Theories of Conversation for Conversational IR

PAUL THOMAS, Microsoft, Australia
MARY CZERWINKSI, DANIEL MCDUFF, and NICK CRASWELL, Microsoft, USA

Conversational information retrieval is a relatively new and fast-developing research area, but conversation itself has been well studied for decades. Researchers have analysed linguistic phenomena such as structure and semantics but also paralinguistic features such as tone, body language, and even the physiological states of interlocutors. We tend to treat computers as social agents—especially if they have some humanlike features in their design—and so work from human-to-human conversation is highly relevant to how we think about the design of human-to-computer applications. In this article, we summarise some salient past work, focusing on social norms; structures; and affect, prosody, and style. We examine social communication theories briefly as a review to see what we have learned about how humans interact with each other and how that might pertain to agents and robots. We also discuss some implications for research and design of conversational IR systems.

CCS Concepts: • **Information systems** → **Users and interactive retrieval**; *Search interfaces*; • **Human-centered computing** → *Natural language interfaces*; *HCI theory, concepts and models;*

Additional Key Words and Phrases: Conversational information retrieval, embodied conversational agent

ACM Reference format:
Paul Thomas, Mary Czerwinksi, Daniel McDuff, and Nick Craswell. 2021. Theories of Conversation for Conversational IR. *ACM Trans. Inf. Syst.* 39, 4, Article 39 (August 2021), 23 pages.
https://doi.org/10.1145/3439869

1 CONVERSATIONS: WITH HUMANS AND WITH MACHINES

"One of the most human things that human beings do is talk to one another." This observation by Labov and Fanshel [1977] is not controversial—indeed it is almost trivial—but it has important consequences for the design and evaluation of conversation-based **information retrieval (IR)** software. People treat their computers as social actors [Nass et al. 1994], and there are rules and conventions by which conversation naturally progresses between actors. We should therefore understand these rules to build more natural, pleasant, and efficient conversational software.

1.1 In Information Retrieval

Conversational user interfaces support human–machine interactions including iteration, refinement, clarification, and grounding; these interactions are typically in natural language and

Authors' addresses: P. Thomas, Microsoft, 6 National Circuit, Barton, ACT 2600, Australia; email: pathom@microsoft.com; M. Czerwinksi, Microsoft, Microsoft Building 99, 14820 NE 36th Street, Redmond, Washington, 98052 USA; email: marycz@microsoft.com; D. McDuff, Microsoft, Microsoft Building 99, 14820 NE 36th Street, Redmond, Washington, 98052 USA; email: damcduff@microsoft.com; N. Craswell, Microsoft, 555 110th Ave NE, Bellevue WA 98004 USA; email: nickcr@microsoft.com.

1046-8188/2021/08-ART39 $15.00
https://doi.org/10.1145/3439869

increasingly are via speech. *Conversational IR* builds on these interfaces to support information seeking, where the human recognises a need for information and the agent's job is to provide it. These ideas are seen in search systems (where some statement of need, or query, is explicit) and in recommenders (where there is no such explicit statement). They are commercialised in systems such as Microsoft's Cortana or Apple's Siri and even to a limited extent on web search engines via suggested queries, refinements, and similar elements.

Advances in natural language processing and in retrieval models have encouraged research in conversational IR. Recent research has considered, most prominently[1]:

Generating and/or ranking conversational responses, including tracking an evolving context and shifting topics. This has seen a great deal of research, including shared exercises in the **Text REtrieval Conference (TREC)**.[2] Approaches to date have either involved retrieving and ranking utterances from a large corpus (e.g., Ji et al. [2014] and Yang et al. [2020]) or generating utterances based on reference text [Gao et al. 2019]. Work has also focussed on understanding topic shifts and knowing what past context is still useful (e.g., Voskarides et al. [2020]), as well as maintaining consistency and coherence [Huang et al. 2020].

Extending systems to mixed initiative, or allowing agents to initiate conversations and allowing agents ask questions of people as well as the other way around (e.g., Aliannejadi et al. [2019] and Rosset et al. [2020]). This extends the concept of "conversation" in IR but does make assumptions about how conversations are, or should be, structured.

Evaluation methodologies for conversation, for example deciding the correct unit of evaluation [Kiseleva et al. 2016] and other methodological questions [Simpson 2020] as well as building instruments for particular sub-tasks such as those listed above.

For the most part, however, research in IR has focussed on the mechanics of a conversation: understanding the questions being asked and generating responses. It has not considered other conversational phenomena well known from other disciplines and from studying natural (human-to-human) exchanges.

1.2 In Other Fields

When people interact with machines, they carry over many of the same expectations, norms, biases, and behaviours as when interacting with humans [Breazeal 2002; Nass and Moon 2000; Reeves 2010; Reeves and Nass 1996]. This is true even when the machines in question have very little in the way of natural language understanding, speech, or other capabilities we associate with humans. It is more true of machines that can "converse" in something like natural language, and even more so if there is a "face" or social component to the interaction. We apply at least some social rules, and conversational norms, even when we are fully aware that we are talking with a computer [Porcheron et al. 2018]. These reactions seem innate and automatic and are difficult to "cure": Reeves and Nass describe our interactions as "*fundamentally social and natural*, just like interactions in real life … everyone expects media to obey a wide range of social and natural rules" [Reeves and Nass 1996, emphasis in original].

For example, research has demonstrated that people are polite to computers, despite being well aware that computers cannot be offended; we prefer computers with "personalities" closer to our own; we also apply human stereotypes, such as those around gender, and biases such as preferring attractive "faces"; and we react to "vulnerabilities" [Hamacher et al. 2016; Nass and Moon 2000;

[1] This is a very brief summary of an active field of research. For a larger survey, see for example references in Gao et al. [2019].

[2] See, e.g., http://ir.cis.udel.edu/sessions/ (TREC session track) and http://www.treccast.ai/ (TREC conversational assistance track).

Nass et al. 1997; Reeves and Nass 1996; Traeger et al. 2020; Yuksel et al. 2016]. Conversational software that acts more like people, and in particular that recognises more of the conversational context [Aneja et al. 2019; Hoegen et al. 2019], is perceived as more trustworthy [Bickmore and Cassell 2001; Hamacher et al. 2016], as well as more engaging [Cassell and Thorisson 1999]. It is also regarded as more intelligent [Shamekhi et al. 2016], and in our own work we have seen suggestions that it is forgiven more when it makes mistakes.

It seems prudent that designers of conversational IR systems should consider these insights. Of course, there are many social phenomena that might also be important: so what do we need to take into account, or perhaps prioritise, if we are building a conversational IR system? How should we decide what is important to build or what key directions to study?

A straightforward approach is suggested by Reichman, writing about information-seeking conversation: "*We shall begin by looking at person-person communication to understand the problem we're dealing with.* Later … only after formulating rules of discourse engagement for people, we shall describe a computer module…" [Reichman 1985, emphasis ours]. A similar approach has been used by Brooks and Belkin [1983] and by Daniels et al. [1985]. In other words, to design a conversational IR system, we could start by understanding how *people* converse; then we can work to understand what carries over to conversations with software agents, and what we need to know as experimenters or as system designers.

1.3 This Paper

The academic study of conversation, the way context is established and shared, and the mechanisms by which conversations are structured, goes back over 40 years to early work by Sacks et al. [1974] and others.[3] In this article, we survey key work from conversation analysis, as well as linguistics, philosophy, and social psychology, in five broad areas: relevant theories of social conversation (Section 2), basic conventions associated therewith (Section 3), the moves and structures in conversations (Section 4), para-linguistic aspects such as prosody and affect (Section 5), and the desirability or otherwise of accurate simulations (Section 6). We attempt to summarise findings or phenomena relevant to conversational IR systems, and finish with suggestions for further innovative research or engineering.

It is important to emphasise that we are not suggesting that there is a need, or even a desire, to precisely mimic an individual person, or even an average person. Rather we believe that the study of human–human interactions can help us understand what might be pertinent in the design of machines. Perhaps the answer is, sometimes, to do nothing or avoid humanlikeness (for example, if that helps evade the pitfalls in the uncanny valley) or to make it clear that the agent is software, not flesh and blood. This is part of a larger conversation that has been sparked by the rapid advancements and "productisation" of artificial intelligence in recent years. Researchers have been increasing their efforts to align the decades of research on human–computer interaction with the design of AI systems—including conversational ones. Design guidelines propose that AI systems should should not "pretend" to be human but should still follow social norms [Amershi et al. 2019],[4] and we agree.

2 SOCIAL COMMUNICATION THEORIES

We start from the large body of work on how humans communicate with each other and posit that it is useless to add natural speech user interfaces to agents and robots for search tasks if the agents are incapable of achieving "common ground" with the searcher. This would require the agents to have an at least rudimentary understanding of concepts that are taken for granted when humans

[3] Ten Have [2007] gives a useful overview of objects and methods.

[4] https://www.microsoft.com/en-us/research/project/guidelines-for-human-ai-interaction/.

make inquiries of one another (e.g., Wilkes-Gibbs and Clark [1992]). There have been quite a few theories that underline this principle, but we will only touch on a few that we think are relevant to developing search interactions with intelligent, conversational agents.

Krämer et al. [2012] provide a thorough summary of why social communication theories are so important not only to the design of systems that communicate with humans but also for our methods for evaluating these types of intelligent systems. Standard usability practices were sufficient before systems became infused with intelligence. However, now humans have to decipher "what the system means" during intelligent communication and interaction, just as with other humans, and this may be more nuanced because of social cues. Krauss and Fussell [1991] have argued that interlocutors must tailor their message for the recipient's assumed knowledge. This means that both the human and the intelligent agent need some fundamental understanding of what each other knows and what actions each can take. In the case of the agent or robot, the system should be able to take the human's perspective, which might avoid communication breakdowns. If the agent has any social affordances at all, then the human will most likely naturally do this anyway, as many studies have shown (e.g., see references in Krämer et al. [2012]).

