ASR module to adopt. Nevertheless, misrecognition and uncertainty are unavoidable when the developed application requires people to interact in outdoor environments and communicate in a language that is not the users' native language.

Hence, some relevant considerations concerning ASR modules should be taken into account in order to implement a robust system that, eventually, can also be exploited in different application scenarios. The percentage of uncertainty, the number of misrecognized words and the environmental noise that can negatively affect communication are some fundamental issues that must be addressed and minimized.

3.2.1.10 *Leveraging Multimodal Signals in Human Motion Data During Miscommunication Instances*

Approaching from a natural dialogue standpoint and inspired by the Running Repairs Hypothesis Healey et al. (2018b), Özkan shared a presentation on why and how we should take advantage of WTF-moments or miscommunications to regulate shared understanding between humans and speech interfaces. Rather than avoiding these moments (which is impossible), if speech interfaces were to identify them and show appropriate behaviour, it could result in more natural, dynamic and effective communication.

Detecting miscommunications from the audio signal can only can be costly in terms of computational load or prone to error due to noise in most environments. Fortunately, repair phenomena manifest themselves in non-verbal signals as well Healey et al. (2015); Howes et al. (2016). Findings regarding speaker motion during speech disfluencies (self-initiated self-repairs) have shown that there are significant patterns in the vicinity of these moments Özkan et al. (2021, 2023); Ozkan et al. (2022). Specifically, the speakers have higher hand and head positions and velocities near disfluencies. This could be treated as a clear indicator for artificial interfaces to identify troubles of speaking in their human partner. For example, to the user input *"Could you check the flights to Paris -uh, I mean- Berlin?"*, the interface, instead of disregarding the uncertain utterance, could offer repair options more actively by returning *"Do you mean Paris or Berlin?"* in a collaborative manner.

Though not in the context of disfluencies, a common example of not allowing repair (in this case other initiated other repair) occurs when the user needs to correct the output of an interface or simply demand another response to a given input. As a WTF moment in the repair context, Özkan demonstrated a frequent problem in their interaction with Amazon Alexa. When asked to play a

501 certain song, Alexa would play another song with the same or similar name. The error is not due to
502 speech recognition, because Alexa understands the name of the song very well. However, it maps
503 the name to a different song that the user does not want to hear. No matter how many times the
504 user tries the same song name input, even with the artist name, Alexa would still pick the one that
505 is the 'first' result of its search. If the conversational repair was embedded in the design, a simple
506 solution to this problem could have been *"Alexa, not that one, can you try another song with the*
507 *same name?"*, but Alexa does not respond to such requests.

508 3.2.2   Technical Aspects of Conversational Failure

509   The following five of the contributions describe technical aspects of failures. Presentations in this
510 section either discuss the technical causes of failures, point out technological attempts to recognize
511 when conversational trouble occurs, or summarize approaches on handling troubles on part of the
512 robot.

513 **3.2.2.1   Chefbot: Reframing Failure as a Dialogue Goal Change**

514   Gkatzia presented their work on *Chefbot*, a cross-platform dialogue system that aims to help users
515 prepare recipes (Strathearn and Gkatzia, 2021a). The task moves away from classic instruction
516 giving and incorporates question-answering for clarification requests, and commonsense abilities,
517 such as swapping ingredients and requesting information on how to use or locate specific utensils
518 (Strathearn and Gkatzia, 2021b). This results in altering the goal of the communication from cooking
519 a recipe to requesting information on how to use a tool, and then returning to the main goal. It
520 was quickly observed that changing the dialogue goal from completing the recipe to providing
521 information about relevant tasks resulted in failure of task completion. This issue was subsequently
522 addressed by *reframing* failure as a temporary dialogue goal change, which allowed the users to
523 engage in question answering that was not grounded to the recipe document, and then forcing the
524 system to resume the original goal.

525 **3.2.2.2   Failure in Speech Interfacing with Local Dialect in a Noisy Environment**

526   Liza (Farhana) presented their ongoing work in capturing the linguistic variation of speech
527 interfaces in real-world scenarios. Specifically, local dialects may impose challenges when modelling
528 a speech interface using an artificial intelligence (deep learning) language modelling system. Deep
529 learning speech interfaces rely on language modelling which is trained on large datasets. A large
530 dataset can capture some linguistic variations; however, dialect-level variation is difficult to capture
531 as a large enough dataset is unavailable. Moreover, very large models require high-performance
532 computation resources (e.g., GPU) and take a long time to respond, which imposes further constraints
533 in terms of deploying such systems in real scenarios. Large data-driven solutions also cannot easily
534 deal with noise as it is impractical to give access to enough real-world data from noisy environments.

535 Overall, state-of-the-art AI models are still not deployable in scenarios with dialect variation and
536 noisy environments. Alharbi et al. (2021) identified several hurdles in training end-to-end Automatic
537 Speech Recognition (ASR) models. Additionally, the conditional interdependence between the
538 acoustic encoder and the language model was emphasized by (Xu et al., 2020). Consequently, while
539 augmenting the standard text training data can enhance the efficacy of general-purpose language
540 models, the limited availability of corresponding acoustic data poses challenges in training end-to-
541 end ASR systems. Moreover, when addressing dialect modeling (Hirayama et al., 2015), the scarcity
542 of training data exacerbates the difficulties in integrating speech interfacing and language modeling
543 (Liza, 2019) within the ASR framework.

544 3.2.2.3 The 'W' in WTF Moments can also be 'When': The Importance of Timing and
545   Fluidity

546   Hough presented WTF moments driven more by inappropriate timing of responses to user
547 utterances, rather than by content misunderstandings. Improving the first-time accuracy of Spoken
548 Language Understanding (SLU) remains a priority for HRI, particularly given errors in speech
549 recognition, computer vision and natural language understanding remain pervasive in real-world
550 systems. However, building systems capable of tolerating errors whilst maintaining *interactive*
551 *fluidity* is an equally important challenge. In human-human situated interactions where an instructee
552 responds to a spoken instruction like "put the remote control on the table" and a follow-up repair
553 like "no, the left-hand table" when the speaker realizes the instructee has made a mistake, there is
554 no delay in reacting to the initial instruction, and adaptation to the correction is instant (Heldner
555 and Edlund, 2010; Hough et al., 2015), in stark contrast to state-of-the-art robots with speech
556 interfaces. Increasing interactive fluidity is vital to give robots with speech understanding more
557 seamless, human-like transitions from processing speech to taking physical action without delay,
558 permitting appropriate overlap between the two, and the ability to repair actions in real-time. Rather
559 than waiting for components to be perfected, preliminary experiments with a pick-and-place robot
560 show users can be tolerant of errors if fluidity is kept high, including appropriate repair mechanisms
561 (Hough and Schlangen, 2016).

