Article

Discourse & Communication
2024, Vol. 18(6) 954–964

DOI: 10.1177/17504813241267071
journals.sagepub.com/home/dcm

Interactive probes: Towards action-level evaluation for dialogue systems

Andreas Liesenfeld
Radboud University, The Netherlands

Mark Dingemanse
Radboud University, The Netherlands

Abstract

Measures of 'humanness', 'coherence' or 'fluency' are the mainstay of dialogue system evaluation, but they don't target system capabilities and rarely offer actionable feedback. Reviewing recent work in this domain, we identify an opportunity for evaluation at the level of action sequences, rather than the more commonly targeted levels of whole conversations or single responses. We introduce *interactive probes*, an evaluation framework inspired by empirical work on social interaction that can help to systematically probe the capabilities of dialogue systems. We sketch some first probes in the domains of tellings and repair, two sequence types ubiquitous in human interaction and challenging for dialogue systems. We argue interactive probing can offer the requisite flexibility to keep up with developments in interactive language technologies and do justice to the open-endedness of action formation and ascription in interaction.

Keywords

Applied conversation analysis, conversational user interfaces, dialogue systems, usability testing

Introduction

What makes a good conversation? This question is increasingly relevant as conversational agents adopt ever larger roles in social life. There is, of course, no simple answer: quantitative measures are reductive and conversation quality depends on myriad factors. This is

Both authors contributed equally to this work.

Corresponding author:
Andreas Liesenfeld, Centre for Language Studies, Radboud University, Erasmusplein 1, 6525 HT Nijmegen, The Netherlands.
Email: andreas.liesenfeld@ru.nl

one reason the field of dialogue system evaluation has not settled on a standardized set of criteria (Finch and Choi, 2020). The challenge is compounded for open-domain conversation, where simple measures of success seem unreachable. Thus, most current evaluation methods for open-domain conversation focus on turn-level or conversation-level metrics that can be expressed in generic terms like 'fluency' or 'humanness'.

In this paper we identify opportunities for evaluation at a level between single turns and full conversations. We designate this *action-level evaluation*, using a broad notion of social action to encompass notions like intents, skills and dialogue acts and incorporating insights from work on the situated accomplishment of social action (Gilbert, 2014). While turn-level and conversation-level metrics can provide second-order inferences about conversational 'quality', few things can replace the situated and co-constructed understanding of joint action that people bring to an interaction as it unfolds turn-by-turn (Sacks, 1992; Suchman, 2007).

The main reason to look at the level of social action is straightforward: it is what people do. People do not rate single turns or judge aggregated conversational quality; they inspect talk for what it might be doing and what response it makes relevant. Additionally, as the dichotomy between task-oriented and open-domain conversation dissolves, action-level evaluation can provide firm ground for probing conversational capabilities across settings and tasks (Young et al., 2022). As conversational agents reach mass diffusion, action-level analysis helps illuminate questions of reported coherence, humanness or even sentience in a way that is informed by research (Garfinkel, 1967).

Harnessing insights from empirical work on human interaction, we formulate an approach that involves *interactive probes*: prompts and projected responses that can bring to light the extent and limits of dialogue systems in ways that turn-level or conversation-level metrics cannot. Our approach relies on reflexive participation by the evaluator, allowing them to use their own expertise (or members' methods) to probe the interactional capabilities of a system in a systematic yet open-ended way.

A survey of evaluation methods

The literature on evaluation of chatbots and conversational agents is growing rapidly (Chalamalasetti et al., 2023; Ni et al., 2022). We identified 37 papers published between 2019 and 2022 that focus entirely or substantially on evaluation techniques as part of discussing conversational systems. Building on a review of work presented at NLP conferences in 2018–2019 (Finch and Choi, 2020), we further diversified our sample, using Google Scholar to identify studies in a wider range of venues and fields. We stopped when we had more than doubled the original sample and the number of venues covered; at that point, new studies did not add evaluation terms or methods not seen before. Our goal was not exhaustive coverage but a representative snapshot of the current state of dialogue evaluation both in methods and content (details at osf.io/qwfzv). We catalogue the terms in which conversations are evaluated and examine the level of analysis at which they are applied: turn, action, or conversation (Figure 1).

