

Expressions of Style in Information-Seeking Conversation with an Embodied Agent

Paul Thomas
Microsoft
Canberra, Australia

Mary Czerwinski
Microsoft Research
Redmond, WA, USA

Daniel McDuff
Microsoft Research
Cambridge, MA, USA

Nick Craswell
Microsoft
Bellevue, WA, USA

ABSTRACT

Past work in information-seeking conversation has demonstrated that people exhibit different *conversational styles*—for example, in word choice or prosody—that differences in style lead to poorer conversations, and that partners actively align their styles over time. One might assume that this would also be true for conversations with an artificial agent such as Cortana, Siri, or Alexa; and that agents should therefore track and mimic a user’s style. We examine this hypothesis with reference to a lab study, where 24 participants carried out relatively long information-seeking tasks with an embodied conversational agent. The agent combined topical language models with a conversational dialogue engine, style recognition and alignment modules. We see that “style” can be measured in human-to-agent conversation, although it looks somewhat different to style in human-to-human conversation and does not correlate with self-reported preferences. There is evidence that people align their style to the agent, and that conversations run more smoothly if the agent detects, and aligns to, the human’s style as well.

ACM Reference Format:

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 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*, July 25–30, 2020, Virtual Event, China. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3397271.3401127>

1 CONVERSATION AND STYLE

Conversational agents have grown in popularity in recent years, powered by significant improvements in speech recognition, natural language understanding, and text-to-speech synthesis. A successful agent must manage several tasks: accurately recording what the searcher says (possibly transcribing an audio stream); understanding the searcher’s utterance (for example, classifying the intent); deciding an appropriate response (for example, retrieving some

information or deciding to ask a question of its own); and rendering the response over an appropriate channel (possibly including a text-to-speech system). To be useful, at each stage it must keep track of the user’s task and keep its responses on-topic.

However, a successful conversation involves more than a mere exchange of on-topic facts. We can impart the same topical information in any number of ways [6]: for example, we have choices of wording, loudness, pitch, and timing as well as how we cooperate and take turns with our partner(s). Some of this is true even in text-only interfaces, but is especially apparent in audio or audio/video interfaces.

In long-running work, Tannen describes and analyses *conversational style*: “the use of specific linguistic devices, chosen by reference to broad conversational principles” [35], and we adopt this definition here: “style” is *how things are said*, rather than what is said, and covers both linguistic and paralinguistic, or non-verbal, properties of conversation. Characterising conversational style can help us understand why a person might be considered rude or abrupt by one listener, but as excited or passionate by another. As designers, we have a large degree of control over an agent’s style and we might ask: what style is best?

The effects of style have been well studied in natural, casual, informal conversation. They have not been so well studied in information-seeking conversation, nor in settings with a software agent, and it is an open question whether similar effects occur in this context.

“Embodiment”—presenting an apparent physical form—allows the presentation of richer non-verbal cues, and so we investigated style using an embodied conversational agent (ECA). In many ways ECAs are a natural extension of voice-only conversational agents, with the benefit that they help the user “locate” the intelligence [9]. Our agent takes an audio/video stream as input, processes user utterances and style, and produces styled audio/video output with a 3D-modelled head (Figure 1). The dialogue engine is capable of extended conversations (around 20 turns from each participant). This allows us to examine how people might behave in long-running, information-seeking conversations with an embodied agent, and compare this to similarly lengthy conversations between humans.

In this work we consider four questions:

RQ1 Do those styles first identified and measured in human-to-human conversation—including in chit-chat and in information seeking—also exist in information-seeking conversations with an embodied software agent?

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR '20, July 25–30, 2020, Virtual Event, China
© 2020 Copyright held by the owner/author(s). Publication rights licensed to the Association for Computing Machinery.
ACM ISBN 978-1-4503-8016-4/20/07...\$15.00
<https://doi.org/10.1145/3397271.3401127>

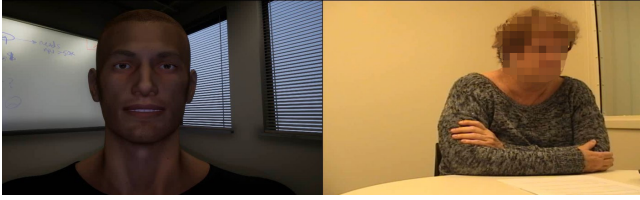


Figure 1: A conversation between our agent and a participant (obscured for publication). We captured audio and video.

RQ2 Do participants’ styles correlate with their perception of the agent?

RQ3 Do we see alignment of the humans’ styles to the agent’s, in the same way humans align to each other?

RQ4 Does it make a difference as to how the agent performs, or is perceived, if it attempts to align to the participants’ conversational style?

We show that style can be measured in our design, although it correlates with different behaviours in human-to-agent than in human-to-human conversation; that participants’ styles do make a small difference; that people do align to the agent’s style; and that when the agent also attempts to align, it is rated higher.

2 RELATED WORK

We draw on work from studies of information seeking and search, conversational style, and embodied agents.

2.1 Spoken Conversational Search

With the increasing popularity of voice-only interfaces, and increasing sophistication in natural language processing, conversational search has emerged as an active research area. In conversational search we have long-form interaction in natural language, asymmetry of role between user and intermediary, and a (possibly ill-defined) task. This contrasts with past work examining chit-chat [16]; cooperative tasks between peers [1, 2]; or slot-filling “dialogue” systems [21, 33].

A few studies have examined the structure information-seeking conversations, mostly between people or otherwise in the absence of running software [e.g. 5, 17, 26, 29, 30, 38]. There have also been attempts to describe what a conversation should look like [4, 11, 28]. Studies with agents have been limited, however, by the agents’ inability to hold a lengthy conversation [12]. There has also been very little discussion of *conversational style* in a retrieval setting.

2.2 Conversational Style

James Pennebaker et al. [20, 25] have extensively studied language, or linguistic, style matching (LSM), observing that increased style matching between humans was correlated with increased interest in the partner [20].

