Theories of conversation for conversational IR

Paul Thomas
Microsoft
Canberra, ACT, Australia

Mary Czerwinksi
Microsoft
Redmond, WA, USA

Daniel McDuff
Microsoft
Cambridge, MA, USA

Nick Craswell
Microsoft
Bellevue, WA, USA

ABSTRACT

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 para-linguistic 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 human-like 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 position paper, we summarise some salient past work, focusing on social norms; structures; and affect, prosody and style. We also discuss some implications for research and design of conversational IR systems.

1 CONVERSATIONS: HUMAN AND MACHINE

"One of the most human things that human beings do is talk to one another". This observation by Labov and Fanshel [41] 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 [48], 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 software.

When people interact with machines, they carry over many of the same expectations, norms, biases, and behaviours as when interacting with humans [47, 57, 58]. 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. But it is even more true of machines that can "converse" in something like natural language. 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" [58, emphasis in original].

For example, research has demonstrated that people are polite to computers, despite being well aware that computers can't be offended; they prefer computers with "personalities" closer to their own; they also apply human stereotypes, such as those around gender, and biases such as preferring attractive "faces" [27, 47, 58, 82]. Conversational software which acts more like people, and in particular which recognises more of the conversational context [3, 32], is perceived as more trustworthy [11, 27], as well as more engaging [17]. It is also regarded as more intelligent [66], 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 be important: so what do we need to take into account, or perhaps prioritise, if we're building a conversational IR system? How should we decide what to build or what 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…" [59, emphasis ours]. A similar approach has been used by Brooks and Belkin [14], and by Daniels et al. [19]. 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.

The academic study of conversation, the way context is established and shared, and the mechanisms by which conversations are structured, goes back over forty years to early work by Sacks et al. [61] and others[1]. In this position paper we will survey key work from conversation analysis, as well as linguistics, philosophy, and social psychology, in three broad areas: basic conventions, the structures of conversations, and extra-textual aspects such as prosody and affect. We will attempt to summarise findings or phenomena relevant to conversational IR systems, and what this suggests for further research or engineering.

2 BASIC CONVENTIONS

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

2.1 Cooperation and Grice's Maxims

In a discussion of how people can unambiguously imply things which are unsaid, Grice notes that there is a structure to natural conversation: "our talk exchanges do not normally consist of a

[1] Ten Have [30] gives a useful overview of objects and methods.

CAIR'20, March 2020, Vancouver, BC

succession of disconnected remarks… at each stage, SOME possible conversational moves would be excluded as conversationally unsuitable" [26, p45; emphasis in original].[2] 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, he suggests, leads to conversation which is natural and easy to follow; violations lead to deception, 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 [58] gives examples in other media. We can assume that these violations lead to a poor experience. From this, Gnewuch et al. [24] 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.

2.2 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 [44, p80], and adds the maxims of tact, generosity, approbation, modesty, agreement, and sympathy. Brown and Levinson [15] independently derive similar behavioural norms from two related notions of "face", and rules for managing this (but see e.g. Watts [80] for a critique). A close reading of virtual reference desks by Radford et al. 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 [23, 60].

Again, we have evidence that politeness is important even when dealing with machines [58], and some suggestions of what "politeness" means. It would be worthwhile formalising these in such a way as an agent could use them, and verifying their effect.

3 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. Considerable work has tried to classify moves and their sequences.

[2] Labov and Fanshel [41] 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.

3.1 Moves

Reichman [59] 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.

Reichman notes that "at particular stages of discourse, some conversational moves are 'expected' and 'most appropriate'" [59, p29], 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 [18]. 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. [70] includes "only" 42 labels, all of which are mutually exclusive. Stolkce 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 [1]. 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 [20], Bunt's DIT++ [16], the COR scheme used by Belkin et al. [9], and a scheme developed by Batliner et al. [8] aimed specifically at detecting communication breakdown. Jiang et al. [37] 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 [56] 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. [5]. Neither of these schemes were grounded in observation, however. Trippas et al. [78] 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—focussing in later work on elicitations from users and intermediaries [62, 68]. To our knowledge this schema has not been used in more recent work.

Earlier work from other fields has developed further alternatives. For example, Bales's Interaction Process Analysis [6] uses twelve 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" [41], 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, we must also consider the prospects of building automatic classifiers.

3.2 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" [59, p7]. Work over many decades has enumerated and explained these structures and the mechanisms by which they are coordinated [61, 63; see also references in Brooks and Belkin [14]].