562 3.2.2.4 Laughter in WTF Moments

563   Maraev presented a hypothesis that laughter can be treated as an indicator of a WTF moment.
564 Laughter can occur in such moments as a) speech recognition failures disclosed to a user via explicit
565 grounding feedback, b) awkwardness due to retrieval difficulties, c) resulting system apologies and
566 down players (e.g., "don't worry"). Along with examples from task-oriented role-played dialogues,
567 Maraev discussed the following constructed example, where laughter communicates a negative
568 feedback to the system's clarification of speech recognition result:

```
569  Usr> I would like to order a vegan bean burger.
570  Sys> I understood you'd like to order a vegan beef burger. Is that correct?
571  Usr> HAHAHA
```

572 Maraev et al. (2021) focused on non-humorous laughs in task-oriented spoken dialogue systems.
573 The paper shows how certain types of laughter can be processed within the dialogue manager and
574 natural language generator, namely: laughter as negative feedback, laughter as a negative answer to
575 a polar question and laughter as a signal accompanying system feedback.

576 *3.2.2.5 To Err is Robot*

577 Giuliani presented findings from six years of research on erroneous human-robot interactions.
578 The team of researchers led by Giuliani has shown that participants in human-robot interaction
579 studies show unique patterns of social signals when they experience an erroneous situation with
580 a robot (Mirnig et al., 2015). The team annotated two large video corpora of 201 videos showing
581 578 erroneous situations and 1200 videos showing 600 erroneous situations, respectively (Giuliani
582 et al., 2015; Cahya et al., 2019). They found that there are two types of errors that do occur in
583 human-robot interaction. Social norm violations are situations in which the robot does not adhere
584 to the underlying social script of the interaction. Technical failures are caused by the technical
585 shortcomings of the robot. The results of the video analysis show that the study participants use
586 many head movements and very few gestures but they often smile when in an error situation with
587 the robot. Another result is that the participants sometimes stop moving at the beginning of error
588 situations. The team was also able to show in a user study for which a robot was purposefully
589 programmed with faulty behaviour that participants liked the faulty robot significantly better than
590 the robot that interacted flawlessly (Mirnig et al., 2017). Finally, the team trained a statistical model
591 for the automatic detection of erroneous situations using machine learning (Trung et al., 2017). The
592 results of this work demonstrate that automatic detection of an error situation works well when the
593 robot has seen the human before.

594 3.2.3 Adjacent Topics in Speech Interfaces

595 The two contributions under this theme do not discuss conversational failures directly but address
596 the related topics of explanatory AI and privacy of speech interfaces.

597 *3.2.3.1 What is a 'Good' Explanation?*

598 Kapetanios presented some thoughts around the long-standing research question of *what is a*
599 *good explanation* in the context of the current buzz around the topics of explainable AI (XAI)
600 and interpretable Machine Learning (IML). Using Amazon's Alexa and Google's Digital Assistant
601 to generate explanations for answers being given to questions being asked of these systems, he

602 demonstrated that both systems, at the technological forefront of voice-based HCI approaches to
603 answering specific questions, fail to generate convincing explanations. Convincing explanations
604 should fit the facts, be relevant, tailored to the recipient, and typically do more than merely describe
605 a situation (Dowden, 2019, chap. 14). It is frequently the latter where digital assistants have been
606 observed to struggle. Hence, when describing the results of running several thousand queries through
607 the most common digital assistants, provides the following example (Enge, 2019):
608 Siri, when being asked the question "Who is the voice of Darth Vader?", instead of providing
609 the name of the (voice) actor, returns a list of movies featuring Darth Vader. While this answer
610 is topically relevant, it certainly is not a proper answer to the question. The same problem of
611 explanation persists with ChatGTP-3/4, despite its fluency in generating precise answers to specific
612 questions in natural language.

613 *3.2.3.2 Privacy and Security Issues with Voice Interfaces*

614 Williams presented privacy and security issues and how these are often underestimated, overlooked,
615 or unknown to users who interact with voice interfaces. What many voice interface users are unaware
616 of is that only three to five seconds of speech are required to create a *voiceprint* of a person's real
617 voice as they are speaking (Luong and Yamagishi, 2020). One of the risks that follows is that
618 voiceprints can be re-used in other voice applications to impersonate or create voice deepfakes
619 (Williams et al., 2021b,a). In the UK and many other countries, this poses a particular security risk
620 as voice-authentication is commonly used for telephone banking and call centres. In addition, some
621 people may be alarmed when a voice interface reveals private information by "speaking out loud"
622 sensitive addresses, birth dates, account numbers, or medical conditions. Anyone in the nearby
623 vicinity may overhear this sensitive information and technology users have no ability to control what
624 kinds of information a voice interface may say aloud (Williams et al., 2022).

625 3.2.4 Summary of Lightning Talks

626 Through their lightning talks, our participants contributed to an initial gathering of different
627 troubles and failures in conversational interactions between humans and robots. Thanks to the
628 description of their memorable failures and their analysis, we could identify the themes of *analysis*,
629 *technical aspects* and *adjacent topics*, which all impact the success (or failure) of a conversation.

630 **3.3 Summary of World Café Session**

631 During the World Café session, four working groups were created based on recurring themes
632 from the lightning talks, participants' answers as to what they perceived as the most pressing issue
633 or the biggest source of failure for speech interfaces, as well as the aim to define the sought after
634 benchmark scenario. Through the initial submissions of the participants, their lightning talks and the
635 keynotes, three main macro-categories have emerged: i) miscommunication, ranging from speech

636 recognition failures to more semantic and conversation-dependent failures; ii) interaction problems,
637 encompassing all those failures that are due to users' expectations and behaviours; iii) context
638 understanding, linked to the fact that interaction is shaped by context and that context changes fast,
639 calling for a need to find more robust ways to establish common ground. While these three themes
640 are highly interdependent and could culminate in the sought after benchmark scenario (the fourth
641 working group), each of them presents peculiarities that we considered worth discussing in detail.

3.3.1 Handling Miscommunication

643 The discussion focused on the need to acknowledge and embrace the concept of miscommunication.
644 One of the open challenges identified by this group was to equip robots with the ability to learn
645 from various forms of miscommunication and to actively use them as an opportunity to establish
646 common ground between users and robots. When communicating with a robot, the human user
647 usually has a goal in mind. The robot could exploit miscommunication to understand this goal
648 better by asking for clarifications at the right moments and updating the common ground. The
649 discussion also acknowledged that miscommunication is only the starting point. Two distinct new
650 challenges and opportunities arise when working on resolving miscommunication: 1) how to explain
651 the miscommunication, and 2) how to move the conversation forward. Both problems are highly
652 context-dependent and related to the severity and type of miscommunication. Moreover, being
653 able to repair a breakdown in conversation may also depend on being able to establish appropriate
654 user expectations in the first place by giving an accurate account of what the robot is really able
655 to accomplish. The final discussion point from this group centered on the possibility of enriching
656 the multimodal and non-verbal component of conversations to help the robot perceive when a
657 miscommunication has happened by detecting and responding to, for example, long pauses or
658 changes in specific types of facial expressions.

