Style and Alignment in Information-Seeking Conversation

Paul Thomas
Microsoft
Canberra, Australia
pathom@microsoft.com

Mary Czerwinski
Microsoft
Redmond, WA, USA
marycz@microsoft.com

Daniel McDuff
Microsoft
Redmond, WA, USA
damcduff@microsoft.com

Nick Craswell
Microsoft
Bellevue, WA, USA
nickcr@microsoft.com

Gloria Mark*
University of California, Irvine
USA
gmark@uci.edu

ABSTRACT

Analysis of casual chit-chat indicates that differences in *conversational style*—the way things are said—can significantly impact a participants' impressions of the conversation and of each other. However, prior work has not systematically analyzed how important style is in task-oriented, information-seeking exchanges of the sort we might have with a conversational search agent. We examine recordings from the MISC data set, where pairs of "users" and "intermediaries" collaborate on information-seeking tasks, and look for indications of style which can be computed at scale.

We find that stylistic markers identified by Tannen in casual chat do exist in information-seeking dialogue, and that participants can be arranged along a single stylistic dimension: "considerate" to "involved". This labelling for style needs no manual intervention. Furthermore, we find that there is no clear best style; but that differences in style, previously thought to impede communication, are only a problem for shorter tasks. This result is likely due to alignment of conversational style over the course of an interaction.

ACM Reference Format:
Paul Thomas, Mary Czerwinski, Daniel McDuff, Nick Craswell, and Gloria Mark. 2018. Style and Alignment in Information-Seeking Conversation. In *CHIIR '18: 2018 Conference on Human Information Interaction & Retrieval, March 11–15, 2018, New Brunswick, NJ, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3176349.3176388

1 INFORMATION-SEEKING CONVERSATION

Recent years have seen a dramatic rise in digital personal assistants such as Alexa, Siri, Cortana, and Google Now, as well as "bots"—software agents interacting in natural language—on messaging platforms such as Messenger, Skype, and Sina Weibo. Such conversational agents are attracting more investment, are gaining capabilities, and are being increasingly used[1]. These agents offer a range of services such as device control (e.g., for making calls), closed-domain task completion (e.g., setting reminders or looking up timetables), and factoid lookup. Importantly, they also support *information-seeking conversation*: multi-turn interactions, in natural language, where the user is looking for information rather than trying to complete a small task.

The information retrieval literature has few ways to systematically describe such a conversation. Time, number of turns, or task success are relatively simple to measure but do not describe well the interactant's experience. We must also consider the *style* of a conversation: was it pleasant?, abrupt?, confusing?, courteous? Visual design provides an analogy: a good visual design may or may not improve efficiency or effectiveness, but it will certainly improve the feel of the overall interaction.

For example, Figure 1 provides three extracts from the Microsoft Information-Seeking Converasation collection [25], recorded between pairs of volunteers working on search tasks. Although the same task is being addressed in each case, clearly the participants have different styles. The first is much more formal and concise, the second verbose with more description, and the third more verbose and informal still with "thinking aloud". People may react very differently to these conversations, regardless of the information being discussed or their overall success. There might not be a single "best" style; in fact prior work would suggest that style needs to adapt to different preferences [3] and cultural norms [20].

It is not yet clear just how important style may be for conversational agents, relative to task performance or other factors. However, we argue that conversational style should be considered in design. Consider, for example, a choice between two conversational agents supporting travel, both on the same channel. If either agent can provide the same information, in about the same time, but one is pleasant to deal with while the other is unemotional, or even rude (e.g. abrupt, or confusing), then the first is clearly preferred.

We expect that in the near future, software agents will be able to maintain a conversation to several turns or even several minutes, and that information-seeking tasks will be more important as this capability develops. Our research questions in this work are, therefore: *can we distinguish different conversational styles, in an information-seeking context and when working with an agent?* If so,

*Work carried out at Microsoft.

CHIIR '18, March 11–15, 2018, New Brunswick, NJ, USA

ACM ISBN 978-1-4503-4925-3/18/03…$15.00
https://doi.org/10.1145/3176349.3176388

[1] See for example http://www.nytimes.com/2015/08/04/science/for-sympathetic-ear-more-chinese-turn-to-smartphone-program.html (Xiaoice); http://venturebeat.com/2016/06/30/facebook-messenger-now-has-11000-chatbots-for-you-to-try/ (Messenger).

User: The two possible treatments for migraine headaches that I want to do first are beta-blockers…
Intermediary: Uh huh
User: …and calcium channel blockers.
Intermediary: Calcium channel blockers.
Ok, so lets start with beta-blockers.
So, beta-blockers are commonly used for treating high blood pressure and other heart issues and are also prescribed to prevent migraines.

(a) Very efficient and to-the-point—little information about what the intermediary is thinking (participants 23/24).

User: So that's my problem, to discuss beta-blockers, calcium channel blockers and diet an exercise as an option
Intermediary: Ok … for migraine headaches … it just says in the … umm … it's not even really a result … it's from migraine.com.
Beta-blockers are commonly used for treating high blood pressure and other heart issues are prescribed to prevent migraines.
Beta-blockers are some times called beta-… eh, well that doesn't matter.

(b) Slightly more verbose language and description of what the searcher is doing—thinking aloud (participants 21/22).

User: I need to research beta-blockers and calcium channel blockers … um, I guess as to their applicability to migraines … and their effectiveness to migraines.
And then after that explore other options, if I don't want to take medicines.
I guess I'd just look for beta-blockers.
Intermediary: (LONG PAUSE)
Yeah … I just go beta-blockers, migraine prevention here … I'm trying to find a vaguely reputable site to go with …
(LONG PAUSE)
I found something called American Family Physician that I have never heard of, I want to go back to WebMD—that can kinda be sketchy but should give some sources …
(LONG PAUSE)
In generally it says beta-blockers work to relax the blood vessels and it is not clear how they work to prevent migraines. …
It says beta-blockers have been shown to prevent migraines.

(c) Very verbose language and description of what the searcher is doing—lots of thinking aloud and informal language, long pauses (participants 26/27).