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) [64].[3] 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?", "$100") or offer-alternative/confirm ("how about Chinese?", "okay"). We would not expect to see e.g. request-facet/confirm ("how much does it cost?", "okay").

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 [14, 19]. 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. 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" which 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 [19].

Later work by Belkin et al. drew on the Conversational Roles Model [67] to further describe the structure of dialogues between searchers and intermediaries [9]. Belkin et al. identify fifteen 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 which, 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's normal can inform any evaluation.

4 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. Researchers have found that gestures and prosody can carry as much emotional content, or sometimes significantly more, than words do (i.e., you can judge emotion well without the words). This then raises the question of how to create an agent that understands and possibly produces these cues.

4.1 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 [29, 42].

There is evidence that if a human is more expressive in a channel (e.g., visual modality) that is not captured by an interlocutor they will be judged as less effective at communication [46]. Thus if an artificial system fails to code data from that modality, and consider it in its representation of the state, its performance might suffer.

While non-verbal behaviours can predict affective states such as frustration [4], 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 [7]. Interpreting vocal cues is subject to similar challenges; for example, a wizard-of-Oz experiment by Batliner et al. [8] mimicked a failing system, to provoke responses, but found prosodic signals were unreliable. 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

[3] Hennoste et al. [31] 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.

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?

4.2 Style and matching

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" [73, p.188]. This is the "how" of a conversation, as opposed to the "what", since we can provide the same information in many ways [10].

Tannen [72] 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 [15], and different strategies for presentation [43].

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. [74] 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.

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 [13, 21] and there is evidence of a similar effect in human-to-computer conversation, as well [12, 13, 40, 53]. In multiple studies of language usage researchers have observed increased linguistic style matching (LSM) between humans [36, 49]. 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 [38, 81]. Would embodied avatars that simulate some of these more subtle signals and also synchronise with humans lead to similar positive outcomes?

5 IMPLICATIONS FOR CONVERSATIONAL IR

We do not suggest that the survey above is complete. Nevertheless, 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 directions for IR research.

5.1 Building on the literature

We have listed above some phenomena from natural human-to-human conversations which are attested to in the literature. For each we might want to ask three sets of questions:

(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?*

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 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 cost more time.

For example, consider the phenomenon of lexical entrainment, demonstrated in human-computer dialogue by Brennan [13]. This suggests adaptations both to input (perhaps ASR 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.

Other ideas worth investigating include Grice's maxims; politeness and face; appropriate coding schemes; affect; and alignment.

5.2 Data

There is a fast-growing number of conversational IR corpora, or other corpora being pressed in to service for conversational IR. For example, SRI have made available transcripts of telephone calls to travel agents, recorded in the late 1980s and manually transcribed [69]. 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. [25]; general chit-chat); HCRC map task (Anderson et al. [2]; instruction and task completion); Verbmobil (Alexandersson et al. [1]; arranging meetings); and Meeting Recorder (International Computer Science Institute [34, 35]). 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. [52] for a list).

Penha et al. [52] 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 which 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.

Past work has 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 [9, 14, 19], Saracevic et al. [62, 68], or Radford et al. [54, 55]. To the best of our knowledge these have not been made generally available. We are aware of only two recent attempts to distribute data from natural, human-to-human, information-seeking contexts: MISC [75] and SCSdata [77]. 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 [76].

In any data collection, especially that of conversations with personal agents or about personal circumstances, ethical considerations also restrict how data is collected and used. Nevertheless further collection of rich, multi-modal and natural, corpora and annotations could be a great help to the work suggested above.

5.3 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. However, while there are some guidelines for computer-mediated interviews [54, 65], there has not been the same attention to metrics in conversational search. Nor has there been much attention to metrics or instruments for 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. 45, 50]. 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 which 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. [39], who found low correlation between success and satisfaction even in simple, highly structured tasks. Earlier work from Tagliacozzo [71] also found little correlation, for mediated searches on MEDLINE.

Time, measured in seconds or turns, is also common: Gibbon et al., for example, lists standard (and recommended) measures including dialogue duration and turn duration [22, Chapter 13] while many conventional IR measures are expressed as gain-per-document (effort) or gain-per-time. Again, this is important but it is unlikely to be the full story.

Conversation measures. Walker et al.'s PARADISE [79] 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. 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. On the other hand, the "efficiency" and "qualitative" measures are more generally appropriate.