3.3.2 Interaction Problems

660 Interaction problems do not only encompass challenges that are specific to the technology used,
661 like issues with automatic speech recognition or the presence of long delays when trying to engage
662 in a "natural" conversation. They are related to perceived failures that longitudinally include all the
663 technical problems identified by the other themes and relate to how the interaction with the human
664 user is managed. In this context, human users play an essential role and the participants of this
665 group emphasized the necessity of creating expectations that allow users to build an adequate mental
666 model of the technology they are interacting with. In Washburn et al. (2020a), authors examine how
667 expectations for robot functionality affected participants' perceptions of the reliability and trust of a
668 robot that makes errors. The hope is that this would lead to an increased willingness and capacity
669 to work with the failures that inevitably occur in conversational interactions. Anthropomorphism
670 was identified as one of the possible causes for the creation of wrong expectations: the way robots

both look and speak risks tricking users into thinking that robots have human-like abilities and are able to follow social norms. Once this belief is abandoned, users could then form an appropriate expectation of the artificial agents, and the severity of the failures would decrease. Setting the right expectations will also enable users to understand when a failure is a technological error in execution or when it is a design problem: humans are unpredictable, and some of the problems that arise in the interactions are due to users' behaviours that were not embedded in the design of robot's behaviours. A related aspect that was considered important by this group is the transparency of the interaction: the rationale behind the failures should be explained and made clear to the users to enable mutual understanding of the situation and prompt recovery. This could, in fact, be initiated by the users themselves. Another need, identified as a possible way to establish better conversational interactions, is the missing link of personalisation. The more the agents are able to adapt to the context and the users they are interacting with, the more they will be accepted, as acceptance plays a fundamental role in failure management. A general consensus converged regarding the fact that we are not yet at the stage where we can develop all-purpose chatbots - or robots - and the general public should be made aware of this, too. Each deployment of conversational agents is context related and the conversation is mainly task-oriented, where a precise exchange of information needs to happen for a scenario to unfold.

3.3.3 Context Understanding

All four groups agreed that context understanding is crucial for reducing or entirely eliminating failures of interactive systems that use spoken language. We determined that capturing and modelling context is particularly challenging since it is an unbound and potentially all-encompassing problem. Moreover, all dialogue, and in fact, interaction as a whole, would be *shaped by* the context while at the same time *renewing* it. Likewise, the volatility of context, in particular, potentially rapid context switches, was also identified as challenging in human-robot conversation. Modelling the interaction partner(s) and evaluating their focus of attention was thereby discussed as one potential approach to reducing context search space.

A precise and consistent representation of the dialogue context was therefore identified as one of the most important problems that would rely on modelling not only the current situation but also any prior experiences of humans with whom the system is interacting. Such previous experience was seen to have significant effects on expectations about the interactive system that would potentially require calibration before or during system runtime to avoid misunderstandings as well as misaligned trust towards the system Hancock et al. (2011). However, even if we assume an optimal representation of context would be possible, the problem of prioritisation and weighting would still persist.

Another challenge discussed was the need for a multi-modal representation of the current situation comprised of nonverbal signals, irregular words, and interjections. Such a model would be required

706 for an appropriate formulation of common ground, whereby it remains unclear what exactly would
707 be required to include. In that context, one group identified the benefits of a typology that could
708 encompass an interaction situation in a multi-modal way, potentially extending work by Holthaus
709 et al. (2023). The exact mapping between a signal or lexical index and their meanings is, however,
710 still difficult to establish.

711 On the other hand, considering the dialogue context was unanimously regarded as beneficial to
712 enrich human-robot conversations offering numerous opportunities to increase its functionality, even
713 if it would not be possible to capture all context comprehensively. With a personalised model of
714 interaction partners, for example, the spoken dialogue could be enhanced by taking into account
715 personal interaction histories and preferences. Conversational agents could be improved for highly
716 constrained settings and converge faster to relevant topics.

717 It is noteworthy to mention that enriching the capabilities of conversational agents with context
718 information poses ethical challenges, e.g. in terms of privacy and data protection. This approach
719 might thus introduce barriers in terms of user acceptance that need to be considered Lau et al. (2018).
720 However, using context appropriately could also help to improve a system's transparency either by
721 designing it with its intended context in mind or by utilising it during a conversation, for example,
722 by providing additional interfaces to transport further information supporting the dialogue or by
723 analysing context to reduce ambiguities and eliminate noise. The context was regarded to often play
724 a vital role in providing the necessary semantic frame to determine the correct meaning of spoken
725 language. Making use of domain and task knowledge was thereby identified as particularly helpful.

726 Moreover, intentionally misapplying context or analysing situations where context has previously
727 misled a conversation, might be avenues to recognize and generate error patterns to help detect
728 future troubles and failures in speech understanding.

729 ### 3.3.4 Benchmark Scenario(s)

730 On this discussion table, participants struggled to devise a single benchmark scenario that would
731 elicit most, if not all, commonly occurring conversational failures. As a main reason for the difficulty
732 of identifying such a prototypical scenario, the lack of a comprehensive taxonomy of conversational
733 failures was determined.
734 An alternative suggestion to the proposed task of identifying one, failure-wise all encompassing,
735 scenario was also made. Rather than seeking to specify a single scenario, it may be necessary
736 to create test plans for each specific interaction task using chaos engineering, with some of the
737 defining characteristics for a scenario being (1) the type(s) of users, (2) the domain of use (e.g.
738 health-related, shopping mall information kiosk), (3) the concrete task of the robot, (4) the types
739 of errors under investigation. Chaos engineering is typically used to introduce a certain level of
740 resilience to large distributed systems (cf. Fomunyam (2020). Using this technique, large online

741 retailers such as Amazon deliberately knock out some of their subsystems, or introduce other kinds
742 of errors, to ensure that the overall service can still be provided despite the failure of one or more
743 of these, typically redundant, components (cf. Siwach et al. (2022)). While both the envisioned
744 benchmark scenario(s) and chaos engineering are meant to expose potential failures of human-made
745 systems, the types of systems and types of failure differ substantially. While failures in technical
746 distributed systems are unilateral, in the sense that the source of failure is typically attributed solely
747 to the system rather than its user, attribution of blame in conversational failure is less unilateral. If a
748 successful conversation is seen to be a joint achievement of at least two speakers, conversational
749 failure is probably also best seen as a joint "achievement" of sorts. In other words, the _user_ of a
750 conversational robot is always also an interlocutor during the interaction. Hence, whatever approach
751 we use to identify and correct conversational failures, the correct level of analysis is that of the dyad
752 rather than of the robot alone.