Humans do not just use the words that they hear or read to hypothesise what their interlocutor means, knows or feels—they have a "model" of the other person's needs, knowledge, and their ability to understand: humans have a "theory of mind" [Baron-Cohen 1997]. Of course, today's agents do not have even a basic understanding of their users' knowledge, abilities, beliefs, or emotions, much less pertinent contextual information. All of this is required to achieve common ground and efficient perspective taking. It seems obvious that we would want to imbue our intelligent agents with as much background knowledge about a user, or a user group, as possible, to improve successful and trustworthy conversations.

Common Ground theory [Wilkes-Gibbs and Clark 1992] provides grounding principles that can help with the design of conversational agents. The heuristic of linguistic co-presence means that anything spoken throughout the conversation should be known to both interlocutors. There is also the principle of closure: interlocutors like evidence that whatever action they have performed, they have achieved (e.g., confirmatory feedback from an agent after a request). Of course the need for feedback is fundamental in all aspects of human–computer interaction, so this is nothing new when it comes to evaluating conversational systems.

3 BASIC CONVENTIONS FOR CONVERSATION

A number of basic conventions have been proposed for (human-to-human) conversation, most prominently the "cooperative principle" and notions of politeness.

3.1 Cooperation and Grice's Maxims

In a discussion of how people can unambiguously imply things that are unsaid, Grice notes that there is a structure to natural conversation: "our talk exchanges do not normally consist of a succession of disconnected remarks… at each stage, SOME possible conversational moves would be excluded as conversationally unsuitable" [Grice 1975, p. 45; emphasis in original].[5] His examination of these regularities led to the "Cooperative Principle," from which in turn he derives further principles for conversation. These "maxims" are in four categories: quantity (make your contribution as informative as required but no more so), quality (make your contribution true), relation ("be relevant"), and manner (avoid obscurity, ambiguity, prolixity; be orderly). Following these maxims,

[5] Labov and Fanshel [1977] have a similar observation: "if almost anything can be said at any time, then the number of choices which are open to the speaker would create a bewildering complexity." Yet most of the time, we can converse without bewilderment.

he suggests, leads to conversation that is natural and easy to follow; violations lead to deception and rudeness or can serve as a signal to the listener (for example, by choosing not to be completely informative we can signal our dislike or disapproval).

It is easy to imagine software violating Grice's maxims. For example, software prompts or messages are often unclear (violating "manner") or have either too much or too little detail (violating "quantity"). Reeves and Nass [1996] give examples in other media. We can assume that these violations lead to a poor experience. From this, Gnewuch et al. [2017] derive design rules for conversational agents, but these have not been tested. We are aware of very little work that explicitly addresses the maxims in search systems.

3.2 Politeness

Grice notes that his maxims exclude norms "aesthetic, social, or moral in character… that are also typically observed by participants." However, these aesthetic and social norms are important. In response, Leech suggests a "politeness principle," a "necessary complement" to Grice's work [Leech 1983, p80], and adds the maxims of tact, generosity, approbation, modesty, agreement, and sympathy. Brown and Levinson [1987] independently derive similar behavioural norms from two related notions of "face" and rules for managing this (but see, e.g., Watts [2003] for a critique). A close reading of virtual reference desks by Radford et al. [2011] suggests similar rules apply in computer-mediated, information-seeking conversations. Norms are not limited to the content of conversations but also to many non-verbal behaviours: These are often referred to as "display rules" and vary by context and culture [Girard and McDuff 2017; Rychlowska et al. 2015].

Again, we have evidence that politeness is important even when dealing with machines [Reeves and Nass 1996], and some suggestions of what "politeness" means. It would be worthwhile to formalise these in such a way as an agent could use them and to verify their effect.

4 MOVES AND STRUCTURES

Even allowing that conversation should be cooperative and polite, a speaker has very many options (possible moves) for each turn or utterance. As Schegloff and Sacks [1973] note, "…there are considerations relevant for conversationalists in ordering and distributing their talk [about various subjects] in a single conversation." Considerable work has tried to classify these moves, their sequences, and the rules for ordering.

4.1 Moves

Reichman [1985] extends Grice's work by deriving and discussing a grammar of "conversational moves," utterances that begin communicative acts and that serve a defined role in structuring discourse—for example, presenting a claim, giving support, or shifting topic. These moves are instances of abstract types, each of which may require certain key information (the claim being made, the authority for support).

Reichman notes that "at particular stages of discourse, some conversational moves are 'expected' and 'most appropriate'" (Reichman [1985], p. 29), suggesting that there are constraints on what should be produced when; she goes on to give a formal grammar specifying when each type of move can appear in conversation. Violations of this grammar could be assumed to be poor form.

There are many other classifications, often presented as annotation schemes for conversation or dialogue. One of the most widely used schemes amongst computer scientists is DAMSL, from the Discourse Resource Initiative [Core and Allen 1997]. This is domain independent and does not focus on information-seeking conversation; it is also complex, with around 50 labels plus "diacritics" available for each utterance. A set this size complicates both labelling (manual or automatic) and deriving low-level structures, but a smaller set from Stolkce et al. [2000] includes "only" 42 labels, all of which are mutually exclusive. Stolcke et al. also report good accuracy from an automatic classifier, which would be necessary to use this at scale.

The VERBMOBIL-2 project has used 33 labels, drawn from a well-documented scheme with comprehensive instructions [Alexandersson et al. 1997]. Classifiers exist for earlier versions of the scheme, but the annotations cover a limited and rather unusual domain: negotiating meeting times. Other labelling schemes are those from the Meeting Recorder project [Dhillon et al. 2004], Bunt's DIT++ [Bunt 2010], the COR scheme used by Belkin et al. [1995], and a scheme developed by Batliner et al. [2003] aimed specifically at detecting communication breakdown. Jiang et al. [2015] have developed a labelling scheme for conversations with software agents but focused on tasks with Cortana and based on observations of Cortana's current capability. More recently, Radlinski and Craswell [2017] enumerated possible utterances for a recommender or filtering system operating over a known domain; this was extended to 22 utterance types by Azzopardi et al. [2018a]. Neither of these schemes were grounded in observation, however. Trippas et al. [2020] used two databases of human-to-human information-seeking conversations to develop yet another annotation schema, but this is relatively new and has not yet been applied to other transcripts. Also in information-seeking, earlier work from Saracevic, Spink, Su, and colleagues used a set of eight categories to code conversations between users of an academic library and professional intermediaries—focusing latterly on elicitations from users and intermediaries [Saracevic et al. 1997; Spink and Saracevic 1997]. To our knowledge, this schema has not been used in more recent work.

Lee [2020], following Austin and Searle, lists five categories of speech acts including imperatives (asking that something be done) and declaratives (describing the world). These are obviously very broad categories, but even with only these five Lee is able to demonstrate large gaps in present conversational technology.

Earlier work from other fields has developed further alternatives. For example, Bales's Interaction Process Analysis [Bales 1950] uses 12 actions (shows solidarity, shows tension release, agrees; gives suggestion, gives opinion, gives orientation, asks for orientation, asks for opinion, asks for suggestion, disagrees, shows tension, and shows antagonism). This has been influential, and the coding scheme is "highly reliable" [Labov and Fanshel 1977], so this may be useful for our purposes although the labels are very broad.

The choice of annotation scheme for conversational IR is by no means settled. A decision must depend on at least two factors: the scheme's relevance for information-seeking, open-domain, conversations and the insight the scheme offers. If annotations are needed at scale, for example for evaluation, then we must also consider the prospects of building automatic classifiers for each human utterance. (A classifier might also be needed for the agent's utterances, if this is not part of discourse planning or if the agent may be prone to error.) This is feasible. Given five action types, Jiang et al. [2015] report on a classifier with F-score 0.9; Stolkce et al. [2000] report accuracies above 70% with a much larger set of 42 types.

4.2 Sequences

Given an alphabet of moves, using any of the schemes above, we can observe that some orderings or sequences are more likely or more "correct" than others: "[t]he fact that some elements or orderings are not regarded as appropriate in discourse suggests assumptions or expectations we have as to what is appropriate" (Reichman [1985], p. 7). Work over many decades has enumerated and explained these structures and the mechanisms by which they are coordinated [Sacks et al. 1974; Schegloff 1968; see also references in Brooks and Belkin 1983].

Pairs. From conversational analysis, we borrow the simplest kind of structure, an adjacency pair. This is a pair of turns, one per participant, where the type of the first turn constrains that of the second (or, the first provokes the second) [Schegloff and Sacks 1973]. For example, adjacency pairs include

- greeting/greeting;
- question/answer; and
- complaint/remedy or complaint/excuse.

What sorts of adjacency pairs might we expect in an information-seeking conversation with a software agent? Drawing on earlier work and on the conventions of existing software, we could imagine utterance types such as request-action, request-facet, provide-facet, offer-alternatives, or confirm-choice. Adjacency pairs might then include request-facet/provide-facet ("how much does it cost?" and "$100") or offer-alternative/confirm ("how about Chinese?" and "okay"). We would not expect to see, e.g., request-facet/confirm ("how much does it cost?" and "okay").[6]

Dialogue goals and structure. At around the same time as Reichman, Daniels et al. were considering methods for an IR system to deal with "*non-specific* enquiries in a *natural* manner" (their emphasis), and for a system to cooperate with the user to find information [Brooks and Belkin 1983; Daniels et al. 1985]. Key to their method was examining the role of human intermediaries, and closely considering the structure of user–intermediary dialogue.