Units of evaluation

The most common unit of evaluation is the *conversation* as a whole, assessed post-hoc or derived as an average of turn-level ratings. The second most common unit is at the

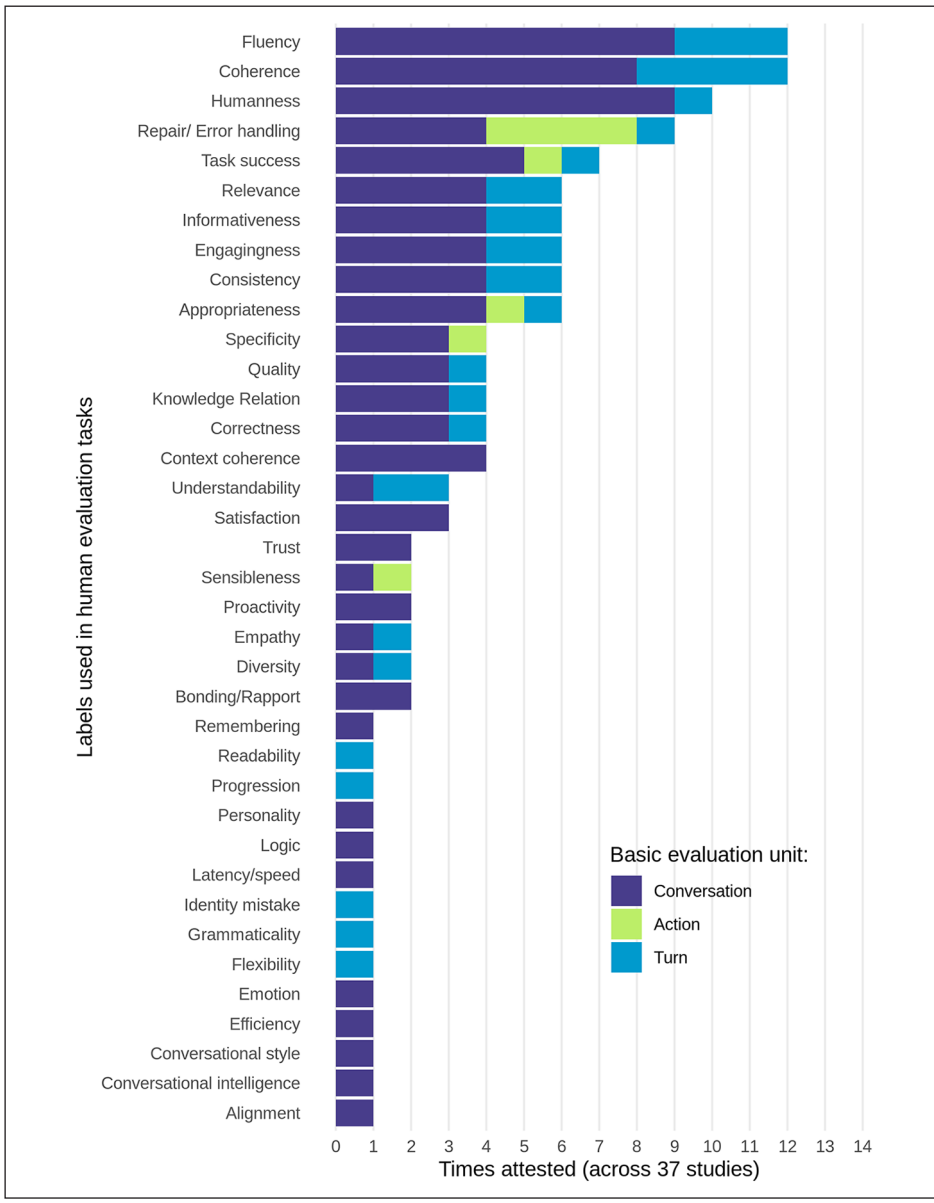

Figure 1. Metrics and labels used in recent work on human evaluation of dialogue systems. Papers often feature multiple metrics and multiple levels of evaluation, so the total number of data points exceeds 100.

level of the *turn*, often with very limited prior context. Both approaches come with advantages and disadvantages. Evaluation at conversation level is easy to carry out at scale, but it waters down the questions that can be asked, and makes it hard to pinpoint

specific pain points. Ratings at turn level may provide some insight into particular turn formats, but the contextual window may be too limited. If a dialogue system asks for clarification, does that count for or against 'coherence'? If a system produces a well-placed *uh-huh*, allowing the user to produce a multi-unit turn, should that turn be penalized as not 'engaging' or 'informative'?