753 Independent of the chaos engineering approach, another suggestion was that at least two benchmarks
754 might be needed in order to distinguish between low-risk and high-risk conversations. Here, low-risk
755 conversations would be the more casual conversations that one may have with a shop assistant whose
756 failure would not carry any hefty consequences. High-risk conversations, on the other hand, would
757 be those where the consequences of conversational failure might be grave - imagine conversational
758 failure between an assistive robot and its human user that are engaged in some joint task of removing
759 radioactive materials from a decommissioned nuclear site. If such a distinction should be made, the
760 logical follow-up question would be how the boundary between low and high-risk scenarios should
761 be determined. Finally, it should be mentioned that at least partial benchmarks such as _Paradise_
762 exist for the evaluation of spoken dialogue systems Walker et al. (1997).

4 DISCUSSION

763 One significant result from the workshop is that no succinct and, more importantly, singular
764 benchmark scenario could be envisioned that would likely elicit all or, at least, a majority of
765 identified failures. A likely reason behind this is the lack of a comprehensive categorization of
766 conversational failures and their triggers in mixed human-machine interactions. Having such a
767 taxonomy would allow us to embed such triggers systematically in benchmark scenarios.

4.1 Wanted: A Taxonomy of Conversational Failures in HRI

769 Honig and Oron-Gilad (2018) recently proposed a taxonomy for failures in HRI based on a
770 literature review of prior failure-related HRI studies. Their survey indicated a great asymmetry in
771 these investigations, in that the majority of previous work focused on technical failures of the robot.
772 In contrast, Honig & Oron-Gilad noticed that no strategies had been proposed to deal with "human
773 errors". From a conversation analytic viewpoint, the dichotomy of technical vs. human error may not

774   always be as absolute when applied to conversational failures, especially since, despite sharing some
775   terminology, CA conceptualizes conversational success and failure quite differently. Conversation
776   analysts conceive of successful conversation as the achievement of joint action by any party (robot or
777   human). In this sense, when a failure occurs, the 'blame' lies with all participants. Similarly, success
778   in CA terms might mean that a joint action is 'successfully' achieved interactionally, even if there
779   are informational errors. For example, an invitation to meet under the clock at Grand Central station,
780   where the recipient misunderstands the time/place might be 'successfully' achieved as an orderly
781   interaction, the error being marked. In HRI, however, this failure of the 'Schelling game' would
782   be considered a classic 'grounding error' Clark (1996), and it would certainly matter who made
783   the error: the human or robot. While not assigning blame for some singular failure simultaneously
784   to both participants, Uchida et al. (2019a) recently used a blame assignment strategy where the
785   responsibility for a sequence of failures was attributed in an alternating fashion to the robot and
786   the human. As indicated by our struggle to find a good general characterisation of conversational
787   failures during the workshop, we advocate the construction of a taxonomy of conversational failures
788   for mixed, that is human-machine dyads and groups. To build such a taxonomy, an interdisciplinary
789   effort is needed, given that the types of relevant failures span the entire spectrum from the very
790   technical (e.g. ASR errors) to the very "relational" (e.g. misunderstanding based on lack of common
791   ground). The relevant disciplines would include linguistics, conversation analysis, robotics, NLP,
792   HRI, and HCI. This workshop represented the first stepping stone towards this interdisciplinary
793   effort. One theory-related advantage of taxonomy building is that it forces us to reconsider theoretical
794   constructs from different disciplines, thereby potentially exposing gaps in the respective theories -
795   similarly to how conversation analysis has exposed shortcomings of speech act theory (cf. Levinson,
796   1983).
797   The process of defining the types of errors could also help us to understand why they arise, measure
798   their impact and explore possibilities and appropriate ways to detect, mitigate and recover from
799   them. If, for example, artificial agents and human users are mismatched conversational partners as
800   suggested by Moore (2007) and Förster et al. (2019), and if this mismatch creates constraints and a
801   "habitability gap" in HRI (Moore, 2017), are their specific types of failures that only occur due to
802   such asymmetric setups? And, if yes, what does that mean for potential error management in HRI?
803   If priors shared between interlocutors matter (Moore, 2022; Huang and Moore, 2022), how does
804   the aligning of interactive affordances help to increase the system's capacity to deal with errors?
805   Moreover, errors can affect people's perception of a robot's trustworthiness and reliability (e.g.,
806   Washburn et al., 2020b), as well as their acceptance and willingness to cooperate in HRI (e.g., Salem
807   et al., 2015). What type of errors matters more? In terms of error recovery, it has been shown that
808   social signals, such as facial action unit (AU), can enhance error detection (Stiber et al., 2023);
809   Users' cooperative intention can be elicited to avoid or repair from dialogue breakdowns (Uchida
810   et al., 2019b). The question is, when facing different errors, do these strategies need to be adaptable

811 to tasks/scenarios, and if so, to what degree? Answering the above questions requires a deeper
812 understanding of conversational failures, and taxonomy building is one possible way to increase our
813 understanding.
814 A more practical advantage of having such a taxonomy is discussed in the next section.

4.2 Benchmarking Multimodal Speech Interfaces

816 One of the intended aims of the workshop was to define, or at least outline, some benchmark
817 scenario that would have the "built-in" capacity to expose, if not all, at least a good number of
818 potential communicative failures of some given speech interface. During the workshop, it became
819 apparent that we would fail to come up with such a single scenario. It questionable whether such a
820 scenario could exist or whether a number of scenarios would be needed to target different settings in
821 which the speech interface is to be deployed. One main reason for our struggle that emerged during
822 the World Café session was the lack of a taxonomy of communicative failures in HRI. Having such
823 a taxonomy would allow the designer, or user, of a speech interface to systematically check whether
824 it could handle the type of situation in which the identified failures are likely to occur prior to testing
825 it "in the wild".
826 Related to the construction of a potential (set of) benchmarks is the question of how to evaluate
827 multimodal speech interfaces. The popular evaluation framework PARADISE Walker et al. (1997),
828 originally designed for the assessment of unimodal dialogue systems, has already been used in
829 multimodal HRI studies (e.g. Giuliani et al., 2013; Hwang et al., 2020; Peltason et al., 2012). Also
830 within the HCI community multimodal alternatives to PARADISE have been proposed (e.g. Kühnel,
831 2012). Given these existing evaluation frameworks for multimodal dialogue systems, what would a
832 failure-based method bring to the table?
833 A characteristic of PARADISE and related frameworks is that they tend to evaluate a past dialogue
834 according to a set of positive performance criteria. PARADISE, for example, uses measurements of
835 *task success*, *dialogue efficiency*, and *dialogue quality* to score a given dialogue. There is likely an
836 inverse relationship between a failure-based evaluation and, for example, *dialogue efficiency* as a
837 dialogue containing more failures, will likely require more turns to accomplish the same task due
838 to repair-related turns. This would mean that the efficiency of this failure-laden dialogue would be
839 reduced. However, despite this relationship, the two methods are not commensurate. A failure-based
840 scoring method could, for example, put positive value on the resilience of some speech interface,
841 by assigning positive values to the number of successful repairs. This would, in some sense, be
842 diametrically juxtaposed to efficiency measures. On the other hand, these two ways of assessing a
843 speech interface are not mutually exclusive and could be applied simultaneously.
844 One interesting observation with respect to the surveyed studies points to a potential limitation
845 of existing evaluation frameworks such as PARADISE. All of the referenced studies are based
846 on turn-based interaction formats. While turn-based interaction is certainly a common format in