Tannen [34] draws a distinction between “high involvement” (HI) and “high consideration” (HC) styles in social conversation. Differing from Pennebaker et al.’s LSM, these styles also include non-verbal or paralinguistic features. A high involvement style emphasises enthusiasm and interpersonal involvement: features

include rapid speech, volume, questions, overlap, and reference to personal experience. A high consideration style, by contrast, emphasises independence and space and features longer pauses, hesitation, and less paralinguistic effect. These notions are also seen in, for example, Lakoff’s “camaraderie” and “distance” strategies [23]. In parallel with Pennebaker, Tannen further suggests that partners with opposing styles will find conversation frustrating.

There is little work on conversational style in information-seeking contexts, as opposed to social contexts. Tannen’s two styles were adopted by Thomas et al. [36], who used audio recordings and transcripts from the MISC data [37] to demonstrate that style could be characterised and computed automatically. They also found that differences in style have an effect on people’s perception of conversations. We draw on this work in the present experiments.

2.3 Style Matching

There is ample evidence of *alignment* in human-human conversation: that is, partners in conversation tend to converge on word choice, syntax, and even accents [15]. There is some evidence of a similar effect when people interact with computers as well [8, 22, 27]. As conversational agents become more sophisticated, we believe based on early evidence [18] that linguistic style matching could influence trust and likeability of the agent for the human user. We know from user-centered design studies that people unconsciously apply human social rules when interacting with computers: they prefer systems which appear to manifest personalities similar to their own [24] and agents that are more human-like [7]. Therefore, it is reasonable to assume that conversational systems would be preferred if they exhibit conversation styles that are more similar to that of the user. We test this hypothesis in the present work.

2.4 Embodied Conversational Agents

Voice-based conversational agents have become ubiquitous in the past few years. However, research into voice-only and embodied systems dates back several decades. Cassell [10] and Bickmore [7] pioneered the use of embodied conversational agents. The research has shown that embodied agents have several advantages over voice or text-only conversational agents. Embodiment helps users locate the source of the interaction and allows for a richer opportunity for the communication of non-verbal cues. An agent that has a physical presence also provides visual affordances, indicating where the user should focus their attention and receive cues from body and hand gestures, eye gaze and other expressions [10]. When artificial conversational agents engage in more natural social behaviours (e.g., using social dialogue), it helps increase user trust [7]. In this work we use an embodied conversational agent for these reasons.

The present work builds on the above, but is different in several aspects. We report on spoken conversations which run for several minutes (as opposed to one or two turns); for this, we use an embodied agent and we track and adapt conversational style in real time. In contrast to past work on conversational style, which uses recordings or transcripts of human-to-human conversation, we draw from recordings of conversations with an actual software agent. We also focus on information-seeking conversation, which has been little-studied in this context.

3 METHOD

We addressed the questions above with a laboratory study in February 2019.¹ Participants were presented with a software agent, represented onscreen as a 3D-modelled head and shoulders, and were asked to have a series of short conversations with the aim of gaining some information (Figure 1). For half of the tasks, the agent ran in “style matching” mode, while for the other half, it did not match. Post-task instruments asked about mental demand, “likeability”, and knowledge gain, while exit instruments canvassed conversational preferences, personality, and demographics. During each task we recorded audio and video, of both the participant and the agent, from which we derived indicators of style and emotion. We describe below our participants, the agent, the assigned tasks, and the data we gathered.

3.1 Overview

Figure 2 gives a typical conversation, where our participant has been asked to find out about a house for sale. This conversation runs to 34 turns: 17 turns each. The agent drew responses from a specialised model for seven utterances; the other ten, marked “★”, were drawn from a neural chit-chat model used when the participant’s intent was not recognised. The chit-chat responses were generated on the fly, allowing the agent a wider range than would have been possible in a scripted interaction.

When intents are correctly recognised, responses are scripted and the agent performs very well (e.g., lines 1–3). The fallback to chit-chat can of course be poorer in performance (line 8), but often works to cover unanticipated questions (e.g., lines 12 or 16–17, where we did not think to write a script). Questions can be repeated where the response was unsatisfactory (lines 4–5 or 7–8), and again chit-chat provides some flexibility for error recovery (line 9).

3.2 Participants

We recruited $N = 24$ participants (11 women, 13 men) from a volunteer database from the US Pacific Northwest. The participants had a range of education levels (high school to postgraduate) and job roles (mostly white-collar). The mean age was 39 years ($sd = 12$ years). Most said they used a virtual assistant daily (13) or weekly (6). All spoke English, although four used a different language at home. The participants were compensated US\$75 for participating in the study. They were free to stop the study at any time and could opt to have their data deleted. No participants exercised these options.

3.3 Tasks and Scripts

Participants were assigned four information-seeking tasks. Three of these (“hotel”, “house”, “car”) were in a fairly small domain, and required the participants to collect a set of facts; the other task (“introduction”) was more open-ended and meant for familiarisation. Each task included a brief motivation and some examples, which triggered scripted responses from the agent. Participants were, however, free to—and did—ask other questions or indeed wander off topic. The four tasks were:

Introduction You have just met the computer agent you will be working with today and you want to find out more about what makes him (the agent) tick. Introduce yourself to the agent and try to find out more about him by asking questions. Start with this: “*What do you do in your free time?*” “*What do you think about robots?*” To continue, you might ask his name, where he lives, what he does in his free time and what he thinks about artificial intelligence.

Hotel Imagine this agent has booked a room for you at the Bellevue Marriott hotel. You’d like to know more about the hotel and the neighbourhood. Find out what you can about the accommodation.

Start with this: “*Can I check into the hotel early?*” “*What’s the address of the hotel?*” To continue, you might ask about the room features, the hotel facilities, or whether there are your kind of attractions nearby.

House Imagine this agent has found a house for you that you are interested in buying. You want to find out as much as you can from the agent about the house.

Start with this: “*What sort of construction is the house?*” “*How many bedrooms does the house have?*” To continue, you might ask about the size of the house, whether it has the sort of features you want (like the number of bathrooms), or about the price.

Car Imagine this agent has found a Volvo car for you that you are interested in buying. Find out whether this Volvo is the type of a car you might like to buy.

Start with this: “*What kind of car is it?*” “*How many miles per gallon does the car get?*” To continue, you might ask about the features or engine details, or of course the price.