Drawing on transcripts of naturally occurring exchanges, Daniels et al. [1985] identified a hierarchy of goals including 23 sub-goals such as "select the databases to be searched" and "literal display of some aspect of the system" that supported eight higher-order goals such as "problem description" or "explain." Both levels exhibited "particular patterns of sequencing," and Daniels et al. could deduce a transition diagram somewhat similar to, although less detailed than, Reichman's.

Later work by Belkin et al. [1995] drew on the Conversational Roles Model [Sitter and Stein 1992] to further describe the structure of dialogues between searchers and intermediaries. Belkin et al. identify 15 general information search strategies, such as browsing for an unspecified item or learning about the system's capabilities, and suggest that each might be served by a different prototypical dialogue with different patterns of exchange. These stereotypical patterns, they suggest, should be determined from case studies of real conversations. Like the earlier work of Daniels et al. and Reichman, this provides schemas for "good" conversation that, although meant for dialogue management, could also be used for evaluation.

Carefully considering the sequences and structures in information-seeking conversation—at the level of simple pairs or higher-level constructs—should be useful for the design of conversational IR agents in at least three ways. First, knowing what a searcher might do next can inform how we interpret their utterance; second, knowing what the structures suggest can inform the agent's response; and, third, knowing what is normal can inform any evaluation.

5 AFFECT, STYLE, AND ALIGNMENT

Non-verbal behaviour and affective expressions have been studied extensively in human–human and human–agent interactions. Non-verbal behaviours include vocal cues such as prosody (tone, stress) and visual cues (for example, facial expression or head gestures). It is difficult to quantify how important specific channels or modalities are, and of course this will vary wildly by context. Having acknowledged this, researchers have found that gestures and prosody can carry as much

[6] Hennoste et al. [2005] do note some common nested structures in natural information-exchange conversations, including for example question-offering-answer/agreement occurring inside question/answer and request/grant, but these could likely still be treated as pairs.

emotional content, or sometimes significantly more, than words do (i.e., you can often judge emotion in a video quite well without the words) and that words are often not necessary to get the "gist" of what people are feeling [Scherer and Ceschi 2000]. This then raises the question of how to create an agent that understands and possibly produces these cues.

5.1 Affect and Emotion

The benefits of understanding and producing non-verbal cues are many. Back-channelling and mimicry of non-verbal cues are associated with increased rapport, liking, and affiliation [Hatfield et al. 1992; Lakin et al. 2003]. There is evidence that if a human is more expressive in a channel (e.g., visual modality) that is not captured by an interlocutor, then they will be judged as less effective at communication [McDuff et al. 2017]. Thus, if an artificial system fails to code data from that modality, and consider it in its representation of the state, then its performance might suffer.

While non-verbal behaviours can predict affective states such as frustration [Ang et al. 2002], they are not always reliable indicators of a single emotion. For example, facial expressions have a lot of variability. While there is modest consistency in how expressions are interpreted by people, how people express emotions also varies considerably in different cultures, social contexts, and even amongst individuals in the same situation [Barrett et al. 2019]. Interpreting vocal cues [Ang et al. 2002] is subject to similar challenges; for example, a wizard-of-Oz experiment by Batliner et al. [2003] mimicked a failing system, to provoke responses, but found prosodic signals were unreliable. This is possibly because many other studies used acted sequences that do not generalise well to more spontaneous and naturalistic interactions, or perhaps because users of Batliner et al.'s system believed they were interacting with a machine and judged it pointless displaying their emotion. This is not to say that non-verbal cues are not important—anyone who has had a sarcastic comment in an email misinterpreted would beg to differ—but rather that they are complex, multimodal and contextual.

It is firmly established that embodied systems have certain advantages over non-embodied systems. One example is that an agent that has a physical presence means that the user can look at it, and this requires less navigation and searching than traditional user interfaces. However, do the same benefits apply in the more specific context of conversational search? Are there different benefits?

5.2 Style and Alignment

Variation in prosody, as well as in word choice and other aspects, together make up someone's *conversational style*. Tannen defines style as "…the use of specific linguistic devices, chosen by reference to broad operating principles or conversational strategies. The use of these devices is habitual and may be more or less automatic" [Tannen 2005, p.188]. This is the "how" of a conversation, as opposed to the "what," since we can provide the same information in many ways [Berg 2014].

Tannen [1987] analysed tape recordings of dinner-party conversation amongst friends. On the basis of features such as "machine-gun questions," displays of enthusiasm, types and frequency of anecdote, and rate of speech, she identifies a distinction between "considerate" and "involved" styles amongst the guests. These unwritten rules put speakers into two camps or styles. While both "styles" aim to build rapport and no one style is better than another, they do so by emphasising different "rules" of conversation, different aspects of face [Brown and Levinson 1987], and different strategies for presentation [Lakoff 1979].

The "high involvement" style emphasises interpersonal involvement and interest, and is characterised by speaking fast, overlapping with your partner, choosing personal topics, and demonstrating enthusiasm. Tannen's "high consideration" prefers independence, and consideration of other speakers, for example by allowing longer pauses, fewer paralingusitic effects, and not imposing topics. From her analysis, Tannen suggests that partners with different styles have more trouble communicating; for example, a high-consideration speaker may find a high-involvement partner overly loud or personal, while a high-involvement speaker may find a high-consideration partner reticent and uninvolved.

While "style" has been studied in various forms in natural, casual, informal conversations, studying them in goal-directed settings has received less attention. However, in recent work these aspects of "style" have been observed in information-seeking conversations between people, and there is some evidence that in this scenario people work to match styles. Thomas et al. [2018] noted that differences in styles lead to less-satisfying conversation. Since they chose variables that can, in principle, be tracked in real time it would be interesting to know whether we see the same phenomena talking to an agent; and whether agents can be programmed to match a person's style. We are experimenting with this at present.

Other stylistic effects are well attested in the literature. For example, agents with a faster speaking rate, more variation in pitch and volume, and fewer hesitations are perceived as more truthful than agents with slower, more uniform, or more hesitant speech even when these agents present exactly the same information [Dubiel et al. 2020]. Such agents are also seen as more involved and even more attractive. While many systems provide a variety of "voices," to our knowledge the voices are not controlled for these factors, and it would be worthwhile to consider the effect of voice style more carefully.

Alignment. When people converse, they tend to *align*: That is, where there is a choice they tend to converge on the same prosody, syntax, or individual words. This is well documented in human-to-human conversation [Brennan 1996; Fusaroli and Tylén 2015] and there is evidence of a similar effect in human-to-computer conversation, as well, including effects on perceived empathy, personalisation, success, and efficiency [Bergmann et al. 2015; Branigan et al. 2010; Brennan 1996; Koulouri et al. 2016; Kühne et al. 2013; Pickering and Garrod 2004; Zepf et al. 2020]. In multiple studies of language usage researchers have also observed increased linguistic style matching between humans [Ireland et al. 2011; Niederhoffer and Pennebaker 2002]. However, there are differing results regarding how matching then impacts other aspects of the conversation, for example the self-report rating of quality or interest in the other person. Research further suggests that alignment goes beyond simply linguistic features but rather includes non-verbal behaviours and possibly even physiological parameters. There is some evidence that people that have synchronised physiological states (e.g., heart rate and respiration) report greater satisfaction [Jun et al. 2019; Woolley et al. 2010]. Would embodied avatars that simulate some of these more subtle signals and also synchronise with humans lead to similar positive outcomes?

6 HUMANLIKE SIMULATIONS, CARTOONS, AND THE "UNCANNY VALLEY"

Much of the discussion above has assumed that researchers and designers should aim for the most humanlike agents possible. This seems self-evident, given the advantages of "natural" interactions, but we must acknowledge some risks.

6.1 The Uncanny Valley

Mori [2012] notes that we feel some affinity for designed objects, such as toys or humanlike robots; and of course we feel great affinity for other people. In between, however—for example, for prosthetic limbs—"we experience an eerie sensation." In Mori's graphical representation of this "uncanny valley" there are several points along the curve that highlight how specific design choices can lead to an agent becoming repulsive. Agents that exhibit physical motion have dramatic transitions in familiarity, probably because they have more dimensions of control. This can help create a more expressive appearance, but can equally cause problems if these changes are badly timed or if some are absent altogether. People can easily spot when behaviours have unnatural intensities or dynamics [Mäkäräinen et al. 2014] or when there is a mismatch between artificial and human features [Kätsyri et al. 2015; Mitchell et al. 2011]. The types of signals that are missing from such agents might be quite subtle. For example, many embodied agents can talk and move; however, the timing and cross-modal synchronisation of these behaviours might be unnatural. As systems begin to have more humanlike dimensions, small discrepancies become apparent, and at this point people can start to find them repulsive. Few embodied agents exhibit realistic physiological changes, including respiration, pupil dilation, or blood flow [Seymour et al. 2019]. Zombies and corpses fall at the very lowest point, this is sometimes known as the "death mask effect." What distinguishes an embodied agent as a healthy person versus a zombie, both of which can move, is the appearance that it is fully alive and modelling physiological processes is most likely a part of this. These challenges present a daunting proposition and raise the question of whether it is the right decision to peruse hyper-realistic avatars.

However, a similar "uncanny valley" effect is possible even with simpler agents. For example, consider a conversation with a computer agent that tries to get to know and befriend you; but where it is clear that the agent is not really understanding what is going on, or where it behaves inappropriately. This may well seem uncomfortable, and would have negative consequences for building trust and engagement.

The transition from one side of the uncanny valley to the other is therefore fraught with danger, and the benefits might be unclear on the other side, so it is quite easy to argue that it is not worth the effort. Mori recommends that designers aim for the first peak, "a moderate degree of human likeness and a considerable sense of affinity." We may observe, however, that on the whole conversational agents are probably still some way off "a moderate degree of human likeness." If affinity is a worthwhile goal, then we still need to understand what is missing from our current models.