Figure 1: Transcripts from the Microsoft Information-Seeking Conversation recordings [25], showing three different styles of conversation.

does conversational style influence perceptions of that conversation? In particular, *does similarity or difference in style influence feelings of effort and engagement?*

2 CONVERSATIONAL STYLE

Conversational style, the way people behave in conversation, is not well understood in information retrieval. In this study, we follow the work of Tannen, who 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" [24, p.188]. "Style" thus includes prosody, word choice, turn-taking, and timing, for example. We distinguish style, the "how", from any topical information transferred, the "what"; we can provide the same information with very different styles [2].

"Style", in various forms, has been considered at length but almost entirely in natural, casual, informal conversation rather than in goal-directed settings or in conversation with agents (human or machine). We discuss some of this work below.

2.1 Involvement and consideration

A key example is long-running work by Tannen [24]. This draws on tape recordings of dinner-party conversation amongst friends, with Tannen as a participant researcher. 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 "considerateness-involvement continuum" amongst the guests. There are no firm rules, but speakers in this model are stereotypically divided into two camps or styles. Both styles try to build rapport with an interlocutor, but they do so by emphasising different "rules" of conversation, different aspects of face [6], and different strategies for presentation [12].

Tannen's "high involvement" style is summarised as one which emphasises interpersonal involvement, interest, approval, understanding, and community. It overlaps with Lakoff's "camaraderie" strategy [12] and a need for positive face:

> *When in doubt, talk. Ask questions. Talk fast, loud, soon. Overlap. Show enthusiasm. Prefer personal topics [23].*

The "high consideration" style, on the other hand, is defined by an emphasis on consideration and independence. It overlaps Lakoff's "distance" strategy and a desire to maintain negative face:

> *Allow longer pauses. Hesitate. Don't impose one's topics, ideas, personal information. Use moderate paralinguistic effects. [23]*

Table 1 lists the characteristics of conversational style, on Tannen's summary [23], and the variables we use in this work.

From her analysis of conversations, Tannen suggests that partners with different styles have more trouble communicating. For example, a high-consideration speaker may find a high-involvement speaker to be pushy or a high-involvement speaker may find their opposite partner standoffish: "the use of … devices that are not understood or expected creates a sense of dissonance, which often leads to negative or mistaken judgements … This, in turn, leads one to walk away from an encounter feeling dissatisfied or disgruntled." [24]

2.2 Style and agents

If we are building conversational search agents, then many of these aspects of style are under our control. If conversational style, or differences in style, make a difference in this setting then we should consider adapting or monitoring agents accordingly. We are not aware of any work which discusses style in information seeking—this is our goal here—but work in other settings suggests that expressions of agent "personality" and style can make a difference.

Category	Characteristics per Tannen	Variable(s) used here
Topic	Prefer personal topics	Pronoun use (ppron)
	Persistence	Repetition (rept, repu)
	Shift topics abruptly	—
	Introduce topics without hesitance	—
Pace	Faster rate	Rate (wps, wpp, wpu)
	Pauses avoided	Pause length (boplen, poplen)
	Faster rate of turntaking	Pause length (poplen)
	Cooperative overlap	Overlap rate (olap)
Expressive paralinguistics	Pitch shifts	Pitch variation (pv)
	Loudness shifts	Loudness variation (lv)
	Marked voice quality	—
	Strategic pauses	—
Genre	Tell more stories	—
	Tell stories in rounds	—
	Point of stories is emotion of teller	—

Table 1: Tannen's characteristics of conversational style [23], and the variables used in this work. Our variables were selected for ease of automation and do not address genre, but otherwise have good coverage. Variables are detailed in Section 3.2.

People "mindlessly" apply human social rules when interacting with computers, including preferring those which appear to manifest personalities similar to their own [14] and preferring to interact with agents that are more human-like [3]. Evidence from work in human-computer interaction also suggests conversational style is likely to be important in our context. For example, Shamekhi et al. [22] gave crowdsourced workers two "agents", each constructed to exhibit high involvement or high consideration by varying both prosody and script. Participants were asked to respond as the "agents" offered to arrange meetings or ask short questions. Participants tended to prefer the agent whose style matched their own: however, participants' styles were not measured directly, and "agents" followed short, fixed, scripts so interaction was not natural.

3 DATA AND ANALYSIS

We drew on the Microsoft Information-Seeking Conversation data set (MISC) [25] to address the questions above. MISC includes recordings of pairs of volunteer participants working together to solve information-seeking problems: in each pair, one "user" was assigned a sequence of information-seeking tasks but no web access; and one "intermediary" had access to the web on the user's behalf. The two participants were connected by an audio link. This is intended to mimic interactions with systems such as Siri or Cortana, but with a much more natural conversational style.

MISC includes time-series data for each participant and task, including basic prosodic signals and transcripts, which we build on below (Section 3.2). It also includes self-reports for effort, engagement, and opinion of the partner, which we use as dependent variables (Section 3.3). To our knowledge, MISC is the only data set with both recordings and self-reports.

We do not at present consider the effect of style on task completion or accuracy as the data to hand does not let us investigate these questions. All tasks were completed, to some degree, and since some tasks were subjective (finding options to match participants' own preferences) there is no notion of "correctness" to judge against.

3.1 Participants and tasks

MISC includes recordings of 22 pairs of participants, each working on five tasks with a ten-minute limit per task. The first task was a warm-up, and we excluded this from our analysis; thus we analyzed four tasks. The tasks varied in difficulty (availability of information) and complexity (cognitive load, or degree of comparison and synthesis required). Summarising Thomas et al. [25], the tasks were:

Heroism "…you want to find accounts of selfless heroic acts by individuals or small groups for the benefit of others or for a cause." (Low difficulty, low complexity.)

Migraine "…You heard about two possible treatments for migraine headaches, beta-blockers and/or calcium channel blockers, and you decided to do some research about them. At the same time, you want to explore whether there are other options for treating migraines without taking medicines, such as diet and exercise." (Low difficulty, high complexity.)