The introduction was always assigned first, but the other three tasks were rotated. The agent’s style matching was also rotated, so each participant had two tasks with and two without matching, and so that matching was balanced across tasks.

Each task was supported by a script: a language model plus a set of intents, each with several examples of participant questions and one canonical response. Each intent covered one fact—e.g., the hotel’s check-out time, or the car’s colour. Whenever the participant’s utterance seemed to match an intent, the corresponding response could be triggered. Exact matches were not required. For example, the hotel task included an intent for room size: questions such as “what sort of room is it?”, “how big is the room?” or “is it a suite?” could trigger “the standard room is 290 square feet”. For the house task, questions about location could trigger any of 23 responses from a description of the suburb to comments on nearby places of worship or bicycle paths. In total we had 223 responses with up to 15 examples each, to cover likely topics and wording variations.

Scripts and intents were supported by the Language Understanding (LUIS) toolkit².

3.4 The Agent

We developed an ECA for this study. Our agent included components for an end-to-end audio conversation: speech recognition,

¹This protocol was approved by the Microsoft Research Institutional Review Board.

²<http://luis.ai>

Turn	Participant (detected)	Agent
1	What sort of construction is the house?	It's a single-storey timber frame with timber siding in good condition.
2	How many bedrooms does the house have?	It's got four bedrooms.
3	How many bathrooms does it have?	There's a bathroom plus an ensuite.
4	Is there a big yard?	There's a garden with a pond, there's a lawn, and there are lots of mature trees around.
5	Is there a big yard?	I will find out and let you know. ★
6	Are the schools in the area good?	It's in the Lake Washington district so the schools are Rose Hill Elementary, Rose Hill Middle, and Lake Washington High. They all seem good.
7	Do you know anything about crime rates in the area?	What do you mean by crime crime? ★
8	A crime rates?	What crime is crime crime? ★
9	Never mind.	That's ok then. ★
10	What is the asking price?	They're asking 748 but you could probably get it for a bit less. The Zestimate is 804.
11	How soon can I move in?	As soon as you can. ★
12	Has the house been inspected?	Yes it has been sorted out. ★
13	Is there an attic?	Yes there is a house. ★
14	How many cars fit in the garage?	There's a double garage and also a carport.
15	I would like if it had a den.	I'm sure you would be able to find one. ★
16	What color is the house?	Grey and white. ★
17	I like blue houses.	I'm sure you do. ★

Figure 2: A sample conversation with our agent (participant 1, house task). Responses labelled ★ are from the neural dialogue generator (chit-chat), others are scripted.

open-domain dialogue generation, closed-domain intent recognition, dialogue management, and speech synthesis. Figure 3 illustrates this pipeline. The agent software also included components to detect and adapt to participants' conversational style, including prosody and word choice. These components are similar to the approach described in Hoegen et al. [18] and leverage an openly available avatar/virtual agent [3], but we offer a summary below.

Sensing. The audio sensing pipeline processed a 16kHz audio stream. The Microsoft Windows system voice activity detector was used with the Microsoft speech API³ for speech recognition. We also extracted linguistic style variables [36] from the user's utterance; these are described in Section 3.5 below.

Synthesis. The agent selected its response with regard to the scripts described above, and to a neural dialogue generator model for chit-chat [32]. The chit-chat model was capable of generating context-sensitive dialogue in response to an open-domain input, and was used whenever the intent module was not able to classify the user's intent. Speech synthesis used the Microsoft API⁴.

The Phoneme Recogniser component automatically output a corresponding sequence of visual groups of phonemes (visemes) from the synthesised speech output. We used PocketSphinx [19] for phoneme recognition and mapped the phonemes to a set of 20 standard visemes [13]. These controlled the agent's facial presets for lip syncing.

"Style matching". In half the tasks our agent matched conversational style via two kinds of variables as described by Hoegen et al. [18]: (1) content of the utterances like the use of pronouns, word repetition, utterance length, and (2) prosodic qualities like speech rate, pitch, and loudness. Both scripts and chit-chat could be adjusted.

3.5 Data Recorded and Derived

We recorded video and audio during each task; self-reported data after each task; and self-reported data during the exit interview.

Task performance. During each task, we recorded audio and video from both the participant and the agent. We also noted events of interest, from a fixed set: the agent speaking over the human or vice versa, the agent or the human veering off topic, the agent freezing or not responding at all, the human repeating her- or himself, the agent talking to itself, or the human asking the agent to stop talking.

Style. Thomas et al. [36] identified eleven variables which can be identified with Tannen's "high consideration" to "high involvement"; we recorded seven of these, chosen for their ease of processing in real time. (Since we have therefore modified the original instruments, we validate our new instrument below.) From the transcripts we extracted four variables describing linguistic features:

Personal pronouns: the rate of use of first- and second-person personal pronouns, as a proportion of words in each utterance.

Term-level repetition: the proportion of terms in one utterance which were repeated from the participant's previous utterance, respecting task boundaries. Disfluencies such as "umm" were removed, as were stopwords, and the remaining words were stemmed before repetitions were counted. Thus, "They've been lecturing" and "I went to a lecture" would count as one overlapped term.

Utterance-level repetition: the proportion of utterances with terms repeated from the previous; that is, the proportion of utterances where term-level repetition is greater than zero.

Length of utterances: the number of words per utterance.

From the audio signal we extracted another three variables, describing prosodic features.

Rate of speech: measured in words per second; the word count, for each utterance, divided by the length in seconds.