847 many forms of human-human and human-robot interaction, it is likely not the only one. Physical
848 human-robot collaboration tasks which require participants to coordinate their actions in a near-
849 simultaneous manner, for example when carrying some heavy object together, do not necessarily
850 follow a turn-based format. While some of the involved communication channels such as speech
851 will likely be turn-based, other channels such as sensorimotor communication (SMC, cf. Pezzulo
852 et al., 2019) may or may not follow this format.

5 CONCLUSION

853 The first workshop on "Working with Troubles and Failures in Conversation between Humans and
854 Robots" was the first effort to gather an interdisciplinary team of researchers interested in openly
855 discuss the challenges and opportunities in designing and deploying speech interfaces for robots.
856 Thanks to insights from conversation analysis, cognitive science, linguistics, robotics, human-robot
857 interaction, and dialogue systems, we initiated a discussion that does not simply dismiss failures in
858 conversational interaction as a negative outcome of the robotic system, but engages with the nature of
859 such failures and the opportunities that arise from using them to improve the interactions. We believe
860 this initial push will spawn a deeper research effort towards the identification of a benchmark for
861 multimodal speech interfaces and the creation of a systematic taxonomy of failures in conversation
862 between humans and robots which could be useful to interaction designers, both in robotics and
863 non-robotics fields.

6 NOMENCLATURE

864 **Voice interfaces**: User interfaces that allow interaction with technology through spoken commands
865 or queries.
866 **Robotic speech interfaces**: Voice interfaces applied on robots that use both speech recognition as
867 well as synthesised or artificial voices to communicate and interact with users.
868 **Chatbots**: Text-based interfaces able to provide information, answer questions, or assist with various
869 tasks.
870 **Agents, artificial agents, conversational agents**: Terms used interchangeably for systems designed
871 to engage in natural language conversations with humans, by employing natural language processing
872 and machine learning to understand and respond to user queries, provide information or assistance.

CONFLICT OF INTEREST STATEMENT

873 Author Ioannis Papaioannou is employed by Alana AI. The remaining authors declare that the
874 research was conducted in the absence of any commercial or financial relationships that could be
875 construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

876 FF, MR, PH, LW, CD, JEF have organised the workshop, the contributions and notes of which form
877 the basis of this article. FF is the lead author and has provided the main structure of the article as
878 well as large parts of the discussion section, parts of the methods section, and overall proof-reading.
879 MR has contributed substantial parts of the methods section, the conclusion, as well as overall
880 proof-reading and improvements. PH, and JEF have contributed to parts of the methods section as
881 well as overall proof-reading and improvements. FFL, SK, JH, BN, DHG, DK, JW, EEÖ, PB, GB,
882 DP, SC, MW, LT, MP, MG, GS, PGTH, IP, DG, SA, GH, VM, EK have contributed subsections in
883 the results section and have contributed to overall proof-reading.

FUNDING

DATA AVAILABILITY STATEMENT

897 The original contributions presented in the study are included in the article/supplementary material,
898 further inquiries can be directed to the corresponding author.

REFERENCES

899 Albert, S. and Hamann, M. (2021). Putting wake words to bed: We speak wake words with
900   systematically varied prosody, but CUIs don't listen. In *CUI 2021 - 3rd Conference on*
901   *Conversational User Interfaces* (New York, NY, USA: Association for Computing Machinery),
902   CUI '21, 1–5. doi:10.1145/3469595.3469608

903 Albert, S. and Ruiter, J. P. d. (2018). Repair: The Interface Between Interaction and Cognition.
904   *Topics in Cognitive Science* 10, 279–313. doi:10.1111/tops.12339

905 Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R., et al. (2021).
906   Automatic speech recognition: Systematic literature review. *IEEE Access* 9, 131858–131876.
907   doi:10.1109/ACCESS.2021.3112535

908 Axelsson, A. and Skantze, G. (2022). Multimodal user feedback during adaptive robot-human
909   presentations. *Frontiers in Computer Science* , 135

910 Axelsson, A. and Skantze, G. (2023). Do you follow? a fully automated system for adaptive robot
911   presenters. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot
912   Interaction*. 102–111

913 Axelsson, N. and Skantze, G. (2020). Using knowledge graphs and behaviour trees for feedback-
914   aware presentation agents. In *Proceedings of the 20th ACM International Conference on Intelligent
915   Virtual Agents*. 1–8

916 Bennett, C. L., Brady, E., and Branham, S. M. (2018). Interdependence as a Frame for Assistive
917   Technology Research and Design. In *Proceedings of the 20th International ACM SIGACCESS
918   Conference on Computers and Accessibility* (New York, NY, USA: Association for Computing
919   Machinery), ASSETS '18, 161–173. doi:10.1145/3234695.3236348

920 Cahya, D. E., Ramakrishnan, R., and Giuliani, M. (2019). Static and temporal differences in social
921   signals between error-free and erroneous situations in human-robot collaboration. In *Social
922   Robotics: 11th International Conference, ICSR 2019, Madrid, Spain, November 26–29, 2019,
923   Proceedings 11* (Springer), 189–199

924 Clark, H. (1996). *Using language* (Cambridge, UK: Cambridge University Press)

925 Colman, M. and Healey, P. (2011). The distribution of repair in dialogue. In *Proceedings of the
926   Annual Meeting of the Cognitive Science Society*. vol. 33, 1563–1568

927 Cuadra, A., Li, S., Lee, H., Cho, J., and Ju, W. (2021). My bad! repairing intelligent voice assistant
928   errors improves interaction. *Proc. ACM Hum.-Comput. Interact.* 5. doi:10.1145/3449101

929 Curry, A. C., Papaioannou, I., Suglia, A., Agarwal, S., Shalyminov, I., Xu, X., et al. (2018). Alana
930   v2: Entertaining and informative open-domain social dialogue using ontologies and entity linking.
931   In *1st Proceedings of Alexa Prize (Alexa Prize 2018)*

932 Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., et al. (2015). Universal
933   principles in the repair of communication problems. *PloS one* 10, e0136100

934 Dingemanse, M., Torreira, F., and Enfield, N. J. (2013). Is "Huh?" a Universal Word? Conversational
935   Infrastructure and the Convergent Evolution of Linguistic Items. *PLoS ONE* 8, e78273. doi:10.
936   1371/journal.pone.0078273