Whatever the unit of analysis, most evaluations are done as *blanket assessments* in terms taken to be uniformly applicable to every single unit evaluated. The terms of evaluation themselves are rather diverse and provide a useful view of the multidimensional nature of dialogue quality (Figure 1). One broad set of terms relates mostly to properties of system responses: fluency, relevance, specificity, readability and grammaticality. Another set contains attributions made about systems on the basis of these responses: humanness, engagingness, personality and proactiveness. A third, smaller set relates to the rater's own stance: satisfaction, trust and rapport. The sheer variety of terms, many of them partly overlapping yet differing in granularity, scope and scale, has led to calls for unification and standardization (as in the 'Sensibility and Specificity Average' of the non-peer-reviewed Google Meena paper (Adiwardana et al., 2020)). But such standardization can easily lead to a sense of false security by obscuring the multidimensional nature of the problem. The full scale of the challenge is never seen in single studies; it only becomes visible in larger surveys.

Our observations on *units of evaluation* and *blanket assessments* form two interrelated challenges. A level of analysis that can only shift between the most coarsegrained (conversation) or the most fine-grained level (turn) makes it hard to formulate actionable insights. And blanket assessments make it hard to see exactly what underlies evaluators' decisions. This leaves room for a method that sits in between in terms of granularity and that focusses on multi-turn actions, or action sequences. Such a method will require a view of interaction that is attentive to patterns of conversational structures beyond single turns.

Insights from work on social interaction

Task success (relating the response of a system to a predefined result) has long been a favoured metric in task-oriented dialogue. Open-domain dialogue is often felt to lack this possibility (Deriu, 2021). But the task-oriented/open-domain distinction is increasingly blurred (Young et al., 2022) and anyway was always an artificial one: delimit some subset of interactions that seems most tractable and in the same move bracket off a vast array of communicative behaviour that seems to offer fewer footholds for evaluation. However, people always and everywhere use language in the service of specific social actions. From this perspective, task-oriented dialogue represents merely a special case of conversation in general. This opens up the way for a reframing of the challenge of evaluation. It is not that there are no tasks or social actions in open-domain dialogue. It is rather that there is an open-ended array of them. Telling a joke (Sacks, 1992) or challenging a sexist remark (Weatherall, 2015) may be different actions than scheduling a meeting or getting a ticket changed, but they are no less intricate in their social and sequential organization. Given the open-endedness of social action, action-level evaluation can be challenging, but it is also our best bet to arrive at generalizable evaluation methods.

Important prior work on action comes from ethnography, ethnomethodology and conversation analysis. While rarely directly focusing on evaluation, this work examines how people interact with dialogue systems (Porcheron et al., 2018), and has investigated core concepts such as progressivity (Fischer et al., 2019) and repair (Li et al., 2019; Raudaskoski, 1990). None of these notions afford straightforward transplantation into design and evaluation; instead, they are critical to understand the fundamentals of the social and sequential organization of action.

Such approaches can observe, for instance, that even if every next response by a dialogue system may be nominally topical and locally responsive – precisely the traits that such a system is optimized for at turn-level – it can still be oblivious to larger action structures that human participants set out to build. In other words, high turn-level ratings (and their conversation-level averages) can obscure a system's lack of ability to orient to larger action structures or language games (Chalamalasetti et al., 2023). Today's large language models also provide good evaluation targets because of their instruction-tuning (Ouyang et al., 2022): fine-tuned using human feedback to provide maximally helpful responses, the conversational patterns they emulate may attract high turn-level or conversational-level ratings, but their actual capabilities can only be probed using action-level evaluation.