Olympics "…Find the venues of the 2024 Olympic Games and the 2016 Winter Olympic Games." (High difficulty, low complexity. There was no 2016 winter games, and at the time the venues for 2024 had not been announced.)

Transport "…your task is to decide on the best form of transportation between cities in North America that would be suitable for you [on a three-month holiday by public transport]." (High difficulty, high complexity.)

3.2 Stylistic variables

We considered aspects of conversational style discussed by Tannen [23, 24] and selected variables which reflected these, and which we could extract from the MISC data. As we are interested in the future in analysing conversational style at scale, we preferred variables which could be derived without manual intervention.

For each participant in each task, we derived eleven variables in six categories. One variable records the extent to which a participant spoke about themselves, their partner, or other people:

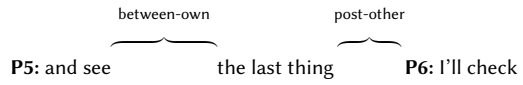

Figure 2: Between-own and post-other pauses. The pause between "see" and "the" is a *between-own* pause for speaker P5; the pause between "thing" and "I'll" is a *post-other* pause for speaker P6. (Participants 5 and 6, migraine task.)

ppron The rate of use of personal pronouns, including first- and second-, but not third-person pronouns. As we needed a finer grain than the Linguistic Inquiry and Word Count (LIWC [17]) analysis included in MISC, this was based on our own list of pronouns. Our list seemed to capture more terms than did LIWC.
The tasks in MISC were assigned, so participants were not completely free to choose whether or not they discussed personal topics. We might therefore expect this to be a relatively weak signal of style, but still indicative.

A single variable records a participant's rate of speech.

wps Speech rate, in words per second. This is an overall (micro-) average, calculated as the number of words in the transcript divided by the total duration of utterances in the transcript. An "utterance" here is simply a line of transcript (i.e., we relied on the speech-to-text system to determine boundaries; note that utterances may not alternate and one participant may have several utterances in a row).

A set of four variables consider pauses, turn-taking, and quickness of response. These variables are based on measurements of two types of pause, illustrated in Figure 2. A *between-own* pause is a period during a single utterance where there is no speech signal (periods where OpenSMILE [8] reports no F0), and a *post-other* pause is the gap between the end of one participant's utterance and the start of their partner's next. In marking post-other pauses, we allowed for cases where participants did not strictly alternate. The variables are:

wpu Mean words per utterance.

wpp Words per between-own pauses; approximately the length of each spoken phrase. Again this is an overall average, calculated as the number of words in the transcript divided by the total number of between-own pauses, and not a per-utterance value.

boplen Mean length of between-own pauses.

poplen Mean length of post-other pauses.

Style can also be indicated in relatively expressive (or flat) phonology, and two variables encode this.

pv Pitch variation, measured as the variance in F0 at those times when there is a speech signal according to OpenSMILE. Again this is per-person, per-task, i.e., this is the variance across the entire recording.

lv Loudness variation, measured the same way.

One variable counts overlap:

olap The proportion of utterances which initiate an overlap: that is, the proportion of one participant's utterances which begin while the other participant is still talking. This need not be an interruption in the usual sense, as overlaps commonly include utterances such as "uh-huh" which indicate agreement but let the partner continue.

Finally, we consider the degree to which topics are re-visited or requests re-stated. We approximate this with two variables counting repetitions.

rept Mean number of terms which are repeated from the same person's previous utterance. Before counting repeats we removed stopwords as well as "um", "uh", and "uh-huh", and stemmed what remained.

repu The fraction of utterances which included at least one repeated term, defined as above.

We did not derive or examine any other variables.

Our features are selected to be computable at scale and without manual intervention. However, they do give reasonable coverage of Tannen's indicators of style [23]. We include two aspects of topic ("prefer personal topics" and "persistence", but not "shift topics abruptly" or "introduce topics without hesitance", both of which would require very sophisticated modelling based on the variety of language used in MISC). We have good coverage of pace (all of "faster rate", "pauses avoided", "faster rate of turn-taking", "cooperative overlap"). and reasonable coverage of expressive paralinguistics (covering pitch and amplitude shifts, but not "marked voice quality" or "strategic pauses"). We do not have any coverage of genre, but we might not expect these signals as much in information-seeking as in chit-chat. We will observe in Section 4.1 that the variables are at any rate adequate for our purposes.

The eleven variables are on very different scales—for example, between-own pauses on the order of 0.1 s and pitch variation on the order of 1000 Hz^2—so in the analysis below they are each rescaled to zero mean, unit standard deviation. Inspection confirmed that each variable is approximately normal, although some have long tails.

3.3 Dependent variables

MISC includes self-report data of four kinds, and we use three here. The fourth, reported emotion, is out of scope for this paper and we leave it for future work:

TLX Effort was recorded using the NASA Task Load Index (TLX) [15], excluding the item on physical effort.

UES Engagement was recorded with a selection of items from O'Brien and Toms's User Engagement Scale (UES) [16]. The MISC data includes questions from the "novelty", "felt involvement", and "endurability" sub-scales.

Opinion MISC includes three custom items soliciting the participants' opinions of their partner: "the other participant helped me work on this task", "…understood what I needed", and "…communicated clearly".

All items were rated on seven-point, Likert-scales. For each of TLX, UES, and opinion, the set of associated items were consistent (GLB = 0.80, 0.89, 0.85),[2] so we used the mean of each set as an overall score. This gave us three dependent variables per participant,

[2]GLB is the "`glb.fa`" greatest lower bound of Revelle [21], which we use in this work as some of our variables are not symmetric and in general they may load differently. This follows Trizano-Hermosilla and Alvarado [26]. Corresponding Cronbach's α were 0.84, 0.85, 0.85, well within the 0.7 to 0.9 suggested by DeVellis [7].

per task. TLX and UES were well distributed (range 1–5.8 and 1–7 respectively, means 3.0 and 5.0), and opinion was clustered at the top of the scale (range 2–7, mean 6.3/7).