6.2 Cartoons vs. Realism

Even if we are designing for affinity, trust, or other social responses, of course we do not need agents to appear identical to humans. As well as the observations of Nass, Reeves, Moon, and colleagues [Nass and Moon 2000; Nass et al. 1994; Reeves 2010; Reeves and Nass 1996] this is well illustrated by Breazeal [2002]. Her "Kismet" robot is cartoonish, in that it does not closely resemble a human, but it has a sophisticated model of emotion and affect, as well as expressive output with 21 degrees of freedom in the head and face. Across several experiments, naïve users demonstrated "affective mirroring" in that they came to reflect the robot's displays of emotion. This close synchronisation of behaviours between robot and human, suggests Breazeal, is "critical to establish a natural flow and rhythm to the humanrobot interaction."

In the same vein, it is reasonable that one goal of conversational IR research could be a cartoon version of conversation: an agent that captures some critical aspects of natural, human-to-human, conversation without chasing unimportant or distracting details. For example, a cartoonish agent might forget everything at the end of each conversation, by design. This is not humanlike, but might nonetheless but might be preferred to an agent that remembers everything: Either because users are wary of "filter bubbles" or because they do not like the idea of an agent keeping tabs on all they do. Again, this argues neither for nor against humanlike interactions: rather, for a better understanding of the design options and tradeoffs.

6.3 Emulations vs. Applications

Shneiderman [2020] draws a useful distinction between what he sees as two possible goals of AI (and we may consider conversational IR systems as a form of AI for his purposes). The *emulation* goal, on his reading, aims to understand human cognition and mimic it; emulated systems may be self-directed and self-monitoring.[7] The *application* goal, by contrast, is to develop useful products by applying AI methods but is characterised by producing tools that are clearly under human control and usually are not anthropomorphised.

Recognising that people often prefer explainable, predictable, and controllable machines, Shneidernman warns against the view that human–human interaction is a good model for human–robot interaction, citing "repeated missteps" in past projects. An overly humanlike agent may lead to the real human losing control, which is paramount; we might add the risk of a humanlike system being trusted more than it should.

However, there are arguments for emulation. For instance, the research presented above indicated that we can increase trust, perceived intelligence and likeability through style matching humans' individual speech and gestures. We do in fact build relationships with software artefacts and apply customary social rules; this may even be necessary "to guarantee meaningful interactions" [Krämer et al. 2012]. In particular, in our context we must also acknowledge the recent success of the emulation programme in voice and natural-language interfaces [Shneiderman 2020].

Some likeness to humans improves communications; some missteps are costly. As conversational IR systems mature, we should acknowledge the benefits of emulation and of more humanlike interaction, but also the risks. We must be honest with users about the systems we are building, but in our view the research question is not "should we emulate?" but a more nuanced "what exactly is good to emulate?" Enumerating the aspects of human–human interaction that we may want to model in software (understanding those aspects and carefully considering the effect of that emulation) will continue to be useful for conversational system design and evaluation.

7 IMPLICATIONS FOR CONVERSATIONAL IR

We do not suggest that the survey above is complete. We hoped to introduce theories, concepts, and empirical evidence that starts to paint a picture for how to design novel, conversational agents that can better partner with humans during conversational search (and other) tasks in a way that is both more natural but also more effective and pleasing. There is much more work to be done, but we hope it is useful and that it gives a flavour of the breadth and depth of work we could draw on. It also suggests important directions for IR research.

Building on the literature. Research has shown that although an agent's perceived accuracy is important, other aspects can lead to acceptance and trust as well. For example, the perceived attractiveness of an agent leads to higher levels of reported intelligence, trust, and intent to continue using [Yuksel et al. 2016]. In addition, as has been reviewed above, the extent to which the agent mirrors the user's linguistic style (and possibly other gestures) is important. Prior research has implied that mimicry of all kinds can be advantageous in terms of human–human outcomes, in addition to the many "honest signals" [Pentland 2010] that humans automatically project. Further exploration should identify what kinds of mimicry and nonverbal social signals agents could exhibit that might lead to increased trust, satisfaction and likeability. We have listed above some phenomena from natural human-to-human conversations that are attested to in the literature. For each we might want to ask three sets of questions:

[7] Shneidernman also describes emulations as featuring humanoid forms, although clearly this is not universal.

(1) *Do we see the same phenomenon in information-seeking, human-to-agent, conversation as we do in general human-to-human conversation? If so, to what extent does it look the same?*

Some, but not all, of the phenomena discussed above have been demonstrated in human-to-agent conversation, but for the most part there is no published work describing these in a conversational IR setting. We might also ask what other aspects of the conversation lead to what phenomena: does the appearance of the agent matter? The mode of input or output? The type of task?

(2) *What does this mean for interaction or system design?*

Can an agent detect the phenomenon, and can it respond or participate appropriately? In many cases, capturing these phenomena would require extra tooling—for example, to capture prosody or affect as well as text. Extensions would be needed for representation and planning. Finally, this may also constrain agents' output, or suggest alternative interactions.

(3) *What would we expect as a consequence, if our designs took this phenomenon into account?*

In some cases, we might expect nearly the same effects with a human–computer conversation as with a human–human one; in other cases, we might think the effect would be different due to agent design fidelity, modality or task. We have also been assuming that more "natural" conversations are generally preferred, meaning that attention to conversational norms will lead to greater satisfaction; but some designs might, for example, increase accuracy or reduce time on task, so the latter dimensions continue to be obvious important design considerations.

For example, consider the phenomenon of lexical entrainment, demonstrated in human–computer dialogue by Brennan [1996]. This suggests adaptations both to input (perhaps speech recognition should assign more likelihood to words the agent itself has used before) and for output (perhaps we should prefer to use words the human has used before). This in turn means recognising where substitutions can be made and perhaps adapting the labelling in any knowledge base. We might expect these adaptations to lead to greater accuracy, more sense of a "respectful" interaction, and more sense of a "natural" conversation, but with no change in time on task or correctness.

Recommendations for research and design. Given the theory and research we have reviewed above, what might be some initial guidelines and recommendations for the design of conversational systems, and what research might be most useful? The steps outlined above give us some suggestions across a number of dimensions.

7.1 Recommendations: Social Communication

It would be interesting, and useful, to explore the notion of how an agent can better take the user's perspective throughout task completion and other interactions. Obviously, this becomes easier over time, as an agent interacts with a single user, so that standardly deployed user references and subsequent agent learning patterns could emerge. It is very likely that agent mimicry could be useful here, trying to get the user to expound upon certain intents, much like Weizenbaum's "therapist" ELIZA [Weizenbaum 1966]. Any efforts to provide the user with better insights into the agent's knowledge should likewise make it easier for the human to know the perspective taken by the agent [Braines et al. 2019]. Further research is needed to explore how to build the agent's knowledge of the user's mental model, though multimodal methods leveraging alternative interaction solutions offered by the agent might provide some hope here [Li et al. 2020].

Error correction. Conversational search is still in its infancy, and system errors or misunderstandings are common. It is therefore important that system designs both recognise errors and respond or correct or appropriately.

From studies using surveys and in-depth interviews, both Ashktorab et al. [2019] and Yuan et al. [2020] found that users prefer error responses with apologetic elements (and sometimes including sentiment), explanations of reasons for the communication error, and further guidance by the system for why the error occurred or how to remedy it. The SOVITE system [Li et al. 2020] was designed to help assist the user in understanding the agent's state when errors of comprehension occur. SOVITE is a multi-modal interaction system that assists users in identifying the causes of an agent's errors, leveraging the user interface of applications that might otherwise have been used to resolve a query. This was shown to help users recover from conversational breakdowns, using existing fill-in froms from existing interfaces to help ground conversation. Using an application's GUI, SOVITE displays the system's understanding of user intents and allows users to leverage the app fields as inputs for intent disambiguation, and enables users to repair breakdowns using direct manipulation on these fields. The results from a small user study suggested that SOVITE's approach is useful and revealed many opportunities for further improvements, since users were unfamiliar with direct manipulation on the application GUIs. The idea of leveraging known application user interfaces to adjust an agent's understanding of user intent is a reasonable one, since users are already familiar with these app interfaces for completing tasks. In general, the combination of graphical user interface elements with free-form conversation allows a nice fallback option for conversational breakdowns. Further exploration of this idea bears exploration.

7.2 Recommendations: Moves and Structures

From experience, we have a good understanding of the basic moves possible in conventional web search. Searcher actions, for example, include entering queries, accepting suggestions, or clicking on results; system actions include presenting ranked lists or offering facets for clarification. We also understand the structure or grammar of these actions and what they may mean for the search experience: for example, clicking on a result from the list, but quickly coming back (a "quick back") is typically taken as a symptom of a misleading and poor result [Agichtein et al. 2006; Hassan et al. 2010].

Natural-language conversations allow a larger space of possible actions from both the user and the system. Despite the work discussed in Section 4, there are at least three outstanding problems. First, can we build *and validate* a vocabulary of basic actions, such as those from Belkin et al. [1995], Bunt [2010], Saracevic et al. [1997], or Trippas et al. [2020]? This certainly seems plausible, although we note that actions must be recognisable by a machine in near-real-time [Jiang et al. 2015]. Second, can we identify conventional structures, such as suggested by Schegloff or Sacks et al. for natural conversation or postulated by Reichman [1985] or Brooks and Belkin [1983] for human–computer dialogue? While we expect a lot of structures and conventions to carry over from human-to-human conversation, we are not aware of any work identifying the similarities or differences with agents. Third, can we use these structures either to understand the human's utterances, to inform the system's output, or to diagnose when conversations go wrong? Such work could build on the labelling schemes above, as well as large conversational corpora, and techniques from the dialogue and the discourse planning literature.