Interactive probes for action-level evaluation

The goal of interactive probes is to determine to what extent a given system is able to partake in the coordination of complex social action, and to measure the degree to which it is collaborative, coherent and consistent in doing so. Any complex social action that requires this kind of close coordination between participants is a candidate domain for generating probes. Given such a domain, the key question is how to characterize its sequential unfolding.

Here we sketch how this might work for two phenomena that are both ubiquitous in interaction and challenging for dialogue systems: tellings and repair. Sometimes you need multiple turns at talk to make a point (explain something, give an example, report an event, tell a story). That is a telling. Sometimes you need to stop the conversation in its tracks to clear something up (ask for clarification, repair a misunderstanding, foreshadow a disagreement) before talk continues. That is repair.

Tellings and repair are complex social actions in that they are delivered in piecemeal, iterative fashion, so that they require close coordination from both parties over multiple turns in live interaction. Both are relevant in open-domain as well as task-oriented dialogue. For repair this may be self-evident: without it, complex interaction of any type would soon be derailed. But tellings, too, are as easily encountered in everyday conversation as in task-oriented situations (Jefferson and Lee, 1981; Stivers and Heritage, 2001). Moreover, tellings and repair can intermesh, as when a listener asks for clarification, or when clarification requires a story. Informed by what is known about tellings and repair in human interaction, we review key elements of these complex social action sequences that can lead to the formulation of interactive probes. Each structural element we single out can be probed on its own. More details are in the online materials at osf.io/qwfzv.

Tellings

Tellings are a frequent feature of conversation (Jefferson, 1978; Sacks, 1992). Structurally defined as ordered multi-turn reports of events, they are one of the main ways in which people share information, signal stance and build rapport in interaction. In the context of open-domain conversation, the main interest of tellings is their interactional delivery, which requires the active involvement of multiple participants and so presents a number of distinct interactional challenges (Jefferson, 1978). Here we highlight only three, related to preface, telling proper and closing.

Preface

Given that interaction is predominantly turnorganised, delivering something in multiple turns requires a switch to a different mode of turn-taking dynamics, one where one person is able to produce a series of turns with the other person acting as a recipient. Prospective tellers usually achieve this by producing a preface with a promise of tellability (Berger, 2017). Such a preface provides a place for recipients to align with the telling activity, for instance by giving a go-ahead. The initiation typically signals what kind of telling it is to be (e.g. a joke, a complaint). It provides recipients with resources they can use to monitor the ongoing telling to find places where responses may be relevant, and to figure out the kind of response warranted (e.g. laughter, commiseration). Not recognizing tellings means that a system can neither participate in co-producing them nor gracefully stop people from launching them.

Telling

Once a telling is under way, a next challenge is related to its delivery over multiple interactional turns. As tellings are delivered, listeners become co-narrators (Bavelas et al., 2002), carefully selecting and positioning response tokens. This includes minimal *mhmm*'s but also more complex signs of stance and affiliation demonstrating involvement. During the progression of the telling, teller and recipient monitor one another and interactively adjust the delivery while they negotiate understanding, alignment and affiliation (Stivers, 2008). Producing relevant feedback at the appropriate moments in a telling requires vigilance and attention (as anyone caught with a misplaced *mhmm* knows) and observing how a dialogue system fares at this can be a very telling type of evidence of its capabilities.

Closing

The final challenge we highlight here relates to places where transition to a next action becomes relevant. The interactional challenges here are about recognizing possible endings and knowing where to go next. Often, prefaces already foreshadow something about possible endings, giving recipients material to work with while they monitor the progression of the telling for signs of them. In free conversation, tellings seldomly come alone: they tend to occur in clusters, and one relevant next thing to do when a telling ends may be to produce a second one (Ryave, 1978).

Each of the interactional phases highlighted here provides a coordination problem that is a possible target for interactive probes. Co-producing tellings in interaction requires the reflexive participation of multiple parties in producing and recognizing elements of the telling activity, with listeners balancing recipiency and co-narration depending on story progression.