Each dependent variable varied with participant and with task. There was no apparent effect of sequence on TLX, UES, or opinion of other (one-way ANOVA $F(3, 161) = 1.7, 0.3, 0.8$; n.s.), that is there was no noticable effect of learning or fatigue on these variables.

3.4 Data cleaning

We examined all apparent outliers, across all the variables discussed above, and removed a small number of cases.

Participant 2, in the transport task, exhibited extremely long post-other pauses (mean more than 20 s, with two of almost 1 min out of only seven post-other pauses). These pauses were not only outliers in the entire set but uncharacteristic for the participant.

Participant 5, in the migraine and Olympics tasks, spoke slowly with very high overlap. A lot of the overlaps were utterances such as "um", "uh", and "uh-huh". Again this was an extreme case.

Finally, participants 17 and 18 did not engage in the transport task as expected. The transcript for the task shows participant 17 (the "user") deciding on an answer early on, with no attempt to search; and then almost a monologue on a variety of related topics. Participant 18, the "intermediary", never offers any information for this task and indeed only speaks 80 of the 1040 words in the data.

Participants in MISC were told they should stop their task, if they were still in conversation at the ten-minute mark. It is likely that this would influence any further conversation (e.g., it would be more rushed), as well as our dependent variables (e.g., participants may report less success), so we also removed any task that had recordings past the ten-minute mark.

This left us with observations from 98 participant-tasks, with 18–34 observations per task (median 23) and 1–4 observations per participant (median 2.5).

4 EVIDENCE FOR STYLE AND ITS EFFECT

On the basis of the variables above, we note that there is evidence for a single "style" dimension, which varies across people and tasks, and which is important to the overall impression of a conversation.

4.1 Making "involvement"

If there is a coherent consideration-involvement dimension in speech, *and* we have selected appropriate methods to capture this, then all eleven variables above would measure the same underlying construct. We tested this in two ways.

First, Revelle's GLB is 0.85 over all eleven variables and 168 observations, suggesting good reliability (corresponding Cronbach's $\alpha = 0.67$). This can be improved somewhat by dropping some variables; however, these improvements would be only marginal, so lacking any theoretical grounds for removing any variable we retained all eleven.

Second, a principal components analysis and scree plot suggested one more important component, explaining 29% of variance, and 2–3 less important components each explaining 14% or less. This first principal component is summarised in Table 2. It is easy to identify with Tannen's "involvement", so we did so: the *involvement* shown by a participant, in a task, is the sum of the eleven variables above weighted by their loadings on the first principal component. The final "involvement" variable is approximately normally distributed. (Some factors with low loadings on involvement do have higher loadings elsewhere, but the components are hard to interpret and vary from person to person much less than does involvement. They also explain relatively little variation, so we do not consider them here.)

Variable	Load
People: ppron	0.13
Rate of speech: wps	0.39
Pauses, turn-taking: wpu	0.45
wpp	0.44
boplen	−0.10
poplen	−0.27
Expressive phonology: pv	0.09
lv	0.21
Overlap: olap	−0.01
Re-statement: rept	0.39
repu	0.39

Table 2: Variables derived from phonology and transcripts, and their loading on "involvement". The sign of each is consistent with predictions, except overlap ratio (see text).

Of the eleven variables, each correlates (or anticorrelates) as expected with the exception of olap, the proportion of a person's utterances which overlap their partner. We expected this would align with rate of speech, use of pronouns, and repetition as an indicator of involvement; instead it aligns with pause length as an indicator of consideration. We note, however, that the loading is very small—the smallest of the eleven, and at −0.01 practically zero—so this reversal is of very little practical significance and overlap effectively carries no signal.

Tannen's "high-consideration"/"high-involvement" distinction was based on observation of a very different situation; and this is one possible instantiation of her description, chosen mainly for processing convenience. Despite this, the formulation used here does result in the same range of conversational styles. Where Tannen describes a division, however, we see more of a continuum.

There is little difference in involvement across roles or tasks (Figure 3), although on average users show somewhat more involvement than intermediaries (0.84 points difference in mean, one-sided $t(95.9) = 2.42$, $p < 0.05$).

We can now ask what might influence involvement, and whether in turn involvement influences how people experience information-seeking conversations.

4.2 Intermediaries' styles

Most software agents have a single style; and certainly people differ in styles, e.g., being more or less helpful, or more or less pleasant, perhaps regardless of who they are interacting with. It is possible that there is a conversational style, on the involvement-consideration axis, which is always (or normally) good. For example,

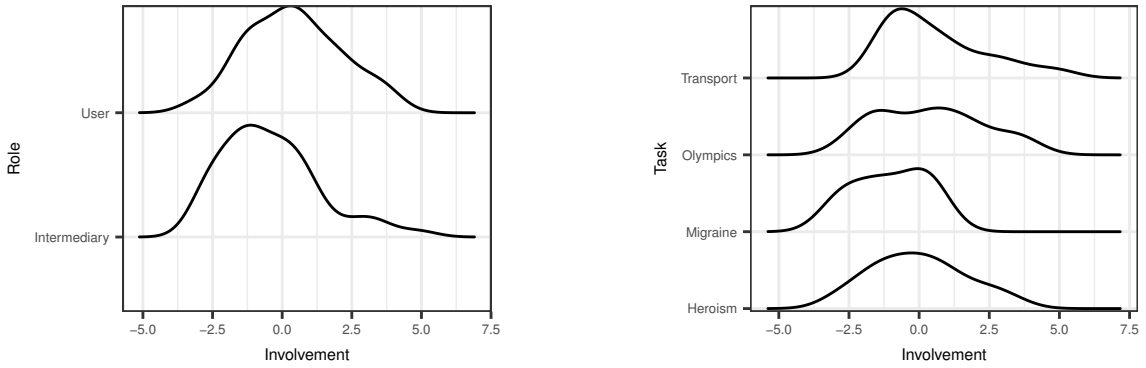

Figure 3: Conversational style is similarly distributed across participant role and task.