7.3 Recommendations: Affect, Style, and Alignment

Affect and emotional responses. Emotions often act as a shortcut in communication. Designing agents with effective expression of emotions could help reduce the cognitive overhead of interactions. However, natural and convincing emotional behaviour is very difficult to synthesize. While the current state-of-the-art in synthesis of expressive agents is still challenging, expressive models for individual modalities are the most mature and could be leveraged in design.

Dialogue. As the nature and structure of dialogue is perhaps the most well studied, methods for generating emotional or empathetic dialogue are quite advanced [Ghosh et al. 2017; Huber et al. 2018; Rashkin et al. 2019; Zhou et al. 2017]. Transformer models offer impressive dialogue generation capabilities [Brown et al. 2020]; however, emotional and topical control of the output of these models still requires work. Almost all dialogue systems require a manager or some level of scripting in order avoid inappropriate, nonsensical or offensive outputs occurring in the course of a conversation. Grounding dialogue in other modalities (e.g., visual information about facial expressions) can be helpful for enabling subtle control over content [Huber and McDuff 2018].

Speech. Controllable speech synthesis has advanced significantly in recent years [Ma et al. 2018; Wang et al. 2017]. Before the development of end-to-end models, systems relied on parameterization of speech statistics (such as the fundamental frequency, range, contours and jitter) [Gunes et al. 2011]. However, it can become very complex and time consuming to develop expressive utterances when controlling the signals at this level of granularity using mark-up languages like SSML.[8] Neural models have replaced much of this fine-grained control with "style matching," in which a sample audio sequence with the intended emotion but different content, or higher-level hyperparameters are used to guide the model.

Visual Expressions. Naturalistic synthesis of facial expressions and body gestures arguably remains the most challenging of the these primary modalities (dialogue, speech, and visual expression). Visual behaviours differ when people are talking compared to listening and speaking with neutral tone compared to emotional tone [Busso et al. 2007]. These differences are subtle and tacit.

Creating natural coherence between multiple modalities of interaction is a very complex task [Gunes et al. 2011] and remains an active research topic. Beyond the technical challenges of creating natural movement, sound and content, the preferred level of expressivity of an interactive agent is something that is highly variable between people [Grover et al. 2020]. Despite the challenges, there is a wealth of evidence that the expression of affect is important for effective agents.

Style and alignment. As with affect and emotions, there is strong evidence for the effect of style, broadly defined, in natural human-to-human conversations [Brennan 1996; Fusaroli and Tylén 2015; Niederhoffer and Pennebaker 2002]. There is also strong evidence that these effects spill over into conversations with software agents [Bergmann et al. 2015; Branigan et al. 2010; Kühne et al. 2013; Pickering and Garrod 2004], as well as evidence for a weak effect in search tasks with software or humans [Thomas et al. 2018, 2020]. Further, there is evidence that aligning styles makes for a conversation that is variously described as less frustrating or more pleasant [e.g. Zepf et al. 2020], although the effect is dominated by basic conversational fluency [Koulouri et al. 2016].

We do not know, however, which aspects of conversational style are most important, nor in which modalities style matters (for example, video or audio or text), although there is early evidence that both word choice and prosody have independent effects [Zepf et al. 2020]. Nor do we know the relative importance of stylistic concerns compared to, for example, efficiency. Each of these questions need to be addressed.

Further, to date we have built mostly simple systems, capable of detecting "style" in a limited way and capable of matching humans only in a limited domain (e.g., Hoegen et al. [2019], Thomas et al. [2020], and Zepf et al. [2020]). An obvious engineering challenge is to extend this, to track style as a conversation evolves (no matter the subject); to adapt output accordingly (while being constrained by the necessities of retrieval, such as fidelity to source documents); and to understand how this effects the information retrieval experience in particular, which is goal-directed but also

[8] http://www.ssml.org/.

only loosely structured and where conversational roles are very asymmetric. Software to do this is in its infancy but we can expect it to be very useful.

7.4 Recommendations: Humanlike Simulation, Appearance, and Embodiment

Embodiment of conversational agents has been identified as helpful for several reasons. First, they help users locate the system. Second, they provide additional channels for expression of emotions and social cues, such as eye gaze or posture [Cassell 2001]. While it is probably unnecessary for some agents to have an embodied presence in certain situations, in others it can be very helpful.

Research on embodied agents has focused on modelling many behaviours, such as lip syncing [Taylor et al. 2017], head gestures [Jin et al. 2019], facial expressions [Kholgade et al. 2011], gait [Lv et al. 2016], and breathing motions [Promayon et al. 1997].

van Pinxteren et al. [2019] reported on a meta-analysis of over 60 papers reviewing conversational agents and their acceptability in the service domain (e.g., chatbots for customer service). Successful design behaviours included (but were not limited to) having a humanlike appearance, resembling the customer in appearance, the use of etiquette or politeness, gesturing cooperatively (e.g., head nods) and the incorporation of laughter. However, even for some of these behaviours, implementing them can come at the cost of decreased task performance by the human, attentional shifts by the user, or the unintended design toward undesirable personas for a user. The authors cautioned, therefore, that care should be taken when attempting to leverage these humanlike qualities in conversational design.

7.5 Recommendations: Data and Corpora

Development of conversational systems relies crucially on corpora: for response retrieval, to train generative models, or for evaluation. There is accordingly a fast-growing number of conversational IR corpora or other corpora being pressed in to service for conversational IR.

Penha et al. [2019] describe several criteria: A corpus should be multi-turn, information-seeking, mixed-initiative, multi-intent and multi-domain, and grounded in some external knowledge base. These are useful goals, but even a corpus that meets all of them may not capture the full range of phenomena above. For example, a corpus of text alone cannot capture prosody. Corpora without video exclude facial expressions. Corpora of only short exchanges (or reference answers) give us no way to talk about longer-range structure. Corpora based on systems such as Siri, Alexa, or Cortana will only tell us how people use agents now, not what a natural conversation looks like. Unfortunately, corpora to date have not been both broad enough to capture a variety of phenomena and simultaneously large enough to be useful for training or evaluation.

For example, SRI has made available transcripts of telephone calls to travel agents, recorded in the late 1980s and manually transcribed [SRI International 2011]. These conversations are a combination of information-seeking and transactional needs—both "how can I get to Chicago?" and "book me a flight"—so although not purely information-seeking they may be a good analogue for IR conversations. Several other corpora are widely used but have even less focus on information seeking. These include Switchboard (Godfrey et al. [1992]; general chit-chat), HCRC map task (Anderson et al. [1991]; instruction and task completion), Verbmobil (Alexandersson et al. [1997]; arranging meetings), and Meeting Recorder International Computer Science Institute [2004a, b]. More recent corpora have drawn on online forums, other online logs, and recordings of human-to-human information-seeking conversation (see, e.g., Penha et al. [2019] for a list or a list from Hauff at https://github.com/chauff/conversationalIR), as well as crowdsourced "conversations," including Question Answering in Context [Choi et al. 2018] and the TREC Conversational Assistance Track.[9]

[9] http://www.treccast.ai/.

Past work has also collected a good deal of data, including annotated transcripts of information-seeking exchanges, which in may in principle be made available for research; for example, see Belkin, Brooks, and Daniels [Belkin et al. 1995; Brooks and Belkin 1983; Daniels et al. 1985], Saracevic et al. [Saracevic et al. 1997; Spink and Saracevic 1997], or Radford et al. [Radford and Connaway 2013; Radford et al. 2011]. To the best of our knowledge, these have not been made generally available.

We are aware of only three recent attempts to distribute data from natural, human-to-human, information-seeking contexts: MISC [Thomas et al. 2017], SCSdata [Trippas et al. 2017], and MetaLWOz [Lee et al. 2019]. Only one (MISC) includes multi-modal and self-report data. However, it covers only four tasks, and what has been recorded is crucially dependent on the precise circumstances of collection [Trippas and Thomas 2019]. The MetaLWOz data are an impressive example, containing 37,884 crowdsourced dialogues recorded between two humans simulating interactions between a human and a bot. The dataset contains a large number of domains (47) and tasks (227), and each dialogue is a minimum of 10 turns. Since participants thought they were talking with a bot, these data may show different or attenuated effects and this difference would in itself be interesting.

In any data collection, especially that of conversations with personal agents or about personal circumstances, ethical considerations will naturally restrict how data are collected and used—for example, annotation by third parties may not be possible—and this makes the project somewhat more difficult. Nevertheless, further collection of rich, multi-modal, and natural corpora and annotations could be a great help to the work suggested above.

7.6 Recommendations: Metrics and Instruments

There has been substantial work on both metrics and instruments for ad hoc search and for semi-structured dialogue with software agents [Simpson 2020]. However, while there are some guidelines for computer-mediated interviews [Radford and Connaway 2013; Shachaf and Horowitz 2008], there has not been the same attention to metrics in conversational search. Nor has there been much attention to metrics or instruments for most of the phenomena described above.

Simple measures. Reference answers—pre-determined best responses at any one point in the interaction—are the basis of much evaluation, including all standard IR measures and many measures from natural language processing (e.g., Lin [2004] and Papineni et al. [2002]). Despite the attraction, it is far from clear that reference answers are useful for evaluating conversation: Quality is a property of the entire conversation, not a single utterance, and there are obvious problems producing a single reference answer—let alone session—when there many turns.

Similarly, success rate is also commonly used for both IR and dialogue systems. It is, however, clearly possible to “succeed” with an agent that is frustrating, or rude, or confusing (or indeed to fail in a task but still enjoy it). This observation is supported by work from Kiseleva et al. [2016], who found low correlation between success and satisfaction even in simple, highly structured tasks. Earlier work from Tagliacozzo [1977] also found little correlation for mediated searches on MEDLINE.

Time, measured in seconds or turns, is also common: Gibbon et al. [1998], for example, list standard (and recommended) measures including dialogue duration and turn duration while many conventional IR measures are expressed as gain-per-document (effort) or gain-per-time [Azzopardi et al. 2018b]. This is important, of course, but given what we know about the diversity of human linguistic styles it is unlikely to be the full story.