Repair

Conversational repair refers to the many ways in which people in interaction can repair, redress or redirect aspects of social action. Out of a larger possibility space of participants initiating repair and positions at which this can be done, we focus here primarily on *other-initiated self-repair*: conversational sequences in which one participant ('other') initiates repair on some prior talk, inviting the producer of that prior talk ('self') to resolve the trouble. One difference with tellings is that there is a long history of awareness of the relevance of repair to dialogue systems. This includes attempts to detect miscommunication or breakdowns (Sugiyama, 2021), empirical studies of repair initiation in human-agent interaction (Frohlich et al., 1994) and implementations of limited forms of repair in chatbots and other interactive interfaces (Ashktorab et al., 2019). This makes repair an all the more interesting target for expert evaluation by means of interactive probes.

Basic repair initiations

Given that contributions to conversation come in turns, the most natural place to find clarification about some prior turn is immediately following it (Kitzinger, 2013). Most dialogue systems can deal with simple repair initiations that invite repetition of all or part of a prior turn (e.g. English 'Huh?', 'Sorry?'). Such formats can offer a useful way to establish baseline capabilities for repetition and clarification.

Repair at longer distances

Sometimes the position immediately following a putative trouble source is not available. In such cases, people use various resources to tie repair initiation to target trouble sources. A straightforward probe could target a turn before the immediately prior one. Responding to such repair initiations often requires a form of reference resolution and dialogue history.

Pursuing a repair

Resolving trouble in conversation sometimes takes multiple attempts. A common structure for this kind of sequence is a shift towards increasing specificity in successive repair initiations (Skedsmo, 2020). This requires keeping track of the state and relative certainty of multiple pieces of information distributed across turns, so probing such cascading repairs may bring to light how conversational interfaces deal with the incremental build-up of mutual understanding.

Resumption of base sequence

Structurally speaking, a repair initiation starts a *side sequence* (Jefferson, 1972) that needs resolving before the ongoing talk is resumed. As with tellings, this is a critical moment that requires orientation to the action level of conversation: it involves transitioning out of one subsidiary action back to the base conversational sequence with its own action dynamics.

The elements singled out here are not exhaustive. Repair is always done in the service of other social actions, used for many purposes beyond simple misunderstandings, including signalling inappositeness or marking surprise (Kitzinger, 2013). Thus, responding to repair initiations is not simply a matter of mechanistically following a template. Creative uses of interactive probes can expose how capable systems are of handling the open-endedness of talk-in-interaction.

Discussion

Surveying existing work on the evaluation of dialogue systems identified a gap: evaluation generally happens at the whole conversation or turn level rather than action level. We argued for the primacy of action in both task-oriented and open-domain conversation and advance a proposal for *action-level evaluation* using *interactive probes*: open-ended participatory prompt templates that evaluate action-level capabilities in dialogue systems.

The probes we have sketched tap into two domains of social action that are ubiquitous in human interaction yet challenging for dialogue systems: tellings and repair. More broadly, interactive probes target elements of conversational patterns that are recognizable to people and that may or may not be supported in human-computer interaction. Empirical work on human interaction provides a wealth of other possible probes and domains, such as *pre-offers*, *lapses*, *extreme case formulations*, *my side tellings*, *laughter*, *responsive list constructions* and so on (Sidnell and Stivers, 2013). Importantly, action in interaction is always provisional: we are never sure quite which action we did, except through how it is taken up by others. Interactive probes should be able to rise to the occasion by doing justice to the open-endedness of action formation and ascription.

Limitations

We are aware of four limitations. First, this is not a benchmark method. Benchmarks are typically precisely delimited tasks with predefined outcomes. Action-level evaluation is different in spirit. It can deliver systematic comparative judgements, but only in the service of understanding capabilities for complex social action. It serves as a starting point for anybody interested in dialogue systems to learn about how people build action sequences together and pinpoint the precise places at which this is challenging for machines.

Second, this method is for evaluation more than design. System-agnostic, interaction-specific feedback can provide precise characterizations of interactional abilities. However, designing dialogue systems that can take part in complex social actions is a different question which requires complementary methods. The actual room for improvement is determined by the design goals and implementational details of actual systems.

Third, tension exists between providing standardized formulations (good for scalability) and respecting the creativity and ambiguity of social action (more like human interaction). While our online materials provide examples of prompts, they only stand for the larger structural elements and positions we target, just as 'once upon a time' only stands for a larger and potentially infinite array of ways to begin what might be a fairy tale.