³<https://azure.microsoft.com/en-us/services/cognitive-services/speech/>

⁴<https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech/>

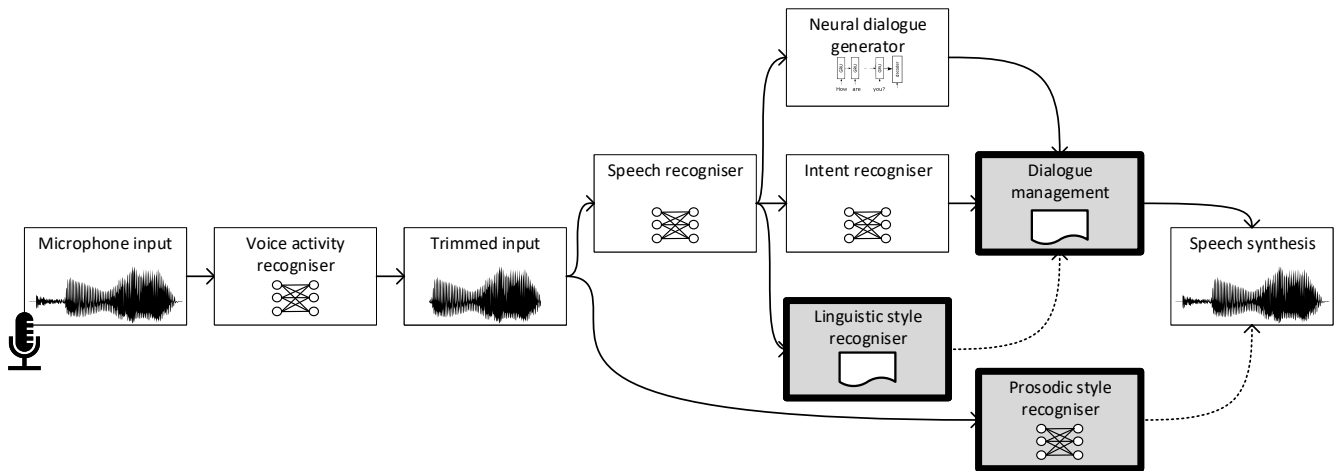


Figure 3: Agent processing of the participant's audio. Data from the shaded boxes is analysed in the present paper. Dialogue management (selecting an utterance) and speech synthesis (choosing rate, pitch, and loudness) made use of participants' own style when the agent was in "style matching" mode (shown with dotted arrows). Excluded from this diagram are video processing, and elements of the graphical output such as head pose modelling and lip sync; see Hoegen et al. [18] for details.

Pitch variation: the variance in fundamental pitch, across the whole utterance. We used OpenSMILE [14] to process the audio signal and ignored samples with no speech.

Loudness variation: the variance in loudness, measured the same way.

Pronoun use corresponds to Tannen's personal topics; repetitions approximate her concept of re-visiting topics; utterance length and rate of speech relate to pauses and turn-taking; and prosodic variation captures "expressive phonology".

Past work by Thomas et al. [36] used observations over a complete task to measure involvement. This is of course impossible with an agent operating in real time, so here we use a rolling mean over five utterances. Accordingly we also disregard utterances 1–4 in our analysis below.

Inspection showed that, with the exception of utterance length and rate of speech, these variables were skewed. We therefore log-transformed each of these remaining five variables. The scale of each variable is also wildly different, so to aid modelling and interpretation we re-scaled each variable to have zero mean and unit standard deviation.

Post-task self-reports. After each task, participants used a separate computer to fill out a questionnaire about their experience and the agent's performance. There were nine items, each scored on a seven-point Likert scale, which measured:

Knowledge: "How would you rate your level of knowledge about this topic after interacting with the system?"

Interest: "How would you rate your interest in this topic?"

Difficulty: "How hard did you have to work to accomplish this level of conversation?"

Fun: "This experience was fun", to assess emotional response.

Control: "To what degree did you control the conversation?"

Comfort: "To what degree was the interaction comfortable?"

"Click": one item to assess the likeability of the agent: "To what degree did you "click" with the agent?"

Anxiety/tension: "To what degree did the interaction make you feel anxious or tense?"

Mental demand: "How mentally demanding was the task you just completed?"

At exit. After all four tasks were finished, we asked participants to complete a final questionnaire. These canvassed demographics; personality; and preferred conversation style. Participants were also asked to rate the agent on a number of items, and invited to leave any final comments. The demographic and personality items are not analysed in the present work.

Eighteen questions asked about our participants' preferred **conversational style**. Six of these questions were drawn from unpublished work from Tannen's lab [personal communication] and used a Likert scale from 1 (strongly disagree) to 7 (strongly agree):

- (1) When I'm having a conversation with someone, and they're talking at the same time as me, I consider them enthusiastic.
- (2) ...and they're talking at the same time as me, I consider them rude.
- (3) ...and they're pausing to let me speak, I consider them polite.
- (4) ...and they're pausing to let me speak, I consider them boring.
- (5) ...I want them to be talking at the same time as me.
- (6) ...I do not want them to be talking at the same time as me.

Items 1, 4, and 6 capture involvement; items 2, 3, and 5 capture consideration. We believe these questions have not been validated, so also used another twelve items due to Shamekhi et al. [31]:

- (1) When I'm having a conversation with someone, and we're talking at the same time, I feel comfortable.
- (2) ...I find the other person a pleasure to talk with.
- (3) ...I find the other person pushy or aggressive.
- (4) ...I find the conversation fun.

- (5) ...I find the conversation overwhelming.
- (6) When I'm having a conversation with someone, and we trade off speaking while the other person pauses to listen, I feel comfortable.
- (7) ...I find the other person a pleasure to talk with.
- (8) ...I find the other person distant or boring.
- (9) ...I find the conversation fun.
- (10) ...I find the conversation boring.
- (11) When I'm having a conversation with someone, I want us to be talking at the same time.
- (12) ...I do not want us to be talking at the same time.

Items 1, 2, 4, 8, 10, and 11 capture involvement, and items 3, 5, 6, 7, 9, and 12 capture consideration.

A final set of items asked participants to **rate the agent** overall. Each item was a five-point scale, anchored by an adjective at either end: fake (1)–natural (5); machinelike–humanlike; unconscious–conscious; unfriendly–friendly; awful–nice; unkind–kind; artificial–lifelike; unpleasant–pleasant; dislike–like; incompetent–competent; ignorant–knowledgeable; irresponsible–responsible; unintelligent–intelligent; and foolish–sensible.

4 GENERAL OBSERVATIONS

In total we had 24 participants undertake 4 tasks each, for a total of 96 recordings. Tasks ran for an average 42.8 utterances (median 42; half each from agents and participants) and an average 8 minutes, 50 seconds (median 7:09).