Other simple measures, specific to conversation and common in the IR literature, include repetition (considered bad), specificity (considered good), and relatedness (considered good). In chit-chat, however, these effects are more nuanced. For example, repetition occurs in human-to-human chat—although typically only in certain ways—and specificity and relatedness are violated [Gilmartin and Saam 2020]. It is reasonable to assume there is similar nuance in information-seeking contexts and therefore that these simple measures will be misleading.

In agents that speak, in particular, we can also measure intelligibility and naturalness of the speech itself [International Telecommunication Union 1994; Nusbaum et al. 1995], although again this can only be a partial measure.

Conversation measures. Walker et al.'s PARADISE [Walker et al. 1997] may be the best-known general framework specifically for evaluating conversational agents. It consists of a combination of task success measures and dialogue costs, the latter including efficiency and quality. Walker et al. simply use the number of repairs, combined with a score for success in a weighted combination. The weights of this combination are learned by linear regression, to predict satisfaction scores provided by judges. This is simple to compute but inadequate for open-ended tasks. It also depends on the particular task and systems being measured, so it is not appropriate for cross-task or cross-system evaluations. However, the "efficiency" and "qualitative" measures are more generally appropriate.

The SERVQUAL method and measures were developed in marketing research [Parasurman et al. 1988] and include a questionnaire and suggested analyses. The questionnaire collects perceptions of a company's (or product's) performance, respondents' expectations, and minimum standards in each of five dimensions—tangibles, reliability, responsiveness, assurance, and empathy [Parasurman et al. 1988]. It was examined for spoken dialogue systems by Hartikainen et al. [2004], who concluded that all five dimensions were appropriate. These also seem useful for conversational IR systems, but we are not aware of any serious attempt to measure systems on these dimensions.

Hone and Grahan took a similar approach to develop their **Subjective Assessment of Speech System Interfaces (SASSI)** [Hone and Graham 2000]. Exploratory factor analysis from an initial set of 50 items revealed six factors: system response accuracy, likeability, cognitive demand, annoyance, habitability, and speed. Only the first three of these seem internally consistent, however, and we note that these three factors do not align particularly well with the five dimensions of SERVQUAL.

Both SERVQUAL and SASSI rely on questionnaires after every conversation, so would scale poorly. It might, however, be appropriate to use them, or related scales such as O'Brien and Tom's User Engagement Scale [O'Brien and Toms 2010] to validate other measures.

Many of the interesting phenomena described above would be hard to measure with our current instruments, and much of what makes a conversation "good" is not covered by our present metrics. Further development here could be particularly useful for research and practice: For example, it would be fruitful to investigate the trade-offs and relative importance of accuracy, responsiveness, and the conversational behaviours discussed in this article. Any thorough evaluation of conversations will necessarily be multi-dimensional, so it also remain to design and validate a suite of metrics to cover conversational phenomena, and to refine this (for example via factor analyses) to a minimal, workable set. To date, there has been heavy reliance on self-reported data for conversation—unlike the hands-off approach to evaluating other search systems—and it may be worthwhile studying camera data, for example, or prosodic signals, to look for signals that predict, e.g., frustration or empathy. It would also be instructive to re-analyse existing metrics and benchmarks with an eye to measuring or discussing these extra criteria.

8 SUMMARY

The literature on conversation suggests many phenomena of interest to conversational search, from general communication theories (e.g., perspective taking and common grounding), to norms

(Gricean maxims, politeness) to structures and sequences (adjacency pairs, higher-level discourse patterns) and aspects of prosody, affect, style, matching, and alignment. As our tools get more sophisticated and better able to manage the basics of conversation; and as we build more sophisticated corpora and measures; we will be better able to investigate these phenomena, and design and build better agents.

ACKNOWLEDGMENTS

We thank our anonymous reviewers for their thoughtful comments on an earlier version of this article. Nick Belkin also provided many useful pointers in an (even) earlier version.

REFERENCES

Eugene Agichtein, Eric Brill, and Susan Dumais. 2006. Improving web search ranking by incorporating user behavior information. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval.* 19–26.

Jan Alexandersson, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Elisabeth Maier, Norbert Reithinger, Birte Schmit, and Melanie Siegel. 1997. *Dialogue Acts in VERBMOBIL-2.* Verbmobil report 204.

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval.* 475–484.

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* 1–13.

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. 1991. The HCRC map task corpus. *Lang. Speech* 34, 4 (1991), 351–366.

Deepali Aneja, Daniel McDuff, and Shital Shah. 2019. A high-fidelity open embodied avatar with lip syncing and expression capabilities. arXiv:1909.08766 [cs.HC]. Retrieved from https://arxiv.org/abs/1909.08766.

Jeremy Ang, Rajdip Dhillon, Ashley Krupski, Elizabeth Shriberg, and Andreas Stolcke. 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proceedings of the International Conference on Spoken Language Processing.* 2037–2040.

Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* 1–12.

Leif Azzopardi, Mateusz Dubiel, Martin Halvey, and Jeffrey Dalton. 2018a. Conceptualizing agent-human interactions during the conversational search process. In *Proceedings of the International Workshop on Conversational Approaches to Information Retrieval.*

Leif Azzopardi, Paul Thomas, and Nick Craswell. 2018b. Measuring the utility of search engine result pages: An information foraging based measure. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval.* 605–614.

Robert F. Bales. 1950. *Interaction Process Analysis: A Method for the Study of Small Groups.* Addison-Wesley Press.

Simon Baron-Cohen. 1997. *Mindblindness: An Essay on Autism and Theory of Mind.* MIT Press, Cambridge, MA.

Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. 2019. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychol. Sci. Publ. Interest* 20, 1 (2019), 1–68.

A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. 2003. How to find trouble in communication. *Speech Commun.* 40, 1–2 (Apr. 2003), 117–143.

N. J. Belkin, C. Cool, A. Stein, and U. Thiel. 1995. Cases, scripts, and information seeking strategies: On the design of interactive information retrieval systems. *Expert Syst. Appl.* 9, 3 (1995), 379–395.

M. M. Berg. 2014. Modelling of natural dialogues in the context of speech-based information and control systems. Ph.D. Dissertation. University of Kiel.

Kirsten Bergmann, Holly P. Branigan, and Stefan Kopp. 2015. Exploring the alignment space—Lexical and gestural alignment with real and virtual humans. *Front. ICT* 2, Article 7 (2015).

T. Bickmore and J. Cassell. 2001. Relational agents: A model and implementation of building user trust. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.*

Dave Braines, Richard Tomsett, and Alun Preece. 2019. Supporting user fusion of AI services through conversational explanations. In *Proceedings of the Annual Conference on Information Fusion.* 1–8.

Holly P. Branigan, Martin J. Pickering, Jamie Pearson, and Janet F. McLean. 2010. Linguistic alignment between people and computers. *J. Pragmat.* 49, 9 (2010), 2355–2368.

Cynthia Breazeal. 2002. Emotion and sociable humanoid robots. *Int. J. Hum.-Comput. Stud.* 59, 1–2 (2002), 119–155.

Susan E. Brennan. 1996. Lexical entrainment in spontaneous dialog. In *Proceedings of the International Symposium on Spoken Dialogue.*

H. M. Brooks and N. J. Belkin. 1983. Using discourse analysis for the design of information retrieval interaction mechanisms. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval.* 31–47.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Use.* Cambridge University Press, Cambridge.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. arXiv:2005.14165 [cs.CL]. Retrieved from https://arxiv.org/abs/2005.14165.

Harry Bunt. 2010. *DIT++ Taxonomy of Dialogue Acts, Release 5.* Retrieved June 2016 from http://dit.uvt.nl/.

Carlos Busso, Zhigang Deng, Michael Grimm, Ulrich Neumann, and Shrikanth Narayanan. 2007. Rigid head motion in expressive speech animation: Analysis and synthesis. *IEEE Trans. Aud. Speech Lang. Process.* 15, 3 (2007), 1075–1086.

Justine Cassell. 2001. Embodied conversational agents: Representation and intelligence in user interfaces. *AI Mag.* 22, 4 (2001), 67–84.

J. Cassell and K. R. Thorisson. 1999. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Appl. Artif. Intell.* 13, 4–5 (1999), 519–538.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. arXiv:1808.07036 [cs.CL]. Retrieved from https://arxiv.org/abs/1808.07036.

Mark G. Core and James Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *Proceedings of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, Vol. 56. Boston, MA.

P. J. Daniels, H. M. Brooks, and N. J. Belkin. 1985. Using problem structures for driving human-computer dialogues. In *Recherche d'Informations Assistée par Ordinateur.* 645–660.

Rajdip Dhillon, Sonali Bhagat, Hannah Carvey, and Elizabeth Shriberg. 2004. *Meeting Recorder Project: Dialog act labeling guide.* Technical Report TR-04-002. International Computer Science Institute.

Mateusz Dubiel, Martin Halvey, Pilar Oplustil Gallegos, and Simon King. 2020. Persuasive synthetic speech: Voice perception and user behaviour. In *Proceedings of the Conference on Conversational User Interfaces.* Article 6.

Riccardo Fusaroli and Kristian Tylén. 2015. Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cogn. Sci.* 40, 1 (2015), 145–171.

Jianfeng Gao, Michel Galley, and Lihong Li. 2019. Neural approaches to conversational AI. *Found. Trends Inf. Retriev.* 13, 2–3 (2019), 127–298. https://doi.org/10.1561/1500000074

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A neural language model for customizable affective text generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* 634–642.

D. Gibbon, R. Moore, and R. Winski (Eds.). 1998. *Handbook of Standards and Resources for Spoken Language Systems.* Walter de Bruyter, Berlin.