Finally, action-level evaluation cannot replace analysis. Evaluation tends to reduce rich, ambiguous and context-dependent conduct to categorical data points. Even if interactive probes are designed to provide wiggle room in formulation and stay linked to actual records of interaction, any categorical decision about capabilities inferred from this necessarily straitjackets multi-interpretable conduct and puts us at a distance from the fine details of unfolding interaction (Birhane, 2021).

Conclusion

As generative AI grows in importance, a lack of direct control over generation makes strong and flexible methods for evaluation all the more important (Albert and Hamann, 2021; Pelikan and Hofstetter, 2022; Takanashi and Den, 2019). Action-level evaluation provides a useful complement to turn-level and conversation-level evaluation. It is experience-near, mapping directly onto the structures and actions that people use when conducting interaction. It is informed by empirical work on human interaction, allowing it to deliver actionable insights and move beyond blanket assessments and reductive metrics. It can offer solid ground even as the boundaries between task-oriented and open-domain conversation blur.

Most dialogue systems today can generate bubbly conversation starters and plausible responses to prompts. But can they partake in tellings delivered over multiple turns, recognize when they end, and produce fitting next actions? Can they deal with complex repair sequences without losing the larger action such sequences are subsidiary to? These and other things are targeted by *interactive probes*. In most current evaluation methods, judgements (whether conversation-level or turn-level) are fundamentally detached from where the action is. In contrast, interactive probes allow evaluators to probe the sequential structure of complex social actions, pinpoint trouble and characterize system capabilities in empirically grounded ways. By helping to make visible the interactional achievement of social action, they point to a future in which dialogue systems are more human-centred, with a division of labour that leans towards tools adapting to people rather than vice versa.

Acknowledgements

We thank the editors and two anonymous reviewers for feedback. Given stringent word limits that include references, we have opted to privilege citations to work by minoritised scholars; citations to classics and usual suspects missing here can be found in their work.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Adiwardana D, Luong MT, So DR, et al. (2020) Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Albert S and Hamann M (2021) Putting wake words to bed: We speak wake words with systematically varied prosody, but CUIs don't listen. In: *CUI 2021 – 3rd conference on conversational user interfaces*. pp.1–5. New York, NY: ACM Press.

Ashktorab Z, Jain M, Liao QV, et al. (2019) Resilient chatbots: Repair strategy preferences for conversational breakdowns. In: *Proceedings of the 2019 CHI conference on human factors in computing systems*. pp.1–12. New York, NY: ACM Press.

Bavelas JB, Coates L and Johnson T (2002) Listener responses as a collaborative process: The role of gaze. *Journal of Communication* 52(3): 566–580.

Berger E (2017) The interactional achievement of tellability: A study of story-openings. *Revue française de linguistique appliquee* XXII(2): 89–107.

Birhane A (2021) The impossibility of automating ambiguity. *Artificial Life* 27(1): 44–61.

Chalamalasetti K, Gotze J, Hakimov S, et al. (2023) Clembench: Using game play to evaluate chat-optimized language models as conversational agents. *arXiv preprint arXiv:2305.13455*.

Deriu J (2021) *Evaluation of dialogue systems*. PhD Thesis, University of Zurich, Zurich.

Finch SE and Choi JD (2020) Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. In: *Proceedings of the 21th annual meeting of the special interest group on discourse and dialogue*. pp.236–245. New York, NY: ACM Press.

Fischer JE, Reeves S, Porcheron M, et al. (2019) Progressivity for voice interface design. In: *Proceedings of the 1st international conference on conversational user interfaces – CUI '19*. Dublin, Ireland, 22–23 August 2019, pp.1–8. New York, NY: ACM Press.

Frohlich D, Drew P and Monk A (1994) Management of repair in human-computer interaction. *Human–Computer Interaction* 9(3–4): 385–425.

Garfinkel H (1967) Studies of the routine grounds of everyday activities. In: Garfinkel H (ed.) *Studies in Ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall, pp.35–75.

Gilbert M (2014) *Joint Commitment: How We Make the Social World*. New York, NY: Oxford University Press.