4.1 Impressions of the Agent

Our exit questions included 14 items where participants were asked to rate the agent overall. Unsurprisingly, these indicate that participants did not find the conversation particularly lifelike: mean responses were 2.1/5 on the scale artificial–lifelike, 2.4 on fake–natural, and 1.8 on machinelike–humanlike; participants did however give a surprisingly high mean 2.8 on the scale unconscious–conscious. (For all exit questions, responses were roughly normal with mean \approx median \approx mode.)

The agent was seen as fair at its job, with means of 2.5 (ignorant–knowledgeable), 3.2 (irresponsible–responsible), 2.8 (unintelligent–intelligent) and 2.4 (incompetent–competent); and it was somewhat pleasant to deal with (mean 3.6 awful–nice, 2.9 dislike–like, 3.6 unkind–kind, 3.6 unpleasant–pleasant). We conclude that, although we are still a long way from lifelike conversation, the agent was at least minimally competent and somewhat pleasant to deal with. We might therefore expect that some social norms from human-to-human conversation could carry over to this setting.

4.2 Conversational Performance

Nine questions, each on a seven-point Likert scale, were administered at the end of each task to ask participants' opinion of the conversation they just had. Average responses indicate moderate performance: knowledge gain scored a mean of 4.1/7, interest in the topic was 4.5, hardness of the task was 4.8, fun in the task was 4.5, feeling in control was 5.1, feeling comfortable scored 4.0, "clicking" with the agent scored 3.4, feeling anxious or tense rated 5.0, and mental demand was good at average of 3.6/7.

These scores showed little difference from task to task, except that the introduction felt less comfortable and was viewed as more difficult (comfortable: mean 3.1/7 compared to 4.3/7 for the other tasks, ANOVA $F(3, 92) = 5.6, p \approx 0.001$; hard: mean 5.6 c.f. 4.6, $F(3, 92) = 8.8, p < 0.0001$). This may be explained by the introduction being first, or simply that the participants still weren't familiar with interacting with the agent. After the first task, we saw no significant effects due to task order—it does not appear that scores changed as participants got more experienced (or more fatigued).

The effect of chit-chat. Our chit-chat model was used to generate responses whenever the classifier was unsure of an intent—for example, whenever the participant strayed from the topical model, when the model was incomplete, or when automatic speech recognition introduced errors. This was common, and 75% of all responses were drawn from the chit-chat model.

Although the chit-chat model could and did produce high-quality responses on many occasions (see e.g., line 16 of Figure 2), it could also produce uninformative fillers (line 5) or non-sequiturs (line 7). We therefore expected that participants would report less satisfaction when chit-chat was invoked more often. This is indeed what we observed: for every 1% of agent responses drawn from chit-chat, self-reported gain in knowledge drops by 0.038 points (of 7); difficulty increases by 0.028 points; "clicking" with the agent drops by 0.031 points; and mental demand increases by 0.026 points. ($F(1, 93) \geq 4.4, p \leq 0.04$ in each case.) We saw no effect on other items. These are moderate effects: a change of 10% more or less chit-chat, about two utterances in a task, corresponds to around a third of a point shift in user perception on a seven-point scale.

The effect of agent disfluencies. Each task was manually annotated for errors and disfluencies: we noted all instances of illogical or off-topic utterances from the agent, freezes, no-answers, or self-talk from the agent, interruptions of the agent by the participant or vice versa, repetitions by the participant, or off-topic utterances from the human. Each of these represents some evidence that the conversation is going awry. From this we derived the time of the first illogical response as well as the total number of errors in the task; each task had at least one such error.

Again, we can observe an effect on the post-task ratings. Self-assessments of difficulty and comfort correlate with the time to the first illogical response: each second before the response corresponds to -0.0007 points of difficulty ($F(1, 93) = 7, p = 0.01$) and 0.0009 points of comfort ($F(1, 93) = 10, p = 0.002$). The longer the agent could go without an illogical response, the easier the task is and the more comfortable our participants felt.

The total number of disfluencies or errors in a task made a bigger difference, impacting ratings for interest (-0.1 points per error), difficulty (+0.1), control (-0.1), and comfort (-0.1; all ANOVA $p < 0.05$). Again, these are moderate effects given the number of possible disfluencies in a long conversation.

5 IDENTIFYING AND MEASURING STYLE

In analyses of the MISC dataset of information-seeking conversations [37], Thomas et al. [36] observed correlations between eleven stylistic variables and identified the underlying factor with Tannen's

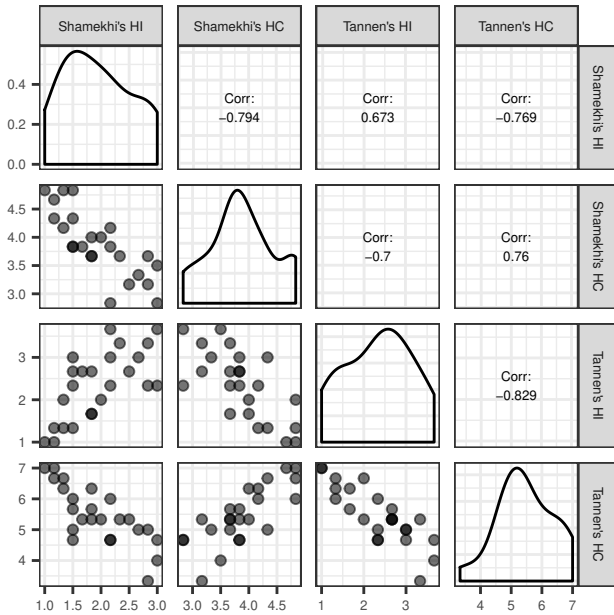


Figure 4: Self-reported style, as measured by high involvement (HI) and high consideration (HC) items on two different scales. In each case, HI and HC anti-correlate; the two scales agree; and there is an approximately normal distribution of scores.

“involvement”. Using this formulation it was possible to compute differences between speakers and their rate of alignment.

That work was motivated by the assumption that, since we often treat machines as people [24], stylistic properties of human-to-human conversation would be applicable to human-to-agent conversation. Our present data lets us ask our RQ1: is that in fact true? Are the style factors identified elsewhere still apparent in conversations with agents?