The SERVQUAL method and measures collects perceptions of performance, respondents' expectations, and minimum standards in each of five dimensions—tangibles, reliability, responsiveness, assurance, and empathy [51]. It was examined for spoken dialogue systems by Hartikainen et al. [28], 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 Graham took a similar approach to develop their Subjective Assessment of Speech System Interfaces (SASSI) [33]. 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.

Both SERVQUAL and SASSI rely on questionnaires after every conversation, so would scale poorly. It might however be appropriate to use them 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.

6 SUMMARY

The literature on conversation suggests many phenomena of interest to conversational search, from general 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 build better agents.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their thoughtful comments. Nick Belkin provided many useful pointers in an earlier version of this work.

REFERENCES

[1] J. Alexandersson, B. Buschbeck-Wolf, T. Fujinami, E. Maier, N. Reithinger, B. Schmit, and M. Siegel. 1997. *Dialogue acts in VERBMOBIL-2.* Verbmobil report 204.

[2] A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert. 1991. The HCRC map task corpus. *Language and Speech* 34, 4 (1991), 351–366.

[3] D. Aneja, D. McDuff, and S. Shah. 2019. A High-Fidelity Open Embodied Avatar with Lip Syncing and Expression Capabilities. *arXiv preprint arXiv:1909.08766* (2019).

[4] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke. 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proc. International Conference on Spoken Language Processing*. 2037–2040.

[5] L. Azzopardi, M. Dubiel, M. Halvey, and J. Dalton. 2018. Conceptualizing agent-human interactions during the conversational search process. In *Proc. Int. W'shop on Conversational Approaches to Information Retrieval.*

[6] R. F. Bales. 1950. *Interaction process analysis: a method for the study of small groups.* Addison-Wesley Press.

[7] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak. 2019. Emotional expressions reconsidered: challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest* 20, 1 (2019), 1–68.

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

[9] 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 Systems with Applications* 9 (1995), 379–395.

[10] M. M. Berg. 2014. Modelling of natural dialogues in the context of speech-based information and control systems. PhD thesis, University of Kiel. (2014).

[11] T. Bickmore and J. Cassell. 2001. Relational agents: a model and implementation of building user trust. In *Proc. SIGCHI.*

[12] H. P. Branigan, M. J. Pickering, J. Pearson, and J. F. McLean. 2010. Linguistic alignment between people and computers. *J. Pragmatics* 42 (2010), 2355–2368.

[13] S. E. Brennan. 1996. Lexical entrainment in spontaneous dialog. In *Proc. Int. Symp. on Spoken Dialogue.*

[14] H. M. Brooks and N. J. Belkin. 1983. Using discourse analysis for the design of information retrieval interaction mechanisms. In *Proc. SIGIR.* 31–47.

[15] P. Brown and S. C. Levinson. 1987. *Politeness: Some universals in language use.* Cambridge University Press, Cambridge.

[16] H. Bunt. 2010. DIT++ taxonomy of dialogue acts, release 5. (2010). Retrieved June 2016 from http://dit.uvt.nl/

[17] J. Cassell and K. R. Thorisson. 1999. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence* 13, 4–5 (1999), 519–538.

[18] M. G. Core and J. Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*, Vol. 56. Boston, MA.

[19] 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.

[20] R. Dhillon, S. Bhagat, H. Carvey, and E. Shriberg. 2004. *Meeting Recorder Project: Dialog act labeling guide.* Technical Report TR-04-002. International Computer Science Institute.

[21] R. Fusaroli and K. Tylén. 2015. Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive Science* 40, 1 (2015), 145–171.

[22] D. Gibbon, R. Moore, and R. Winski (Eds.). 1998. *Handbook of standards and resources for spoken language systems.* Walter de Bruyter, Berlin.

[23] J. M. Girard and D. McDuff. 2017. Historical heterogeneity predicts smiling: Evidence from large-scale observational analyses. In *2017 12th IEEE International Conference on automatic face & gesture recognition (FG 2017).* IEEE, 719–726.

[24] U. Gnewuch, S. Morana, and A. Maedche. 2017. Towards designing cooperative and social conversational agents for customer service. In *Proc. Conf. on Information Systems.*

[25] J. J. Godfrey, E. C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proc. IEEE Conf. Acoustics, Speech, and Signal Processing*, Vol. 1. 517–520.