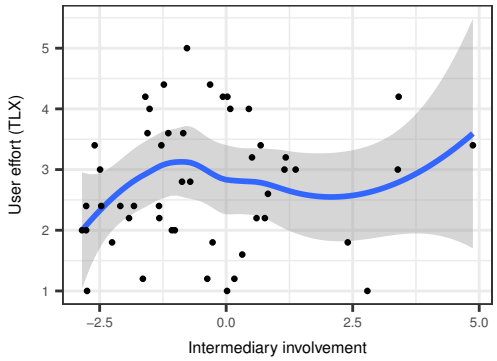

Figure 4: Intermediaries' style (consideration–involvement) and users' reported effort. There is no single "best style" for an intermediary.

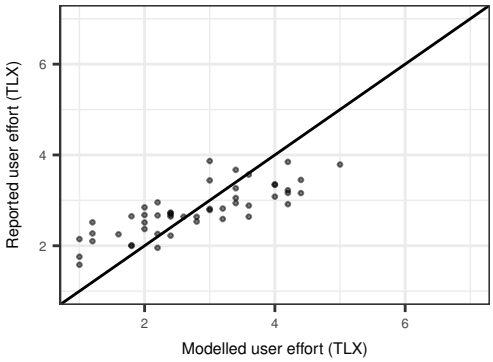

Figure 5: Modelling users' effort as a response to difference in style with random effects per-participant and per-task. Mean error 0.58 points out of 7.

we might believe that an agent should generally show consideration, and that it is inappropriate to display too much enthusiasm.

If there is in fact a "generally good" style, we should design our agents for this. Figure 4 plots intermediaries' style against users' reported effort. There is no apparent correlation, and certainly no "sweet spot" across the range: that is, there is no single "best style". There is no apparent best style either for UES or opinion-of-other (not shown).

4.3 Difference in style

Since there is no single good style, it is possible that effort is at least partly explained by *differences* in style between "user" and "intermediary": that is, that more effort is reported when the two participants have different conversational styles [22, 24]. We test this idea by constructing, for each pair of participants and each task, a difference-in-style variable which is simply the absolute difference between the user's and the intermediary's style on the task. If observations from casual chit-chat [24] and crowdsourcing experiments [22] are borne out here, we would expect reported effort to correlate with difference. This is in fact what we observe.

To investigate the relationship, we built a mixed-effects (linear) model to predict user's effort, as their per-task TLX score, as a response to difference in style. As TLX varies with participant and task, the model also included random effects (both intercept and slope) for both. Model fitting used the `lme4` package from Bates et al. [1], in R 3.3.1 [19].

The fitted model is remarkably accurate (Figure 5). The mean error is only 0.58 points on the seven-point scale, and no pattern was apparent in the residuals. The overall (fixed) effect of difference is 0.09: that is, effort is 0.09 points higher for every one-point difference in involvement. The variance due to participant is greater than that due to task, as is common in interactive retrieval. Per-participant effects of difference range from −0.13 to 0.27; per-task effects range from −0.03 to 0.16.[3]

(As an aside, we note that the model would not be useful as a metric or predictor "in the wild": it includes random effects for task, which is not normally observable. As we will see, the model

[3] A similar model can be built for intermediaries' effort, with similar effects (fixed effect 0.09; per-participant random effects −0.07–0.30; per-task random effects 0.06–0.15; mean error higher at 0.82). In this study we are interested in the users' experience, not the intermediaries', so we leave further investigations to future work.

is useful for exploring and understanding the relationship between style and effort, and this can lead to design guidelines.)

Direction of difference. Differences in style do make a difference to the "user's" experience. Since the roles of user and intermediary are asymmetric, it is reasonable to ask whether the direction of the difference is significant: that is, does it matter whether it is the user or the intermediary who is more involved? For example, if the intermediary exhibits higher consideration than the user, then she might be seen as courteous; while an intermediary who exhibits higher involvement might be seen as pushy.

We built a version of the model above with two fixed effects, one for the degree to which the intermediary was higher in consideration and one for the degree to which they were higher in involvement. Both were clamped to zero: that is, if the intermediary exhibited higher involvement, then the "higher consideration" variable would be set to zero rather than be negative. This let us model the effect of each direction separately.

The learned model, in this case, had near-identical effects for both "higher involvement" and "higher consideration". This indicates that the *amount* of difference, not the *direction*, is important for perceptions of effort.

Task effects. Although the effect due to task is less than that due to participant, it is still significant. Intercepts range from 2.22 (heroism and Olympics) to 3.35 (transport), as we might expect; but the effect of style differences ranges on both sides of zero. For the heroism and Olympics tasks, difference in style does indeed correlate with effort (effect 0.16). For the migraine task it makes less difference, although differences still increase effort (0.05), and for the transport task we see overall *lower* effort when there is more difference (−0.03).

As possible explanations for this, we note that the migraine and transport tasks are the "high complexity" tasks, which required participants to compare, aggregate, and synthesise information. They are also the tasks which tended to take the longest, even after removing all instances past ten minutes (Figure 6). We consider each of these aspects below.

Complexity and difficulty effects. To examine the effect of task complexity and difficulty, we built a second mixed-effects linear model with fixed effects for style difference; and included a style difference-complexity interaction; and style-difference-difficulty interaction. (That is, the effect of style difference was allowed to vary for each level of complexity and of difficulty). Note that whereas the previous model allowed each task to vary independently, as a random effect, this model postulates a pair of underlying effects and links tasks accordingly. The new model had no random effect for task but retained a random intercept and slope per participant. Final accuracy was good, with mean error 0.58 points, suggesting we have not given up any explanatory power by representing a task by its complexity and difficulty.

Figure 7 plots the fixed effects in the resulting model. We can see again that style differences are most clearly bad for the heroism task, moderate for the Olympics and migraine task, and seem to help the transport task; however we can also see that both increasing task difficulty (left to right) and increasing task complexity (top to bottom) reduces the effect of style mismatch.