5.1 Self-Reports

Our exit questions included six Likert-type items from Tannen’s lab [pers. comm.]. These were aggregated into two scales, which measured our participants’ preference for “high involvement” or “high consideration” style. We also used a further twelve items from Shamekhi et al. [31], also on two scales, which have been validated against self-reported criterion variables. In each case, we took the mean scores over relevant items to produce single values for high involvement (HI) and high consideration (HC).

Tannen [34] suggests a partition between high involvement and high consideration styles, such that any person expresses one style or the other. In common with Thomas et al. [36], however, we see a continuum with most participants somewhere in the middle (Figure 4). Tannen’s HI and HC scores anti-correlate, as do Shamekhi et al.’s ($r = -0.77, -0.79$), and Tannen’s HI and HC scores correlate with Shamekhi et al.’s ($r = 0.67, 0.76$). Cronbach’s $\alpha = 0.89$ on Tannen’s six items, 0.91 on Shamekhi et al.’s twelve items, and 0.92 on all eighteen considered together.

Variable	Load
<i>People</i> : pronoun use	−0.32
<i>Rate of speech</i> : words/s	0.26
<i>Pauses</i> : words/utterance	0.33
<i>Phonology</i> : variance in pitch	−0.18
variance in loudness	−0.08
<i>Re-statement</i> : repeated terms	0.56
utterances with repeats	0.60

Table 1: Variables derived from phonology and transcripts, and their loadings on the first principal component.

We conclude that, on self-reported data, we again see evidence of a single underlying factor ranging from high involvement to high consideration. We next ask whether we can detect this preference automatically, in real time and in a running agent.

5.2 Automatically Deriving “Style”

Past work on human-to-human conversation identified a single factor underlying such variables as rate of speech, repetition, and variance in pitch, and it is possible to identify this with Tannen’s consideration/involvement dimension [36]. If the same habits of speech apply when people are talking to software agents as when they are talking to other humans, we should see something similar in the style data described above.

The seven style items do seem to be measuring a consistent variable (Cronbach’s $\alpha = 0.59$, Gutmann’s $\lambda_6 = 0.70$ over all 1716 utterances with data). A principal components analysis suggested one important component, accounting for 33% of variance, and this is summarised in Table 1. Rate of speech, length of utterance, repeated terms per utterance, and fraction of utterances with repeats load positively (that is, they correlate with the underlying factor). Pronoun use and prosodic variables load negatively (anti-correlate).

We are using a subset of the variables in Thomas et al. but the loadings here are very different. Thomas et al. saw all of these variables correlate with an underlying component, and could call this “involvement”; here, however, pronoun use and prosody (which we might expect to indicate involvement) anti-correlates with rate of speech and repetitions (also involvement). The loading on loudness is low, but the other variables are harder to explain.

Part of the reason may be the differences in protocol. In particular, here we have a software agent interlocutor, not a person; and there is a clear asymmetry of roles between seeker and intermediary which was much less pronounced in the earlier work. That, we expect, will give rise to a much lower feeling of belonging and hence different use of pronouns (and indeed first person pronouns dominate our recordings). We might also expect more stilted language, as people consciously speak so as to be understood by the computer. We also note a difference in processing, as we are tracking style in near-real-time rather than at the end of a task.

Regardless of the differences, our analysis does suggest an underlying “style” factor. Since the seven variables do seem to be measuring a single construct—at least in part—we formed a single “style” variable for each utterance by weighting each component according to Table 1 and summing. The resulting variable, like the

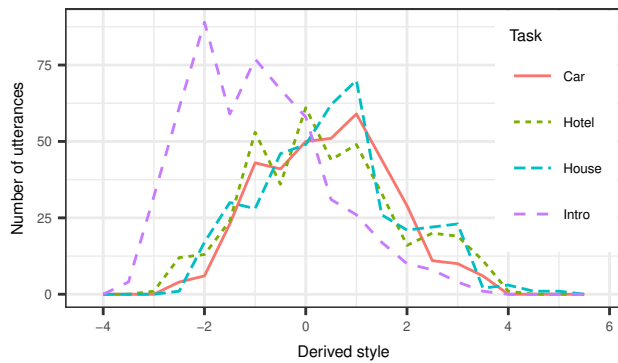


Figure 5: The distribution of our derived “style” variable. Similarly to self-reports, it is approximately normal.

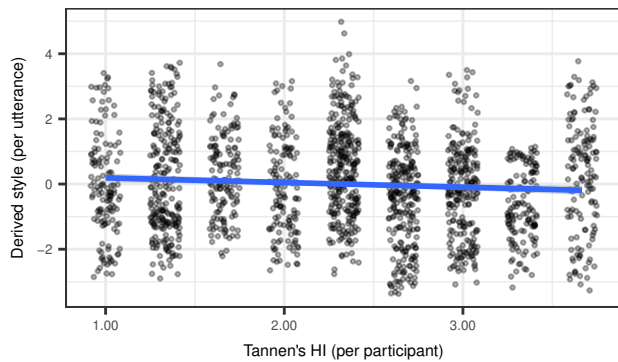


Figure 6: Derived “style” per utterance, against involvement according to Tannen’s instruments. There is only a very weak correlation: $r = -0.07$, $p = 0.002$.

HI/HC variables, is approximately normally distributed with some variation according to task (Figure 5).

5.3 Comparing Derived Style to Self-Reports

We have self-reported data for each participant, on the one hand; and derived style data for each five-utterance window, on the other hand. We can now ask whether these two instruments measure the same phenomenon. That is: can we use pitch, repetition, and so on to measure the involvement–consideration dimension in real time?

Figure 6 plots, for participants with each HI score according to Tannen’s instrument, the distribution of utterance styles. There is no clear pattern. In particular, it seems that derived “style” correlates with self-declared preference only very weakly. Per-utterance derived style correlates with self-reported HI (on Tannen’s scale) with Pearson’s $r = -0.07$, $p = 0.002$. Similarly weak correlation was seen against self-reported HI and HC on Shamekhi’s scales ($r = 0.06$ and -0.05 , $p = 0.02$ and 0.04).

The weakness and inconsistency of these correlations lead us to believe there is no real connection between Tannen’s involvement–consideration dimension and the style measured here.