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

[27] 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 *Proc. Robot and Human Interactive Communication.*

[28] M. Hartikainen, E.-P. Salonen, and M. Turunen. 2004. Subjective evaluation of spoken dialogue systems using SERVQUAL method. In *Proc. Interspeech.* 2273–2276.

[29] E. Hatfield, J. T. Cacioppo, and R. L. Rapson. 1992. Primitive emotional contagion. *Review of personality and social psychology* 14 (1992), 151–177.

[30] P. ten Have. 2007. *Doing conversation analysis* (2nd ed.). SAGE, London.

[31] 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 *Proc. Second Baltic Conf. on Human Language Technologies.* 249–254.

[32] R. Hoegen, D. Anjea, D. McDuff, and M. Czerwinski. 2019. An end-to-end conversational style matching agent. In *Proc. Intelligent Virtual Agents.* 111–118.

[33] K. S. Hone and R. Graham. 2000. Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering* 6, 3-4 (Sept. 2000), 287–303.

[34] International Computer Science Institute. 2004. The ICSI meeting corpus. (2004). Retrieved June 2016 from http://www1.icsi.berkeley.edu/Speech/mr/

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

[36] M. E. Ireland, R. B. Slatcher, P. W. Eastwick, L. E. Scissors, E. J. Finkel, and J. W. Pennebaker. 2011. Language style matching predicts relationship initiation and stability. *Psychological science* 22, 1 (2011), 39–44.

[37] 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 *Proc. WWW.* 506–516.

[38] E. Jun, D. McDuff, and M. Czerwinski. 2019. Circadian Rhythms and Physiological Synchrony: Evidence of the Impact of Diversity on Small Group Creativity. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 60.

[39] 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 *Proc. CHIIR.* 121–130.

[40] V. Kühne, A. M. R. von der Pütten, and N. C. Krämer. 2013. Using linguistic alignment to enhance learning experience with pedagogical agents: The special case of dialect. In *Proc. Int. W'shop on Intelligent Virtual Agents.* Springer, 149–158.

[41] W. Labov and D. Fanshel. 1977. *Therapeutic discourse: Psychotherapy as conversation.* Academic Press, New York.

[42] J. L. Lakin, V. E. Jefferis, C. M. Cheng, and T. L. Chartrand. 2003. The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of nonverbal behavior* 27, 3 (2003), 145–162.

[43] R. T. Lakoff. 1979. Stylistic strategies within a grammar of style. *Annals of the New York Academy of Sciences* 327, 1 (1979), 53–78.

[44] G. N. Leech. 1983. *Principles of pragmatics.* Longman, London.

[45] C.-Y. Lin. 2004. ROUGE: a package for automatic evaluation of summaries. In *Proc. Workshop on Text Summarization Branches Out.*

[46] D. McDuff, P. Thomas, M. Czerwinski, and N. Craswell. 2017. Multimodal analysis of vocal collaborative search: A public corpus and results. In *Proc. Multimodal Interaction.*

[47] C. Nass and Y. Moon. 2000. Machines and mindlessness: Social responses to computers. *J. Soc. Issues* 56, 1 (2000), 81–103.

[48] C. Nass, J. Steuer, and E. R. Tauber. 1994. Computers are social actors. In *Proc. SIGCHI.* ACM, 72–78.

[49] K. G. Niederhoffer and J. W. Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology* 21, 4 (2002), 337–360.

[50] K. Papineni, S. Roukos, T. Ward, and W. jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL.* 311–318.

[51] A. Parasurman, V. A. Zeithaml, and L. L. Berry. 1988. SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *J Retailing* 64, 1 (1988), 12–40.

[52] G. Penha, A. Balan, and C. Hauff. 2019. Introducing MANtIS: a novel multi-domain information seeking dialogues dataset. arXiv:1912.04639v1 [cs.CL]. (2019).

[53] M. J. Pickering and S. Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brian Sciences* 27, 2 (2004), 169–225.

[54] M. L. Radford and L. S. Connaway. 2013. Not dead yet! A longitudinal study of query type and ready reference accuracy in live chat and IM reference. *Library and Information Science Research* 35, 1 (2013), 2–13.

[55] M. L. Radford, G. P. Radford, L. S. Connaway, and J. A. DeAngelis. 2011. On virtual face-work: An ethnography of communication approach to a live chat reference interaction. *Library Quarterly* 81, 4 (2011), 431–453.