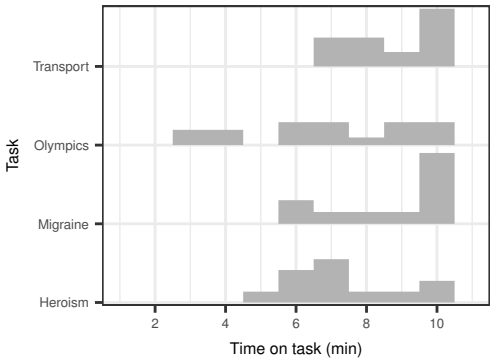

Figure 6: Task times, measured as time of the last recorded conversation. Participants were asked to stop at 10:00 and any longer tasks were excluded (see text). The transport and migraine tasks took considerably longer, typically, than the heroism or Olympics tasks.

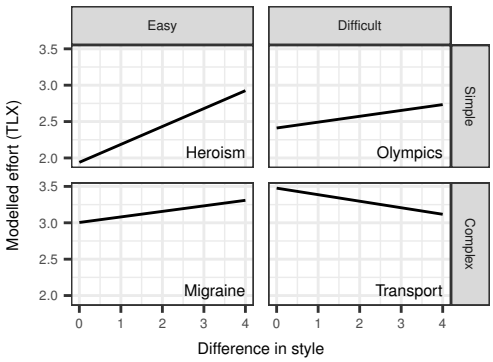

Figure 7: Modelling users' effort as a response to difference in style, complexity, and difficulty (only fixed effects shown). Both complexity and difficulty change the effect of style differences.

It may be that as tasks get more complex and difficult, they require more mental processing and become more "intellective", leading people to focus more on the task content than on factors such as conversational style. Similar effects of task complexity have been noted in a group setting [13]. Another possibility is that as tasks become more difficult, it becomes harder for people to align their styles [10]. As a result, participants may have focussed more on solving the task and less on their and their partner's style, and mismatch could perhaps have become less of an issue.

Task length effects. Since the complex and difficult tasks tended to take longer, we can use similar modelling to examine the effect of time on task. In this case we replace the random effect of task with a fixed effect for time on task, where we use the total time in conversation as a proxy for time on task. Again the result matches closely, with mean error 0.57 on the seven-point scale.

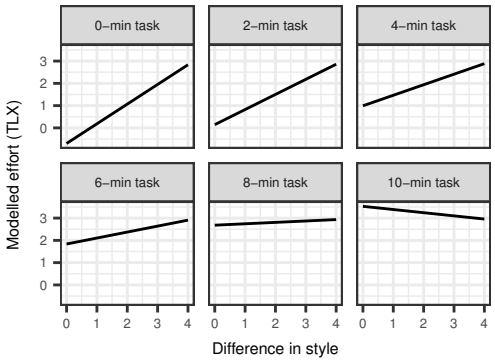

Figure 8: Modelling users' effort as a response to difference in style and time spent on task (only fixed effects shown). As tasks get longer, style differences are less important; past about nine minutes, style differences actually reduce effort.

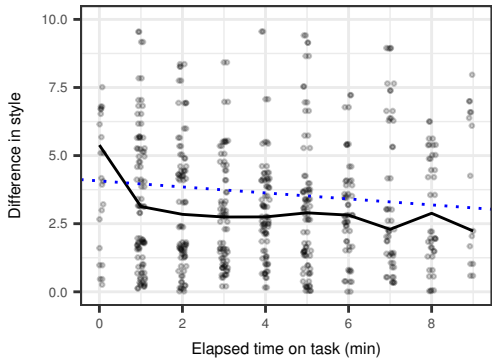

Figure 9: Alignment of conversational style. Solid line shows median difference between "user" and "intermediary", over the course of a task. Dotted line is fixed effect from a mixed-effects model (see text). Alignment is just over 0.11 units per minute on task.

Figure 8 shows the resulting model. For short tasks, up to about six minutes, our model has difference correlating with effort; for medium tasks, it correlates less (and makes no difference for tasks taking about nine minutes); for long tasks, past about nine minutes, it starts to anticorrelate.

If difference in style does become less important as tasks get longer, this anticorrelation may of course be an artefact of our linear model: a linear model cannot represent, for example, a "flattening out" of effect as tasks get longer.

The fact that differences in style do have an effect on experience is congruent with Tannen's model, although to our knowledge it has not been observed before in a search setting. The fact that differences have less impact when tasks are longer, however, warrants further attention and we turn to this next.

5 ALIGNMENT

The effect of task length might be explained several ways. The simplest case is just that "style", as defined here, depends on elapsed time. For example, there is more chance for a measurement to vary given more samples, so the pv (pitch variance) and lv (loudness variance) could be higher for longer tasks due to glitches in the recording or outbursts from the speakers. Similarly, a longer task might involve discussion of more aspects and ppron might vary accordingly. This could lead to more noise in the difference between styles, and therefore less apparent correlation with effort or engagement. However, this simple hypothesis does not hold: neither pv nor lv, the two variance-based variables, correlate with total time on task ($r = -.01$ and .03, $t(96) = -.12$ and .32, n.s.). Further, of the eleven variables, only one correlates with total time (repu, proportion of utterances with repeated terms: $r = -0.27$, $t(96) = -2.70$, $p < 0.05$). Finally, we see no apparent increase in variability of involvement across task lengths (regressing variance in style on task length to the nearest minute $r = -0.20$, $t(6) = -0.50$, n.s.).

A second possible explanation is that users "got their ear in"—that is, that over time they became more accustomed to their partners' conversational style and it became less of a hindrance. This would not be observable in the MISC data, except possibly by looking for markers of puzzlement such as "huh?" or frowning.

Alignment. A further possibility is *alignment*. Alignment is the largely unconscious process by which speakers converge on common ways of speaking: for example, by preferring different syntax, lexical choices, articulations, prosodic styles, and even accents [5, 11, 18]. It has been observed in both human-human interaction and human-computer interaction [5], so it is reasonable to expect it in the present case. If MISC participants did align their conversational style over the course of a task, the effect of style mismatch would be reduced.

Evidence for alignment. If there were alignment in MISC, in features such as speech rate or pronoun use, we would see the difference in involvement between "user" and "intermediary" drop during the course of a task, or across the whole series of tasks.