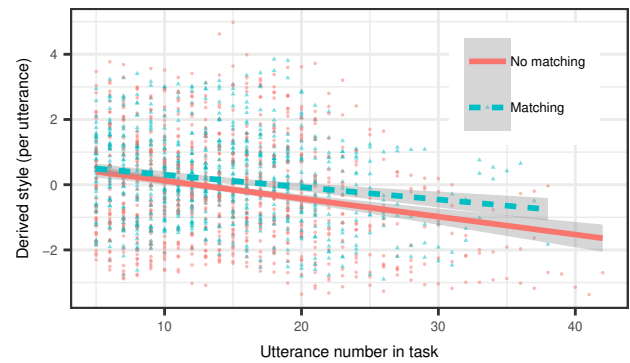


Figure 7: Derived style, per utterance, as tasks progress. On average the value decreases, but it decreases less sharply when the agent is in style matching mode.

5.4 Summary

Our results show that while there is recognisable variation in style in our human-agent interactions, it is not the same as observed in human-human interactions and in particular cannot be identified with Tannen’s “involvement” and “consideration”. Nevertheless, the derived “style” is internally consistent and we can, as before, use a single variable to capture this.

6 PARTICIPANTS’ STYLES

It is possible that the agent’s style—even allowing for our attempts at matching—suited some participants better than others. It is also possible that style reflects personality or other preferences, which in turn make the agent a better match for some participants. Having established that there is a consistent “style” which we can derive, we now ask our RQ2: does this style correlate with our post-task or exit instruments? That is, do people’s experiences of the tasks or opinions of the agent vary with their style?

To answer this we took the median derived style over a whole task, as a per-task variable, and compared it to our post-task items. We saw modest effects on a participants’ sense of control and comfort (0.26 and 0.24 points per point of “style”, $F(1, 94) = 6.1$ and 4.3 , $p = 0.02$ and 0.04). There was no significant effect on any other post-task items, nor was there any relationship between derived style and any of our post-experiment ratings of the agent.

We also saw little difference in post-task scores based on self-reported style. There was a significant relation only with self-reported levels of anxiety or tension, which decreased 0.43 points per point on Tannen’s HI scale ($F(1, 94) = 4.3$, $p = 0.04$). Again, we saw no difference on any of the post-experiment ratings.

7 ALIGNMENT TO THE AGENT

Past work [8, 22, 27] suggests that alignment should happen in human-computer interaction as it does between humans, so we would expect to see a similar effect here. (This is our RQ3.)

Figure 7 plots, for each utterance in our recordings, the calculated style against the utterance number within the task. Our agent has a slower and relatively flat speaking style, with little prosodic

variation or rapid speech. We would expect to see participants' rate of speech and vocal variation decrease over time, as they aligned with the agent, and this is indeed the case.

We can also see, however, a difference due to the agent's matching: for tasks where the agent is matching participants' styles, the participants' own style changes less (effect of utterance number -0.055 , effect of interaction between utterance number and matching 0.012). That is, when the agent is "meeting participants in the middle", participants changed their own style about 22% slower. This effect is modest but statistically significant (utterance number $F(2, 1713) = 86.5$, $p \ll 0.0001$, interaction $F(2, 1713) = 15.4$, $p < 0.0001$). If anything, the results suggest our agent could have varied its style more, which may have created a stronger effect.

8 ALIGNMENT TO THE PARTICIPANT

In half of the tasks—two per participant—our agent was in "style matching" mode and aligned, somewhat coarsely but in real time, to the participants' detected style (Section 3.4).

There is evidence that people prefer agents that match their conversational style. Shamekhi et al. [31] reported a slight preference for agents which matched participants on the involvement-consideration dimension, although unlike our work they used only two levels ("HI" and "HC"). The agent used by Shamekhi et al. also used hand-written responses and a short, hard-coded conversation, so the agent's two styles were distinctive. Despite the differences in our setup, since the underlying phenomena are general, we might expect similar effects here; this was our fourth question from §1. If so—and if the derived style above is valid, despite not being identical to Tannen's "involvement"—we would expect the agent to do better in style matching mode.

This is indeed what we observe. Amongst the post-task items, there is a significant difference in participants' sense of control between the matching and non-matching conditions: on average, participants scored 0.52 points higher on a scale of 1–7 (Welch's two-sample $t(93.7) = 2.04$, $p = 0.02$). Style matching also leads to higher scores for interest in the topic (mean 0.54 points higher, $t(93.3) = 1.75$, $p = 0.04$). This is consistent with results from Shamekhi et al. [31], who saw an effect only on "engaging interaction".

Many of the effects noted in Section 4.2 were also attenuated when the agent was matching styles. Falling back to the chit-chat model had an effect on self-reported knowledge gain, "clicking" with the agent, and mental demand. All these effects were reduced when the agent is matching, by up to 50%, but none of these interaction effects reach statistical significance at $\alpha = 0.05$. Similarly, disfluencies were observed to impact difficulty, comfort, interest, and control: all of these effects were also attenuated, by up to 77%, in style matching mode. None of these interactions, however, reached statistical significance except for feeling in control of the conversation $F(3, 92) = 4.7$, $p = 0.03$, 77% reduction in effect size).

We must emphasise that our current software has only a crude notion of matching; and indeed we may not have captured what is most important about style. We are also chasing a subtle effect. The results above give us some confidence that alignment will be as beneficial in human-to-agent as human-to-human conversations, and that conversational agents should include this feature.

9 CONCLUSIONS

These are, to our knowledge, the first experiments in long-form information-seeking conversations with an embodied software agent. We have long conversations—42 turns and 8 minutes each, on average—and our agent can fall back to chit-chat to keep the conversation moving. This in turn has allowed us to ask whether the conversational styles seen in human-to-human conversation are also seen in conversations with an embodied software agent, what the effect is of each party's style, and whether we see alignment.

Using prosodic and lexical variables inspired by Tannen [34], we do indeed see an underlying style component in our recordings. This "derived style" is not the same as Tannen's, and some components look different to those of Thomas et al.. Nevertheless, this derived style can be tracked and matched by an agent.