937 Dowden, B. H. (2019). *Logical Reasoning* (LibreTexts)

938 Enfield, N. (2017). *How We Talk: The Inner Workings of Conversation* (Hachette UK)

939 Enge,   E.   (2019).   Rating   the   smarts   of   the   digital   personal   assistants
940   in   2018.   `https://blogs.perficient.com/2018/05/01/`
941   `2018-digital-personal-assistants-study/`, last accessed 14 June 2023

942 Fischer, J. E., Reeves, S., Porcheron, M., and Sikveland, R. O. (2019). Progressivity for voice
943   interface design. In *Proceedings of the 1st International Conference on Conversational User*
944   *Interfaces* (New York, NY, USA: Association for Computing Machinery), CUI '19. doi:10.1145/
945   3342775.3342788

946 Fomunyam, K. G. (2020). Chaos engineering (principles of chaos engineering) as the pathway to
947   excellence and relevance in engineering education in africa. *International Journal of Engineering*
948   *and Advanced Technology (IJEAT)* 10, 146–151. doi:10.35940/ijeat.B3266.1010120

949 Förster, F., Saunders, J., Lehmann, H., and Nehaniv, C. L. (2019). Robots learning to say "no":
950   Prohibition and rejective mechanisms in acquisition of linguistic negation. *ACM Transactions on*
951   *Human-Robot Interaction* 8. doi:10.1145/3359618

952 Foster, M. E., Craenen, B., Deshmukh, A. A., Lemon, O., Bastianelli, E., Dondrup, C., et al. (2019).
953   Mummer: Socially intelligent human-robot interaction in public spaces. *ArXiv* abs/1909.06749

954 Giuliani, M., Mirnig, N., Stollnberger, G., Stadler, S., Buchner, R., and Tscheligi, M. (2015).
955   Systematic analysis of video data from different human–robot interaction studies: a categorization
956   of social signals during error situations. *Frontiers in Psychology* 6. doi:10.3389/fpsyg.2015.00931

957 Giuliani, M., Petrick, R. P., Foster, M. E., Gaschler, A., Isard, A., Pateraki, M., et al. (2013).
958   Comparing task-based and socially intelligent behaviour in a robot bartender. In *Proceedings*
959   *of the 15th ACM on International Conference on Multimodal Interaction* (New York, NY, USA:
960   Association for Computing Machinery), ICMI '13, 263–270. doi:10.1145/2522848.2522869

961 Goodwin, C. (2004). A Competent Speaker Who Can't Speak: The Social Life of Aphasia. *Journal*
962   *of Linguistic Anthropology* 14, 151–170. Publisher: [American Anthropological Association,
963   Wiley]

964 Goodwin, C. (2007). Interactive footing. In *Reporting Talk*, eds. E. Holt and R. Clift (Cambridge:
965   Cambridge University Press), Studies in Interactional Sociolinguistics. 16–46. doi:10.1017/
966   CBO9780511486654.003

967 Green, H. N., Islam, M. M., Ali, S., and Iqbal, T. (2022). Who's laughing nao? examining perceptions
968   of failure in a humorous robot partner. In *2022 17th ACM/IEEE International Conference on*
969   *Human-Robot Interaction (HRI)*. 313–322. doi:10.1109/HRI53351.2022.9889353

970 Gunson, N., Hernández García, D., Sieińska, W., Dondrup, C., and Lemon, O. (2022). Developing
971   a social conversational robot for the hospital waiting room. In *2022 31st IEEE International*
972   *Conference on Robot and Human Interactive Communication (RO-MAN)*. 1352–1357. doi:10.
973   1109/RO-MAN53752.2022.9900827

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., and Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors* 53, 517–527. doi:10.1177/0018720811417254. PMID: 22046724

Healey, P. (2008). Interactive misalignment: The role of repair in the development of group sub-languages. *Language in Flux. College Publications* 212

Healey, P., Plant, N., Howes, C., and Lavelle, M. (2015). When words fail: Collaborative gestures during clarification dialogues. In *2015 AAAI Spring Symposium Series*

Healey, P. G. (1997). Expertise or expertese?: The emergence of task-oriented sub-languages. In *Proceedings of the 19th annual conference of the cognitive science society* (Stanford University Stanford, CA), 301–306

Healey, P. G., De Ruiter, J. P., and Mills, G. J. (2018a). Editors' introduction: miscommunication. *Topics in Cognitive Science* 10, 264–278

Healey, P. G., Mills, G. J., Eshghi, A., and Howes, C. (2018b). Running repairs: Coordinating meaning in dialogue. *Topics in cognitive science* 10, 367–388

Heldner, M. and Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics* 38, 555–568

Hirayama, N., Yoshino, K., Itoyama, K., Mori, S., and Okuno, H. G. (2015). Automatic speech recognition for mixed dialect utterances by mixing dialect language models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, 373–382

Holtgraves, T. M. (2013). *Language as social action: Social psychology and language use* (Psychology Press)

Holthaus, P., Schulz, T., Lakatos, G., and Soma, R. (2023). Communicative Robot Signals: Presenting a New Typology for Human-Robot Interaction. In *International Conference on Human-Robot Interaction (HRI 2023)* (Stockholm, Sweden: ACM/IEEE), 132–141. doi:10.1145/3568162.3578631

Honig, S. and Oron-Gilad, T. (2018). Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in Psychology* 9. doi:10.3389/fpsyg.2018.00861

Hough, J., de Kok, I., Schlangen, D., and Kopp, S. (2015). Timing and grounding in motor skill coaching interaction: Consequences for the information state. In *Proceedings of the 19th SemDial Workshop on the Semantics and Pragmatics of Dialogue (goDIAL)*. 86–94

Hough, J. and Schlangen, D. (2016). Investigating fluidity for human-robot interaction with real-time, real-world grounding strategies. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (Los Angeles: ACL), 288–298

Howes, C. and Eshghi, A. (2021). Feedback relevance spaces: Interactional constraints on processing contexts in dynamic syntax. *Journal of Logic, Language and Information* 30, 331–362

Howes, C., Lavelle, M., Healey, P., Hough, J., and McCabe, R. (2016). Helping hands? gesture and self-repair in schizophrenia. In *Proceedings of the Resources and Processing of Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric Impairments (RaPID-2016)*. 9–13

Huang, G. and Moore, R. K. (2022). Is honesty the best policy for mismatched partners? aligning multi-modal affordances of a social robot: an opinion paper. *Frontiers in Virtual Reality*

Hwang, E. J., Kyu Ahn, B., Macdonald, B. A., and Seok Ahn, H. (2020). Demonstration of hospital receptionist robot with extended hybrid code network to select responses and gestures. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. 8013–8018. doi:10.1109/ICRA40945.2020.9197160