Emer Gilmartin and Christian Saam. 2020. Pragmatics research and non-task dialog technology. In *Proceedings of the Conference on Conversational User Interfaces.* Article 22.

Jeffrey M. Girard and Daniel McDuff. 2017. Historical heterogeneity predicts smiling: Evidence from large-scale observational analyses. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition.* IEEE, 719–726.

Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2017. Towards designing cooperative and social conversational agents for customer service. In *Proceedings of the International Conference on Information Systems.*

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. 517–520.

H. Paul Grice. 1975. Logic and conversation. In *Syntax and Semantics*, Peter Cole and Jerry L Morgan (Eds.). Vol. 3. Academic Press, New York, NY, 41–58.

Ted Grover, Kael Rowan, Jina Suh, Daniel McDuff, and Mary Czerwinski. 2020. Design and evaluation of intelligent agent prototypes for assistance with focus and productivity at work. In *Proceedings of the 25th International Conference on Intelligent User Interfaces.* 390–400.

Hatice Gunes, Björn Schuller, Maja Pantic, and Roddy Cowie. 2011. Emotion representation, analysis and synthesis in continuous space: A survey. In *Face and Gesture 2011.* IEEE, 827–834.

A. Hamacher, N. Bianchi-Berthouze, A. G. Pipe, and K. Eder. 2016. Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical Human-robot interaction. In *Proceedings of the Robot and Human Interactive Communication.*

Mikko Hartikainen, Esa-Pekka Salonen, and Markku Turunen. 2004. Subjective evaluation of spoken dialogue systems using SERVQUAL method. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'04)*. 2273–2276.

Ahmed Hassan, Rosie Jones, and Kristina Lisa Klinkner. 2010. Beyond DCG: User behavior as a predictor of a successful search. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. 221–230.

Elaine Hatfield, John T. Cacioppo, and Richard L. Rapson. 1992. Primitive emotional contagion. *Rev. Personal. Soc. Psychol.* 14 (1992), 151–177.

Paul ten Have. 2007. *Doing Conversation Analysis* (2nd ed.). SAGE, London.

T. Hennoste, O. Gerassimenko, R. Kasterpalu, M. Koit, A. Rääbis, K. Strandson, and M. Valdisoo. 2005. Information-sharing and correction in estonian information dialogues: Corpus analysis. In *Proceedings of the 2nd Baltic Conference on Human Language Technologies*. 249–254.

Rens Hoegen, Deepali Anjea, Daniel McDuff, and Mary Czerwinski. 2019. An end-to-end conversational style matching agent. In *Proceedings of the Intelligent Virtual Agents* (Paris). 111–118.

Kate S. Hone and Robert Graham. 2000. Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Nat. Lang. Eng.* 6, 3–4 (Sep. 2000), 287–303.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Trans. Inf. Syst.* 38, 3, Article 21 (Jun. 2020).

Bernd Huber and Daniel McDuff. 2018. Facial expression grounded conversational dialogue generation. In *Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG'18)*. IEEE, 365–372.

Bernd Huber, Daniel McDuff, Chris Brockett, Michel Galley, and Bill Dolan. 2018. Emotional dialogue generation using image-grounded language models. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

International Computer Science Institute. 2004a. The ICSI Meeting Corpus. Retrieved June 2016 from http://www1.icsi.berkeley.edu/Speech/mr/.

International Computer Science Institute. 2004b. Meeting Recorder Dialog Act (MRDA) Database. Retrieved June 2016 from http://www1.icsi.berkeley.edu/~ees/dadb/.

International Telecommunication Union. 1994. A method for subjective performance assessment of the quality of speech voice output devices. ITU-T recommendation P.85.

Molly E. Ireland, Richard B. Slatcher, Paul W. Eastwick, Lauren E. Scissors, Eli J. Finkel, and James W. Pennebaker. 2011. Language style matching predicts relationship initiation and stability. *Psychol. Sci.* 22, 1 (2011), 39–44.

Zongcheng Ji, Zhengdong Lu, and Hang Li. 2014. An information retrieval approach to short text conversation. arXiv:1408.6988v1 [cs.IR]. Retrieved from https://arxiv.org/abs/1408.6988v1.

J. Jiang, A. H. Awadallah, R. Jones, U. Ozertem, I. Zitouni, R. G. Kulkarni, and O. Z. Khan. 2015. Automatic online evaluation of intelligent assistants. In *Proceedings of the International Conference on World Wide Web*. 506–516.

Aobo Jin, Qixin Deng, Yuting Zhang, and Zhigang Deng. 2019. A deep learning-based model for head and eye motion generation in three-party conversations. *Proc. ACM Comput. Graph. Interact. Techn.* 2, 2 (2019), 1–19.

Eunice Jun, Daniel McDuff, and Mary Czerwinski. 2019. Circadian rhythms and physiological synchrony: Evidence of the impact of diversity on small group creativity. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (2019), 60.

Jari Kätsyri, Klaus Förger, Meeri Mäkäräinen, and Tapio Takala. 2015. A review of empirical evidence on different uncanny valley hypotheses: Support for perceptual mismatch as one road to the valley of eeriness. *Front. Psychol.* 6 (2015), 390.

Natasha Kholgade, Iain Matthews, and Yaser Sheikh. 2011. Content retargeting using parameter-parallel facial layers. In *Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. ACM, 195–204.

J. Kiseleva, K. Williams, J. Jiang, A. H. Awadallah, A. C. Crook, I. Zitouni, and T. Anastasakos. 2016. Understanding user satisfaction with intelligent assistants. In *Proceedings of the Conference on Human Information Interaction and Retrieval*. 121–130.

Theodora Koulouri, Stanislao Lauria, and Robert D. Macredie. 2016. Do (and Say) as I say: Linguistic adaptation in human-computer dialogs. *Hum.-Comput. Interact.* 31, 1 (2016), 59–95.

Nicole C. Krämer, Astrid von der Pütten, and Sabrine Eimler. 2012. Human-agent and human-robot interaction theory: Similarities to and differences from human-human interaction. In *Human-computer Interaction: The Agency Perspective*, M. Zacarias and J. V. de Oliveira (Eds.). Number 396 in Studies in Computational Intelligence. Springer.

R. M. Krauss and S. R. Fussell. 1991. Perspective taking in communication: Representation of others' knowledge in reference. *Soc. Cogn.* 9, 1 (1991), 2–24.

Vivien Kühne, Astrid Marieke Rosenthal von der Pütten, and Nicole C. Krämer. 2013. Using linguistic alignment to enhance learning experience with pedagogical agents: The special case of dialect. In *Proceedings of the International Workshop on Intelligent Virtual Agents*. Springer, 149–158.

William Labov and David Fanshel. 1977. *Therapeutic Discourse: Psychotherapy as Conversation.* Academic Press, New York.

Jessica L. Lakin, Valerie E. Jefferis, Clara Michelle Cheng, and Tanya L. Chartrand. 2003. The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *J. Nonverb. Behav.* 27, 3 (2003), 145–162.

Robin Tolmach Lakoff. 1979. Stylistic strategies within a grammar of style. *Ann. N. Y. Acad. Sci.* 327, 1 (1979), 53–78.

Minha Lee. 2020. Speech acts redux: Beyond request-response interactions. In *Proceedings of the Conference on Conversational User Interfaces*. Article 13.

Sungjin Lee, Hannes Schulz, Adam Atkinson, Jianfeng Gao, Kaheer Suleman, Layla El Asri, Mahmoud Adada, Minlie Huang, Shikhar Sharma, Wendy Tay, and Xiujun Li. 2019. Multi-domain task-completion dialog challenge. In *Dialog System Technology Challenges 8*.

Geoffrey N. Leech. 1983. *Principles of Pragmatics*. Longman, London.

Toby Jia-Jun Li, Jingya Chen, Haijun Xia, Tom M. Mitchell, and Brad A. Myers. 2020. Multi-modal repairs of conversational breakdowns in task-oriented dialogs. In *Proceedings of the ACM Symposium on User Interface Software and Technology*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*.

Xiaolei Lv, Jinxiang Chai, and Shihong Xia. 2016. Data-driven inverse dynamics for human motion. *ACM Trans. Graph.* 35, 6 (2016), 163.

Shuang Ma, Daniel Mcduff, and Yale Song. 2018. Neural TTS stylization with adversarial and collaborative games. In *Proceedings of the International Conference on Learning Representations*.

Meeri Mäkäräinen, Jari Kätsyri, and Tapio Takala. 2014. Exaggerating facial expressions: A way to intensify emotion or a way to the uncanny valley? *Cogn. Comput.* 6, 4 (2014), 708–721.

Daniel McDuff, Paul Thomas, Mary Czerwinski, and Nick Craswell. 2017. Multimodal analysis of vocal collaborative search: A public corpus and results. In *Proceedings of the Annual Conference on Multimodal Interaction*.

Wade J. Mitchell, Kevin A. Szerszen Sr, Amy Shirong Lu, Paul W. Schermerhorn, Matthias Scheutz, and Karl F. MacDorman. 2011. A mismatch in the human realism of face and voice produces an uncanny valley. *i-Perception* 2, 1 (2011), 10–12.

Masahiro Mori. 2012. The uncanny valley. *IEEE Robotics and Automation* (June 2012), 98–100. Translated by Karl F. MacDorman and Norri Kageki.

Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *J. Soc. Issues* 56, 1 (2000), 81–103.

Clifford Nass, Youngme Moon, and Nancy Green. 1997. Are machines gender neutral? Gender-stereotypic responses to computers with voices. *J. Appl. Soc. Psychol.* 27, 10 (1997), 864–876.

Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 72–78.

Kate G. Niederhoffer and James W. Pennebaker. 2002. Linguistic style matching in social interaction. *J. Lang. Soc. Psychol.* 21, 4 (2002), 337–360.