In our experiment, people had different experiences of the tasks according to their self-reported (HI or HC) style, and also according to their derived style. There is no corresponding difference in their ratings of the agent. We do, however, see evidence that people align to the agent's style, and that this alignment happens slower if the agent is simultaneously matching the human. The agent's matching also has an effect on reported knowledge gain, "clicking" with the agent, and mental demand, and these effects are larger than the effect of the participants' own style. Matching also seems to have an effect on how people perceive disfluencies or chit-chat.

The effect of alignment is subtle and in our experiments it does not always reach the threshold of statistical significance. (We have performed many hypothesis tests, so there is as well an increased risk of false positives.) However, even in this constrained setup we have seen effects, these effects are consistent with each other, and these effects are consistent with observations in the literature. We also note that the effects of matching are larger, and more often seen, than the effects of participants' own style. This is highly suggestive and motivates continued research, with more responsive and capable agents.

Caveats. The introduction task is more open-ended than the others, and was always given first, so we might expect some differences in our analyses. We have run all the analyses here with the first task excluded, and saw the same effects. In particular, derived "style" and its components look similar; there is still no difference due to the user's own style; and the effects of matching are consistent, although of course the effect sizes are slightly different. We are therefore confident the results are robust to this task difference.

We used an embodied agent in this experiment, to better induce social responses, but we must acknowledge extra complications—importantly, users are less familiar with the interface. We believe we would see similar effects with an audio-only or text-only interface [e.g. 18], but this is not tested in the current work.

Although relatively small-scale, this study provides converging lines of evidence that "conversation style" is real, even in task-driven conversations with a software agent; and also that attending to and matching this style makes for a better conversation. We believe this will become increasingly important as agents gain competence and are more capable of longer exchanges. As well as *what* agents say, we should also be paying attention to just *how* they are saying it.

ACKNOWLEDGMENTS

We are very grateful to Kael Rowan, Deepali Aneja and Rens Hoegen for their contributions to the avatar framework. We thank our participants for their time.

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] 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*.
- [5] N. J. Belkin. 1987. Discourse analysis of human information interaction for specification of human-computer information interaction. *Canadian Journal of Information Science* 12, 3–4 (1987), 31–42.
- [6] M. M. Berg. 2014. Modelling of natural dialogues in the context of speech-based information and control systems. PhD thesis, University of Kiel. (2014).
- [7] T. Bickmore and J. Cassell. 2005. Social dialogue with embodied conversational agents. In *Advances in natural multimodal dialogue systems*, Jan C J Kuppevelt, Laila Dybkjær, and Niels Ole Bernsen (Eds.). Springer.
- [8] H. P. Branigan, M. J. Pickering, J. Pearson, and J. F. McLean. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics* 42 (2010), 2355–2368.
- [9] J. Cassell. 2001. Embodied conversational agents: representation and intelligence in user interfaces. *AI magazine* 22, 4 (2001), 67–67.
- [10] J. Cassell, J. Sullivan, E. Churchill, and S. Prevost. 2000. *Embodied conversational agents*. MIT press.
- [11] P. J. Daniels, H. M. Brooks, and N. J. Belkin. 1985. Using problem structures for driving human-computer dialogues. In *RLAO-85: Actes: Recherche d'Informations Assistée par Ordinateur*. 645–660.
- [12] M. Dubiel, M. Halvey, L. Azzopardi, and S. Daronnat. 2018. Investigating how conversational search agents affect user's behaviour, performance and search experience. In *Proc. Int. W'shop on Conversational Approaches to Information Retrieval*.
- [13] P. Edwards, C. Landreth, E. Fiume, and K. Singh. 2016. JALI: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 127.
- [14] F. Eyben, F. Weninger, F. Gross, and B. Schuller. 2013. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In *Proc. ACM Multimedia*. ACM Press, 835–838.
- [15] R. Fusaroli and K. Tylén. 2015. Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive Science* (2015), 1–27.
- [16] 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.
- [17] 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.
- [18] R. Hoegen, D. Aneja, D. McDuff, and M. Czerwinski. 2019. An end-to-end conversational style matching agent. In *ProcIntelligent Virtual Agents*. 111–118.
- [19] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky. 2006. Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Vol. 1. IEEE, 1–1.
- [20] 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.
- [21] S. Kim, L. F. D'Haro, R. E. Banchs, J. Williams, M. Henderson, and K. Yoshino. 2016. Dialog state tracking challenge 5 handbook. (2016). Retrieved February 2017 from https://github.com/seokhwankim/dstc5/raw/master/docs/handbook_DSTC5.pdf
- [22] 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.
- [23] R. T. Lakoff. 1979. Stylistic strategies within a grammar of style. *Annals of the New York Academy of Sciences* 327, 1 (1979), 53–78.
- [24] C. Nass and Y. Moon. 2000. Machines and mindlessness: Social responses to computers. 56, 1 (2000), 81–103.
- [25] 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.
- [26] R. Nordlie. 1999. “User revelation”—A comparison of initial queries and ensuing question development in online searching and in human reference interactions. 11–18.
- [27] M. J. Pickering and S. Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences* 27, 2 (2004), 169–225.
- [28] F. Radlinski and N. Craswell. 2017. A theoretical framework for conversational search. In *Proc. CHIIR*. ACM, 117–126.
- [29] R. Reichman. 1985. *Getting computers to talk like you and me*. MIT Press, Cambridge, Massachusetts.
- [30] T. Saracevic, A. Spink, and M.-M. Wu. 1997. Users and intermediaries in information retrieval: What are they talking about?. In *User Modeling*. Springer, 43–54.
- [31] 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.
- [32] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714* (2015).
- [33] SRI International. 2011. SRI's Amex Travel Agent Data. (2011). Retrieved June 2016 from <http://www.ai.sri.com/~communic/amex/amex.html>
- [34] D. Tannen. 1987. Conversational style. In *Psycholinguistic models of production*, Hans W Dechert and Manfred Raupach (Eds.). Ablex, Norwood, NJ.
- [35] D. Tannen. 2005. *Conversational style: Analyzing talk among friends* (new ed.). Oxford University Press, New York.
- [36] P. Thomas, M. Czerwinski, D. McDuff, N. Craswell, and G. Mark. 2018. Style and alignment in information-seeking conversation. In *Proc. CHIIR*. 42–51.
- [37] 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*.
- [38] 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.