Kendrick, K. H. and Drew, P. (2016). Recruitment: Offers, Requests, and the Organization of Assistance in Interaction. *Research on Language and Social Interaction* 49, 1–19. doi:10.1080/08351813.2016.1126436. Publisher: Routledge _eprint: https://doi.org/10.1080/08351813.2016.1126436

Kontogiorgos, D., Pereira, A., Sahindal, B., van Waveren, S., and Gustafson, J. (2020a). Behavioural responses to robot conversational failures. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 53–62

Kontogiorgos, D., Tran, M., Gustafson, J., and Soleymani, M. (2021). A systematic cross-corpus analysis of human reactions to robot conversational failures. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 112–120

Kontogiorgos, D., Van Waveren, S., Wallberg, O., Pereira, A., Leite, I., and Gustafson, J. (2020b). Embodiment effects in interactions with failing robots. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14

Krauss, R. M. and Fussell, S. R. (1990). Mutual knowledge and communicative effectiveness. *Intellectual teamwork: Social and technological foundations of cooperative work* , 111–146

Kühnel, C. (2012). *Quantifying Quality Aspects of Multimodal Interactive Systems* (Springer Science & Business Media)

Kwon, M., Huang, S. H., and Dragan, A. D. (2018). Expressing robot incapability. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 87–95

Lau, J., Zimmerman, B., and Schaub, F. (2018). Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on Human-Computer Interaction* 2, 1–31

Levinson, S. C. (1983). *Pragmatics* (Cambridge, UK: Cambridge University Press)

Liza, F. F. (2019). *Improving Training of Deep Neural Network Sequence Models* (University of Kent (United Kingdom))

Luong, H.-T. and Yamagishi, J. (2020). Nautilus: a versatile voice cloning system. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28, 2967–2981

1047 Maraev, V., Bernardy, J.-P., and Howes, C. (2021). Non-humorous use of laughter in spoken dialogue
1048   systems. In *Linguistic and Cognitive Approaches to Dialog Agents (LaCATODA 2021)*. 33–44
1049 Marge, M., Espy-Wilson, C., Ward, N. G., Alwan, A., Artzi, Y., Bansal, M., et al. (2022). Spoken
1050   language interaction with robots: Recommendations for future research. *Computer Speech &*
1051   *Language* 71, 101255. doi:https://doi.org/10.1016/j.csl.2021.101255
1052 Mirnig, N., Giuliani, M., Stollnberger, G., Stadler, S., Buchner, R., and Tscheligi, M. (2015). Impact
1053   of robot actions on social signals and reaction times in hri error situations. In *Social Robotics:*
1054   *7th International Conference, ICSR 2015, Paris, France, October 26-30, 2015, Proceedings 7*
1055   (Springer), 461–471
1056 Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., and Tscheligi, M. (2017). To err
1057   is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and*
1058   *AI* , 21
1059 Moore, R. K. (2007). Spoken language processing: Piecing together the puzzle. *Speech*
1060   *communication* 49, 418–435
1061 Moore, R. K. (2017). Is spoken language all-or-nothing? implications for future speech-based human-
1062   machine interaction. *Dialogues with Social Robots: Enablements, Analyses, and Evaluation* ,
1063   281–291
1064 Moore, R. K. (2022). Whither the priors for (vocal) interactivity? *arXiv preprint arXiv:2203.08578*
1065 Nielsen, J. (1995). 10 usability heuristics for user interface design. `https://www.nngroup.`
1066   `com/articles/ten-usability-heuristics/`, last accessed 14 June 2023
1067 Ozkan, E. E., Gurion, T., Hough, J., Healey, P. G., and Jamone, L. (2022). Speaker motion patterns
1068   during self-repairs in natural dialogue. In *Companion Publication of the 2022 International*
1069   *Conference on Multimodal Interaction* (New York, NY, USA: Association for Computing
1070   Machinery), ICMI '22 Companion, 24–29. doi:10.1145/3536220.3563684
1071 Papaioannou, I., Cercas Curry, A., Part, J. L., Shalyminov, I., Xu, X., Yu, Y., et al. (2017). Alana:
1072   Social dialogue using an ensemble model and a ranker trained on user feedback. *Proc. AWS re:*
1073   *INVENT*
1074 Park, S., Healey, P. G. T., and Kaniadakis, A. (2021). Should robots blush? In *Proceedings*
1075   *of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA:
1076   Association for Computing Machinery), CHI '21. doi:10.1145/3411764.3445561
1077 Peltason, J., Riether, N., Wrede, B., and Lütkebohle, I. (2012). Talking with robots about objects: A
1078   system-level evaluation in hri. In *Proceedings of the Seventh Annual ACM/IEEE International*
1079   *Conference on Human-Robot Interaction* (New York, NY, USA: Association for Computing
1080   Machinery), HRI '12, 479–486. doi:10.1145/2157689.2157841
1081 Pezzulo, G., Donnarumma, F., Dindo, H., D'Ausilio, A., Konvalinka, I., and Castelfranchi, C. (2019).
1082   The body talks: Sensorimotor communication and its brain and kinematic signatures. *Physics of*
1083   *Life Reviews* 28, 1–21. doi:https://doi.org/10.1016/j.plrev.2018.06.014

1084 Porcheron, M., Fischer, J. E., Reeves, S., and Sharples, S. (2018). Voice interfaces in everyday life.
1085   In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (New York,
1086   NY, USA: Association for Computing Machinery), CHI '18, 1–12. doi:10.1145/3173574.3174214

1087 Purver, M. (2004). *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. thesis, King's
1088   College, University of London

1089 Purver, M., Eshghi, A., and Hough, J. (2011). Incremental semantic construction in a dialogue
1090   system. In *Proceedings of the ninth international conference on computational semantics (IWCS
1091   2011)*. 365–369

1092 Ragni, M., Rudenko, A., Kuhnert, B., and Arras, K. O. (2016). Errare humanum est: Erroneous
1093   robots in human-robot interaction. In *2016 25th IEEE International Symposium on Robot and
1094   Human Interactive Communication (RO-MAN)*. 501–506. doi:10.1109/ROMAN.2016.7745164

1095 Ram, A., Prasad, R., Khatri, C., Venkatesh, A., Gabriel, R., Liu, Q., et al. (2018). Conversational ai:
1096   The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*

1097 Salem, M., Lakatos, G., Amirabdollahian, F., and Dautenhahn, K. (2015). Would you trust a
1098   (faulty) robot? effects of error, task type and personality on human-robot cooperation and trust. In
1099   *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*.
1100   141–148

1101 Schegloff, E. A. (1992a). Repair after next turn: The last structurally provided defense of
1102   intersubjectivity in conversation. *American journal of sociology* 97, 1295–1345

1103 Schegloff, E. A. (1992b). Repair after next turn: The last structurally provided defense of
1104   intersubjectivity in conversation. *American Journal of Sociology* 97, 1295–1345