Howard C. Nusbaum, Alexander L. Francis, and Anne S. Henly. 1995. Measuring the naturalness of synthetic speech. *Int. J. Speech Technol.* 1 (1995), 7–19.

Heather L. O'Brien and Elaine G. Toms. 2010. The development and evaluation of a survey to measure user engagement. *J. Assoc. Inf. Sci. Technol.* 61, 1 (2010), 50–69.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 311–318.

A. Parasurman, Valarie A. Zeithaml, and Leonard L. Berry. 1988. SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *J Retail.* 64, 1 (1988), 12–40.

Gustavo Penha, Alexandru Balan, and Claudia Hauff. 2019. Introducing MANtIS: A novel multi-domain information seeking dialogues dataset. arXiv:1912.04639v1 [cs.CL]. Retrieved from https://arxiv.org/abs/1912.04639v1.

Alex Pentland. 2010. *Honest Signals: How they Shape Our World*. MIT Press.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behav. Brian Sci.* 27, 2 (2004), 169–225.

Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice interfaces in everyday life. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Emmanuel Promayon, Pierre Baconnier, and Claude Puech. 1997. Physically-based model for simulating the human trunk respiration movements. In *Proceedings of the 1st Joint Conference; Computer Vision, Virtual Reality and Robotics in Medicine, CVRMed, and Medical Robotics and Computer-Assisted Surgery (CVRMed-MRCAS'97)*. Springer, 379–388.

Marie L. Radford and Lynn Silipigni Connaway. 2013. Not dead yet! A longitudinal study of query type and ready reference accuracy in live chat and IM reference. *Libr. Inf. Sci. Res.* 35, 1 (2013), 2–13.

Marie L. Radford, Gary P. Radford, Lynn Silipigni Connaway, and Jocelyn A. DeAngelis. 2011. On virtual face-work: An ethnography of communication approach to a live chat reference interaction. *Libr. Quart.* 81, 4 (2011), 431–453.

Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the Conference on Human Information Interaction and Retrieval*. ACM, 117–126.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y.-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5370–5381.

Byron Reeves. 2010. People do like people: The benefits of interactive online characters. *Madison Ave. J.* (13 Apr. 2010).

Byron Reeves and Clifford Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press, New York.

Rachel Reichman. 1985. *Getting Computers to Talk Like You and Me*. MIT Press, Cambridge, MA.

Corbin Rosset, Chenyan Xiong, Xia Song, Daniel Campos, Nick Craswell, Saurabh Tiwary, and Paul Bennett. 2020. Leading conversational search by suggesting useful questions. In *Proceedings of the International Conference on World Wide Web*. 1160–1170.

Magdalena Rychlowska, Yuri Miyamoto, David Matsumoto, Ursula Hess, Eva Gilboa-Schechtman, Shanmukh Kamble, Hamdi Muluk, Takahiko Masuda, and Paula Marie Niedenthal. 2015. Heterogeneity of long-history migration explains cultural differences in reports of emotional expressivity and the functions of smiles. *Proc. Natl. Acad. Sci. U.S.A.* 112, 19 (2015), E2429–E2436.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 50, 4 (1974).

Tefko Saracevic, Amanda Spink, and Mei-Mei Wu. 1997. Users and interme-diaries in information retrieval: What are they talking about? In *Proceedings of the Annual Conference on User Modeling*. 43–54.

Emanuel A. Schegloff. 1968. Sequencing in conversational openings. *Am. Anthropol.* 70, 6 (1968), 1075–1095.

Emanuel A. Schegloff and Harvey Sacks. 1973. Opening up closings. *Semiotica* 8, 4 (1973), 289–327.

Klaus R. Scherer and Grazia Ceschi. 2000. Criteria for emotion recognition from verbal and nonverbal expression: Studying baggage loss in the airport. *Personal. Soc. Psychol. Bull.* 26, 3 (2000), 327–339.

Michael Seymour, Kai Riemer, and Judy Kay. 2019. Mapping beyond the uncanny valley: A delphi study on aiding adoption of realistic digital faces. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS'19)*. 1–11.

Pnina Shachaf and Sarah M. Horowitz. 2008. Virtual reference service evaluation: Adherence to RUSA behavioral guidelines and IFLA digital reference guidelines. *Libr. Inf. Sci. Res.* 30, 2 (2008), 122–137.

Ameneh Shamekhi, Mary Czerwinski, Gloria Mark, Margeigh Novotny, and Gregory A. Bennett. 2016. An exploratory study toward the preferred conversational style for compatible virtual agents. In *Proceedings of the International Conference on Intelligent Virtual Agents*. Springer, 40–50.

Ben Shneiderman. 2020. Drawing lessons from AI's two grand goals: Human emulation and useful applications (unpublished).

James Simpson. 2020. Are CUIs just GUIs with speech bubbles? In *Proceedings of the Conference on Conversational User Interfaces*. Article 23.

Stefan Sitter and Adelheit Stein. 1992. Modeling the illocutionary aspects of information-seeking dialogues. *Inf. Process. Manage.* 28, 2 (1992), 165–180.

Amanda Spink and Tefko Saracevic. 1997. Interaction in information retrieval: Selection and effectiveness of search terms. *J. Assoc. Inf. Sci. Technol.* 48, 8 (1997), 741–761.

SRI International. 2011. *SRI's Amex Travel Agent Data.* . Retrieved June 2016 from http://www.ai.sri.com/~communic/amex/amex.html.

A. Stolkce, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. van Ess-Dykema, and M. Meeter. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Comput. Ling.* 26, 3 (2000), 339–373.

Renate Tagliacozzo. 1977. Estimating the satisfaction of information users. *Bull. Med. Libr. Assoc.* 65, 2 (1977), 243–249.

Deborah Tannen. 1987. Conversational style. In *Psycholinguistic Models of Production*, Hans W. Dechert and Manfred Raupach (Eds.). Ablex, Norwood, NJ.

Deborah Tannen. 2005. *Conversational Style: Analyzing Talk Among Friends* (New Ed.). Oxford University Press, New York, NY.

Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. 2017. A deep learning approach for generalized speech animation. *ACM Trans. Graph.* 36, 4 (2017), 93.

Paul Thomas, Mary Czerwinski, Daniel McDuff, Nick Craswell, and Gloria Mark. 2018. Style and alignment in information-seeking conversation. In *Proceedings of the Conference on Human Information Interaction and Retrieval*. 42–51.

Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2017. MISC: A data set of information-seeking conversations. In *Proceedings of the International Workshop on Conversational Approaches to Information Retrieval*.

Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2020. Expressions of style in information-seeking conversation with an embodied agent. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1171–1180.

Margaret L. Traeger, Sarah Strohkorb Sebo, Malte Jung, Brian Scassellati, and Nicholas A Christakis. 2020. Vulnerable robots positively shape human conversational dynamics in a humanrobot team. *Proc. Natl. Acad. Sci. U.S.A.* 117, 12 (2020), 6370–6375.

Johanne Trippas and Paul Thomas. 2019. Data sets for spoken conversational search. In *Proceedings of the Workshop on Barriers to Interactive IR Resources Re-use.*

Johanne R. Trippas, Lawrence Cavedon, Damiano Spina, and Mark Sanderson. 2017. How do people interact in conversational speech-only search tasks: A preliminary analysis. In *Proceedings of the Conference on Human Information Interaction and Retrieval.* 325–328.

Johanne R. Trippas, Damiano Spina, Paul Thomas, Mark Sanderson, Hideo Joho, and Lawrence Cavedon. 2020. Towards a model for spoken conversational search. *Inf. Process. Manage.* 57, 2 (2020).

Michelle M. E. van Pinxteren, Ruud W. H. Wetzels, Jessica Rüger, Mark Pluymaekers, and Martin Wetzels. 2019. Trust in humanoid robots: Implications for services marketing. *J. Serv. Market.* 33, 4 (2019), 507–518.

Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, and Maarten de Rijke. 2020. Query resolution for conversational search with limited supervision. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval.* 921–930.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics.* 271–280.

Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. arXiv:1703.10135 [cs.CL]. Retrieved from https://arxiv.org/abs/1703.10135.

Richard J. Watts. 2003. *Politeness.* Cambridge University Press, Cambridge, UK.

Joseph Weizenbaum. 1966. ELIZA—A computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (1966), 36–45.

Deanna Wilkes-Gibbs and Herbert H. Clark. 1992. Coordinating beliefs in conversation. *J. Mem. Lang.* 31, 2 (1992), 183–194.

Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *Science* 330, 6004 (2010), 686–688.

Liu Yang, Minghui Qiu, Chen Qu, Cen Chen, Jiafeng Guo, Yongfeng Zhang, W. Bruce Croft, and Haiqing Chen. 2020. IART: Intent-aware response ranking with transformers in information-seeking conversation systems. In *Proceedings of the International Conference on World Wide Web.* 2592–2598.

Sihan Yuan, Birgit Brüggemeier, Stefan Hillmann, and Thilo Michael. 2020. User preference and categories for error responses in conversational user interfaces. In *Proceedings of the Conference on Conversational User Interfaces.* Article 5.

Beste F. Yuksel, Mary Czerwinski, and Penny Collisson. 2017. Brains or beauty: How to engender trust in user-agent interactions. *ACM Trans. Internet Technol.* 17, 1 (2017).

Sebastian Zepf, Arijit Gupta, Jan-Peter Krämer, and Wolfgang Minker. 2020. EmpathicSDS: Investigating lexical and acoustic mimicry to improve perceived empathy in speech dialogue systems. In *Proceedings of the Conference on Conversational User Interfaces.* Article 2.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2017. Emotional chatting machine: Emotional conversation generation with internal and external memory. arXiv:1704.01074 [cs.CL]. Retrieved from https://arxiv.org/abs/1704.01074.

Received May 2020; revised October 2020; accepted November 2020