Jefferson G (1972) Side sequences. In: Sudnow DN (ed.) *Studies in Social Interaction*. New York, NY: Macmillan, pp.294–338.

Jefferson G (1978) Sequential aspects of storytelling in conversation. In: Schenkein J (ed.) *Studies in the Organization of Conversational Interaction*. New York, NY: Academic Press, pp.219–248.

Jefferson G and Lee JRE (1981) The rejection of advice: Managing the problematic convergence of a 'troubles-telling' and a 'service encounter'. *Journal of Pragmatics* 5(5): 399–422.

Kitzinger C (2013) Repair. In: Sidnell J and Stivers T (eds) *The Handbook of Conversation Analysis*. Malden, MA: Wiley, pp.229–256.

Li CH, Chen K and Chang YJ (2019) When there is no progress with a task-oriented chatbot: A conversation analysis. In: *Proceedings of the 21st international conference on human-computer interaction with mobile devices and services*. Taipei, Taiwan, pp.1–6. New York, NY: ACM Press.

Ni J, Young T, Pandelea V, et al. (2022) Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial Intelligence Review* 56(4): 3055–3155.

Ouyang L, Wu J, Jiang X, et al. (2022) Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35, 27730–27744.
Pelikan H and Hofstetter E (2022) Managing delays in human-robot interaction. *ACM Transactions on Computer-Human Interaction* 30(4): 1–42.
Porcheron M, Fischer JE, Reeves S, et al. (2018) Voice interfaces in everyday life. In: *Proceedings of the 2018 CHI conference on human factors in computing systems*. Montreal, QC, Canada. pp.1–12. New York, NY: ACM Press.
Raudaskoski P (1990) Repair work in human-computer interaction: A conversation analytic perspective. In: Luff P, Gilbert N and Frohlich D (eds) *Computers and Conversation*. London: Academic Press. pp.151–171.
Ryave AL (1978) On the achievement of a series of stories. In: Schenkein J (ed.) *Studies in the Organization of Conversational Interaction*. Language, Thought, and Culture. New York, NY: Academic Press. pp.113–132.
Sacks H (1992) *Lectures on Conversation*. London: Blackwell.
Sidnell J and Stivers T (eds) (2013) *The Handbook of Conversation Analysis*. Malden, MA: Wiley.
Skedsmo K (2020) Multiple other-initiations of repair in Norwegian sign language. *Open Linguistics* 6(1): 532–566.
Stivers T (2008) Stance, alignment, and affiliation during storytelling: When nodding is a token of affiliation. *Research on Language and Social Interaction* 41(1): 31–57.
Stivers T and Heritage J (2001) Breaking the sequential mould: Narrative and other methods of answering "more than the question" during medical history taking. *Text* 21(1): 151–185.
Suchman LA (2007) *Human-Machine Reconfigurations: Plans and Situated Actions*. 2nd edn. Cambridge: Cambridge University Press.
Sugiyama H (2021) Dialogue breakdown detection using BERT with traditional dialogue features. In: Marchi E, Siniscalchi SM, Cumani S, et al. (eds) *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*. Singapore: Springer, pp.419–427.
Takanashi K and Den Y (2019) Field interaction analysis: A second-person viewpoint approach to Maai. *New Generation Computing* 37(3): 263–283.
Weatherall A (2015) Sexism in language and talk-in-interaction. *Journal of Language and Social Psychology* 34(4): 410– 426.
Young T, Xing F, Pandelea V, et al. (2022) Fusing task-oriented and open-domain dialogues in conversational agents. *Proceedings of the AAAI Conference on Artificial Intelligence* 36(10): 11622–11629.

Author biographies

Andreas Liesenfeld is Assistant Professor at the Centre for Language Studies at Radboud University. He has a background in computational linguistics (Nanyang Technological University, Singapore) and expertise in data science and conversational AI. His work focuses on technology assessment and has appeared in venues like ACL, INTERSPEECH, LREC and SIGDIAL.

Mark Dingemanse is Associate Professor in Language and Communication and head of the Elementary Particles of Conversation research project. His work on interactive repair, interjections and conversational structure has revealed a number of candidate pragmatic universals that his team studies using qualitative, cross-linguistic and computational approaches.