1105 Schegloff, E. A. (1997). Third turn repair. *AMSTERDAM STUDIES IN THE THEORY AND
1106   HISTORY OF LINGUISTIC SCIENCE SERIES 4* , 31–40

1107 Schegloff, E. A. (2007). *Sequence organization in interaction: Volume 1: A primer in conversation
1108   analysis* (New York: Cambridge University Press)

1109 Schegloff, E. A., Jefferson, G., and Sacks, H. (1977a). The preference for self-correction in the
1110   organization of repair in conversation. *Language* 53, 361–382

1111 Schegloff, E. A., Jefferson, G. D., and Sacks, H. (1977b). The preference for self-correction in the
1112   organization of repair in conversation. *Language* 53, 361 – 382

1113 Shriberg, E. (1994). *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. thesis, University of
1114   California, Berkeley

1115 Siwach, G., Haridas, A., and Chinni, N. (2022). Evaluating operational readiness using chaos
1116   engineering simulations on kubernetes architecture in big data. In *2022 International Conference
1117   on Smart Applications, Communications and Networking (SmartNets)*. 1–7. doi:10.1109/
1118   SmartNets55823.2022.9993998

1119 Skantze, G. (2005). Exploring human error recovery strategies: Implications for spoken dialogue
1120   systems. *Speech Communication* 45, 325–341

1121   Skantze, G. and Doğruöz, A. S. (2023). The open-domain paradox for chatbots: Common ground as
1122   the basis for human-like dialogue. *arXiv preprint arXiv:2303.11708*

1123   Stiber, M., Taylor, R. H., and Huang, C.-M. (2023). On using social signals to enable flexible error-
1124   aware hri. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot
1125   Interaction* (New York, NY, USA: Association for Computing Machinery), HRI '23, 222–230.
1126   doi:10.1145/3568162.3576990

1127   Strathearn, C. and Gkatzia, D. (2021a). Chefbot: A novel framework for the generation of
1128   commonsense-enhanced responses for task-based dialogue systems. In *Proceedings of the 14th
1129   International Conference on Natural Language Generation* (Aberdeen, Scotland, UK: Association
1130   for Computational Linguistics), 46–47

1131   Strathearn, C. and Gkatzia, D. (2021b). Task2Dial dataset: A novel dataset for commonsense-
1132   enhanced task-based dialogue grounded in documents. In *Proceedings of the 4th International
1133   Conference on Natural Language and Speech Processing (ICNLSP 2021)* (Trento, Italy:
1134   Association for Computational Linguistics), 242–251

1135   Tian, L. and Oviatt, S. (2021). A Taxonomy of Social Errors in Human-Robot Interaction. *ACM
1136   Transactions on Human-Robot Interaction (THRI)* 10, 1–32

1137   Tomasello, M. (2009). *Why we cooperate* (MIT press)

1138   Trung, P., Giuliani, M., Miksch, M., Stollnberger, G., Stadler, S., Mirnig, N., et al. (2017). Head
1139   and shoulders: Automatic error detection in human-robot interaction. In *Proceedings of the 19th
1140   ACM International Conference on Multimodal Interaction* (New York, NY, USA: Association for
1141   Computing Machinery), ICMI '17, 181–188. doi:10.1145/3136755.3136785

1142   Uchida, T., Minato, T., Koyama, T., and Ishiguro, H. (2019a). Who is responsible for a dialogue
1143   breakdown? an error recovery strategy that promotes cooperative intentions from humans by
1144   mutual attribution of responsibility in human-robot dialogues. *Frontiers in Robotics and AI* 6.
1145   doi:10.3389/frobt.2019.00029

1146   Uchida, T., Minato, T., Koyama, T., and Ishiguro, H. (2019b). Who is responsible for a dialogue
1147   breakdown? an error recovery strategy that promotes cooperative intentions from humans by
1148   mutual attribution of responsibility in human-robot dialogues. *Frontiers in Robotics and AI* 6, 29

1149   Walker, M. A., Litman, D. J., Kamm, C. A., and Abella, A. (1997). Paradise: A framework for
1150   evaluating spoken dialogue agents. In *Proceedings of the 35th Annual Meeting of the Association
1151   for Computational Linguistics and Eighth Conference of the European Chapter of the Association
1152   for Computational Linguistics* (USA: Association for Computational Linguistics), ACL '98/EACL
1153   '98, 271–280. doi:10.3115/976909.979652

1154   Washburn, A., Adeleye, A., An, T., and Riek, L. D. (2020a). Robot errors in proximate hri:
1155   How functionality framing affects perceived reliability and trust. *J. Hum.-Robot Interact.* 9.
1156   doi:10.1145/3380783

1157 Washburn, A., Adeleye, A., An, T., and Riek, L. D. (2020b). Robot errors in proximate hri: how
1158   functionality framing affects perceived reliability and trust. *ACM Transactions on Human-Robot*
1159   *Interaction (THRI)* 9, 1–21
1160 Williams, J., Fong, J., Cooper, E., and Yamagishi, J. (2021a). Exploring Disentanglement with
1161   Multilingual and Monolingual VQ-VAE. In *Proc. 11th ISCA Speech Synthesis Workshop (SSW*
1162   *11)*. 124–129. doi:10.21437/SSW.2021-22
1163 Williams, J., Pizzi, K., Das, S., and Noé, P.-G. (2022). New challenges for content privacy in speech
1164   and audio. In *Proc. 2nd ISCA Symposium on Security and Privacy in Speech Communication*.
1165   1–6. doi:10.21437/SPSC.2022-1
1166 Williams, J., Zhao, Y., Cooper, E., and Yamagishi, J. (2021b). Learning disentangled phone
1167   and speaker representations in a semi-supervised vq-vae paradigm. In *ICASSP 2021-2021*
1168   *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE),
1169   7053–7057
1170 Xu, H., Khassanov, Y., Zeng, Z., Chng, E. S., Ni, C., Ma, B., et al. (2020). Independent language
1171   modeling architecture for end-to-end asr. In *ICASSP 2020-2020 IEEE International Conference*
1172   *on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE), 7059–7063
1173 Özkan, E. E., Gurion, T., Hough, J., Healey, P. G., and Jamone, L. (2021). Specific hand motion
1174   patterns correlate to miscommunications during dyadic conversations. In *2021 IEEE International*
1175   *Conference on Development and Learning (ICDL)*. 1–6. doi:10.1109/ICDL49984.2021.9515613
1176 Özkan, E. E., Healey, P. G., Gurion, T., Hough, J., and Jamone, L. (2023). Speakers raise their
1177   hands and head during self-repairs in dyadic conversations. *IEEE Transactions on Cognitive and*
1178   *Developmental Systems* , 1–1doi:10.1109/TCDS.2023.3254808