We used the same variables, scaling, and loading as described above to calculate involvement on a minute-by-minute basis for each participant and task. Partial minutes at the end of each task were dropped, and minutes with no utterance were recorded with undefined style.[4] The variation in style is much higher in this version—the underlying variables are not being averaged over the entire task, so cover a greater range—but the overall distribution remains approximately normal.

Figure 9 plots the difference between users' and intermediaries' styles, across all pairs and tasks, minute by minute. Although there is a good deal of variation, the median difference drops a good deal in the first minute and does tend down overall ($r = -0.09$, $t(654) = -2.33$, $p < 0.05$). Again building a linear mixed-effects model with random intercept and slope for each pair, we see a fixed effect of a 0.11-unit drop in difference per minute on the task—i.e., if a typical pair started a task with 3 units of difference in style, this would be reduced to 1.9 units by the 10-minute mark.

[4]For example, words per utterance (wpu) is undefined if there is no utterance, and between-own pause length (boplen) is undefined if there is no pause.

We would see this if styles all converged on the same point, as time went on—for example, if fatigue or some quirk of recording caused participants to all adopt the same style over time. There is no evidence of this, however: variance in involvement scores does not decrease across time and in fact is lowest between minutes 3 and 5. Instead, pairs appear to be converging on different styles. This is evidence of alignment over the course of a task: stylistic differences are being smoothed over, consciously or not, by small changes in such factors as rate of speech and use of pronouns.

There is no indication of alignment over longer periods, in the MISC data. That is, there is no overall alignment of style over the whole hour or so of the exercise ($r = 0.01$, $t(654) = 0.19$, n.s.). The MISC protocol had participants work separately after each task, answering questions on their experience, which took a few minutes. It seems likely that this enforced break in conversation, plus the abrupt change of topic when the next task began, meant participants somehow "reset" their adaptations.

Alignment and time on task. Observing alignment partly explains the interaction between time on task, difference in style, and users' reported effort. Spending longer on a task gives more time to align styles, that is to change styles; so a single style computed over the whole task is not as representative. This in turn means this single style is not a good predictor of reported effort, and we would expect to see less effect.

Alignment by role. Since there is evidence that style varies slightly with role, it is possible that alignment also varies—for example, that intermediaries make more effort (even unconsciously) to meet users than vice versa.

Minute to minute, users changed their involvement more than did intermediaries. Median jumps were 2.8 units for users, 1.8 for intermediaries, a significant difference (two-sided $t(441.06) = 5.27$, $p < 0.05$). For both roles, changes in style tended to be in the direction of the partner: that is, if in minute n the participant showed less involvement than their partner, then in minute $n + 1$ they would increase involvement and vice versa (one-sided $t(515) = 2.59$, $p < 0.05$). However, users made larger shifts, closing 46% of the gap minute to minute compared to 18% for intermediaries. This difference was significant (paired $t(48) = 2.20$, $p < 0.05$). This is probably an artefact of the MISC protocol: "users", without any resources to tackle the task themselves, had less to do so may have made greater use of verbal cues to encourage a solution. This also put users in a submissive role, which would result in greater attempts at alignment [4, 9].

6 ENGAGEMENT AND OPINION

The discussion above has focussed on reports of effort as measured by the TLX. We also considered reports of engagement and opinions of the partner.

6.1 Engagement

MISC includes items on engagement for each task and participant, drawn from the User Engagement Scale [16]. The effect of style on engagement is consistent with that on effort, but smaller overall: differences in style are more important to the sense of effort than the sense of engagement.

In particular, there is no single style which maximises engagement; rather, engagement varies with difference in style between the partners. Modelling engagement as a response to style difference, with random effects as in Section 4.3, again gives a good fit (mean error 0.50 units out of seven) but with a smaller effect (0.02 units *less* engagement per unit of style difference, compared to 0.09 units more effort). We observed similar effects of complexity, difficulty, and time, although these effects were smaller all around.

6.2 Opinion

Our third dependent variable is the user's opinion of the intermediary, based on whether they felt understood; whether they felt helped; and whether the intermediary communicated clearly. Again, there is no best style; and other effects are consistent but they are smaller still.

A model of opinion as a response to style difference was remarkably accurate, with mean error only 0.29 units (c.f. 0.58 units for effort). The effects of complexity and difficulty remain, but are smaller still, and the effect of time is negligible.

These items asked about the partner in particular—not the task, or the participant themselves—so smaller effects make sense. They can also be explained by bias in the data, as most responses were 6 or 7 on the seven-point scale.

7 CONCLUDING REMARKS

It is possible to measure the conversational styles adopted by "users" and "intermediaries" in this information-seeking context, and to distinguish differences in a single axis from involvement to consideration. These styles—in particular, differences between the partners' styles—do make a difference to users' reports of effort and engagement.

7.1 Style and style difference

This study focuses on eleven stylistic variables, based on Tannen's description of styles in casual chit-chat [23] but chosen partly for processing convenience. These variables do point to a single factor, which we identify with "involvement". To the best of our knowledge this is the first work to measure involvement in this way, and the first to provide a process for doing so automatically and at scale.

We believe it is also the first study to consider conversational style in information-seeking contexts, and we see a similar pattern to that reported elsewhere. In particular, differences in style contribute to a sense of effort, as reported by the MISC "user" on the NASA task load index. Allowing for per-participant and per-task differences, tasks where partners exhibited similar styles were those which took less effort.

There is also an effect of task complexity and difficulty (differences are less important in more complex or difficult tasks), and of time (overall differences are less important when tasks take longer).

All these effects are also at play for users' engagement, although the effects are smaller, and for users' opinion of their partner, where the effects are smaller still.

7.2 Alignment

The interaction between effort, style, and time on task is at least partly explained by alignment, whereby the user and intermediary

work (perhaps unconsciously) to match each others' expressions of involvement or consideration. We see some evidence for this in the MISC data: participants' styles tend to change to close the gap. This makes per-task measures of style less representative for longer tasks. We also see evidence that users, rather than intermediaries, change their expression more. This may be because users were relatively submissive and intermediaries dominant; if so, we would expect the opposite effect for people interacting with software agents.