[56] F. Radlinski and N. Craswell. 2017. A theoretical framework for conversational search. In *Proc. CHIIR.* ACM, 117–126.

[57] B. Reeves. 2010. People do like people: The benefits of interactive online characters. *Madison Avenue J* (13 April 2010).

[58] B. Reeves and C. Nass. 1996. *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge University Press, New York.

[59] R. Reichman. 1985. *Getting computers to talk like you and me.* MIT Press, Cambridge, Massachusetts.

[60] M. Rychlowska, Y. Miyamoto, D. Matsumoto, U. Hess, E. Gilboa-Schechtman, S. Kamble, H. Muluk, T. Masuda, and P. M. Niedenthal. 2015. Heterogeneity of long-history migration explains cultural differences in reports of emotional expressivity and the functions of smiles. *Proceedings of the National Academy of Sciences* 112, 19 (2015), E2429–E2436.

[61] H. Sacks, E. A. Schegloff, and G. Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 50 (1974).

[62] T. Saracevic, A. Spink, and M.-M. Wu. 1997. Users and interme-diaries in information retrieval: What are they talking about?. In *Proc. User Modeling*. 43–54.

[63] E. A. Schegloff. 1968. Sequencing in conversational openings. *American Anthropologist* 70 (1968), 1075–1095.

[64] E. A. Schegloff and H. Sacks. 1973. Opening up closings. *Semiotica* 8 (1973), 289–327.

[65] P. Shachaf and S. M. Horowitz. 2008. Virtual reference service evaluation: Adherence to RUSA behavioral guidelines and IFLA digital reference guidelines. *Library and Information Science Research* 30 (2008), 122–137.

[66] A. Shamekhi, M. Czerwinski, G. Mark, M. Novotny, and G. A. Bennett. 2016. An exploratory study toward the preferred conversational style for compatible virtual agents. In *Proc. Int. Conf. on Intelligent Virtual Agents*. Springer, 40–50.

[67] S. Sitter and A. Stein. 1992. Modeling the illocutionary aspects of information-seeking dialogues. *Inf. Proc. Mgmt* 28, 2 (1992), 165–180.

[68] A. Spink and T. Saracevic. 1997. Interaction in information retrieval: Selection and effectiveness of search terms. *J. Assoc. Information Science and Technology* 48, 8 (1997), 741–761.

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

[70] 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. *Computational Linguistics* 26, 3 (2000), 339–373.

[71] R. Tagliacozzo. 1977. Estimating the satisfaction of information users. *Bulletin of the Medical Library Association* 65, 2 (1977), 243–249.

[72] D. Tannen. 1987. Conversational style. In *Psycholinguistic models of production*, Hans W Dechert and Manfred Raupach (Eds.). Ablex, Norwood, NJ.

[73] D. Tannen. 2005. *Conversational style: Analyzing talk among friends* (new ed.). Oxford University Press, New York.

[74] P. Thomas, M. Czerwinski, D. McDuff, N. Craswell, and G. Mark. 2018. Style and alignment in information-seeking conversation. In *Proc. CHIIR*. 42–51.

[75] P. Thomas, D. McDuff, M. Czerwinski, and N. Craswell. 2017. MISC: A data set of information-seeking conversations. In *Proc. Int. W'shop on Conversational Approaches to Information Retrieval*.

[76] J. Trippas and P. Thomas. 2019. Data sets for spoken conversational search. In *Proc. W'shop on Barriers to Interactive IR Resources Re-use*.

[77] J. R. Trippas, L. Cavedon, D. Spina, and M. Sanderson. 2017. How do people interact in conversational speech-only search tasks: A preliminary analysis. In *Proc. ACM SIGIR Conf. Human Information Interaction and Retrieval*. 325–328.

[78] J. R. Trippas, D. Spina, P. Thomas, M. Sanderson, H. Joho, and L. Cavedon. 2020. Towards a model for spoken conversational search. *Inf. Proc. Mgmt* 57, 2 (2020).

[79] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *Proc. ACL*. 271–280.

[80] R. J. Watts. 2003. *Politeness.* Cambridge University Press, Cambridge.

[81] A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *science* 330, 6004 (2010), 686–688.

[82] B. F. Yuksel, M. Czerwinski, and P. Collisson. 2016. Brains or beauty: How to engender trust in user-agent interactions. *Trans. Internet Technology* (Sept. 2016).