7.3 System design

These results suggest a few design principles for speech-based, conversational, agents. First, there is no single best style—there is no amount of involvement or consideration which gets uniformly good feedback. Second, it should be possible to monitor a user's conversational style: the variables used here are by no means the only choices, but are all computable at scale and in real-time. Having identified a style, these results suggest adapting to that style will help the user: the same tasks are reported as needing lower effort, and being slightly more engaging, when styles match more closely. This effect is more pronounced when tasks are shorter.

Again, although there are many other possibilities, the variables used here could all be under the control of software systems: speech rate and pauses are easy to adapt, loudness and pitch variation likewise, and dialogue could be varied to include more or fewer personal references.

7.4 Limitations and future work

We note some limitations of the analyses above. Most importantly, the MISC data set was collected from conversations between strangers, working on assigned tasks; it does not reflect how people currently talk to software agents or how people might work on their own tasks (especially shorter tasks). At present, of course, software agents are not capable of carrying on such long conversations nor (outside some experimental systems) are they capable of varying their style in any substantial way; so this limitation is forced on us. The data set is also possibly too small to note subtle effects, and our linear models cannot represent non-linear effects if they exist. Manual inspection of the residuals has not revealed any obvious patterns, however.

The present study is purely descriptive, not experimental. An obvious follow-up would be to measure participants' styles, and adapt an agent to match (or not) before recording effort and engagement; perhaps using a Wizard-of-Oz strategy for dialogue management. This would be an extension of work by Shamekhi et al. [22], for example, with more experimental control and with a more mechanistic notion of "style". We hope to run such experiments in future.

AVAILABILITY

The code used to derive the stylistic variables, the resulting data, and the code used for analysis is available on request.

ACKNOWLEDGMENTS

The variables listed in Section 3.2 build on suggestions from Gregory A. Bennett. Trudy O'Connor contributed to the discussion in Section 5. We thank the MISC participants for their time.

REFERENCES

[1] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48.
[2] M M Berg. 2014. Modelling of natural dialogues in the context of speech-based information and control systems. PhD thesis, University of Kiel. (2014).
[3] Timothy Bickmore and Justine 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.
[4] Frances R Bilous and Robert M Krauss. 1988. Dominance and accommodation in the conversational behaviours of same- and mixed-gender dyads. *Language & Communication* 8, 3/4 (1988), 183–194.
[5] Holly P. Branigan, Martin J. Pickering, Jamie Pearson, and Janet F. McLean. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics* 42 (2010), 2355–2368.
[6] Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language use.* Cambridge University Press, Cambridge.
[7] Robert F DeVellis. 2003. *Scale development: Theory and applications* (2nd ed.). Sage, Thousand Oaks, California.
[8] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In *Proc. ACM Multimedia*. ACM Press, 835–838.
[9] Cindy Gallois, Tania Ogay, and Howard Giles. 2006. Communicaition accommodation theory: A look back and a look ahead. In *Theorizing about communication and culture*, W B Gudykunst (Ed.). Sage, Thousand Oaks, 121–148.
[10] Simon Garrod and Martin J Pickering. 2009. Joint action, interactive alignment, and dialog. *Topics in Cognitive Science* 1, 2 (2009), 292–304.
[11] Vivien Kühne, Astrid Marieke Rosenthal von der Pütten, and Nicole C. Krämer. 2013. Using linguistic alignment to enhance learning experience with pedagogical agents: The special case of dialect. In *Proc. Int. W'shop on Intelligent Virtual Agents*. Springer, 149–158.
[12] Robin Tolmach Lakoff. 1979. Stylistic strategies within a grammar of style. *Annals of the New York Academy of Sciences* 327, 1 (1979), 53–78.
[13] Joseph E McGrath. 1984. *Groups: Interaction and performance.* Prentice-Hall, Englewood Cliffs, NJ.
[14] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. 56, 1 (2000), 81–103.
[15] National Aeronautics and Space Administration Human Systems Integration Division. 2016. TLX @ NASA Ames. (2016). Retrieved January 2017 from https://humansystems.arc.nasa.gov/groups/TLX/
[16] Heather L O'Brien and Elaine G Toms. 2010. The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology* 61, 1 (2010), 50–69.
[17] J W Pennbaker, R L Boyd, K Jordan, and K Blackburn. 2015. *The development and psychometric properties of LIWC2015.* Technical Report. University of Texas at Austin.
[18] Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brian Sciences* 27, 2 (2004), 169–225.
[19] R Core Team. 2016. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
[20] Matthias Rehm, Elisabeth André, Yukiko Nakano, Toyoaki Nishida, Nikolaus Bee, Birgit Endrass, Hung-Hsuan Huan, Michael Wissner, I Mayer, and H Mastik. 2007. The CUBE-G approach-Coaching culture-specific nonverbal behavior by virtual agents. *Proceedings of ISAGA* (2007).
[21] William Revelle. 2017. *psych: Procedures for psychological, psychometric, and personality research.* Northwestern University, Evanston, Illinois. R package version 1.7.3.
[22] Ameneh Shamekhi, Mary Czerwinski, Gloria Mark, Margeigh Novotny, and Gregory A Bennett. 2016. An exploratory study toward the preferred conversational style for compatible virtual agents. In *Proc. Int. Conf. on Intelligent Virtual Agents*. Springer, 40–50.
[23] Deborah Tannen. 1987. Conversational style. In *Psycholinguistic models of production*, Hans W Dechert and Manfred Raupach (Eds.). Ablex, Norwood, NJ.
[24] Deborah Tannen. 2005. *Conversational style: Analyzing talk among friends* (new ed.). Oxford University Press, New York.
[25] Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2017. MISC: A data set of information-seeking conversations. In *Proc. Int. W'shop on Conversational Approaches to Information Retrieval.*
[26] Italo Trizano-Hermosilla and Jesús M Alvarado. 2016. Best alternatives to Cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. *Frontiers in Psychology* 7, Article 769 (2016).