A Workflow Analysis of Context-driven Conversational Recommendation

Shengnan Lyu
University of Toronto
Toronto, ON, Canada
zoey.lyu@mail.utoronto.ca

Arpit Rana
University of Toronto
Toronto, ON, Canada
arana@mie.utoronto.ca

Scott Sanner
University of Toronto
Toronto, ON, Canada
ssanner@mie.utoronto.ca

Mohamed Reda Bouadjenek
Deakin University
Geelong, VIC 3216, Australia
reda.bouadjenek@deakin.edu.au

ABSTRACT

A number of recent works have made seminal contributions to the understanding of user intent and recommender interaction in conversational recommendation. However, to date, these studies have not focused explicitly on context-driven interaction that underlies the typical use of more pervasive Question Answering (QA) focused conversational assistants like Amazon Alexa, Apple Siri, and Google Assistant. In this paper, we aim to understand a general workflow of natural context-driven conversational recommendation that arises from a pairwise study of a human user interacting with a human simulating the role of a recommender. In our analysis of this intrinsically organic human-to-human conversation, we observe a clear structure of interaction workflow consisting of a preference elicitation and refinement stage, followed by inquiry and critiquing stages after the first recommendation. To better understand the nature of these stages and the conversational flow within them, we augment existing taxonomies of intent and action to label all interactions at each stage and analyze the workflow. From this analysis, we identify distinct conversational characteristics of each stage, e.g., (i) the preference elicitation stage consists of significant iteration to clarify, refine, and obtain a mutual understanding of preferences, (ii) the inquiry and critiquing stage consists of extensive informational queries to understand features of the recommended item and to (implicitly) specify critiques, and (iii) explanation appears to drive a substantial portion of the post-recommendation interaction, suggesting that beyond the purpose of justification, explanation serves a critical role to direct the evolving conversation itself. Altogether, we contribute a novel qualitative and quantitative analysis of workflow in conversational recommendation that further refines our existing understanding of this important frontier of conversational systems and suggests a number of critical avenues for further research to better automate natural recommendation conversations.

WWW '21, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442381.3450123

ACM Reference Format:
Shengnan Lyu, Arpit Rana, Scott Sanner, and Mohamed Reda Bouadjenek. 2021. A Workflow Analysis of Context-driven Conversational Recommendation. In *Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3442381.3450123

1 INTRODUCTION

As natural language conversational assistants are on the rise, there is a growing interest in conversational recommender systems [5, 12, 24] that leverage natural language dialog. To date, some user studies [3, 14, 18] and some implementations [9, 15] have started to investigate aspects of this research area. However, none of these studies or methods focus explicitly on context-driven recommendation settings where the user interacts with a fixed intent and context; yet this context-driven setting is critical to bring personalized recommendation to the next generation of conversational assistants such as Amazon Alexa, Apple Siri, and Google Assistant.

Hence, our objective in this work is to understand the workflow of natural context-driven recommendation conversations through the analysis of dialog transcripts from human-to-human recommendation interactions. Specifically, we ran a pairwise study of a human user interacting with a human simulating the role of a recommender. Following this, we built on existing taxonomies of user intents and recommendation actions [3] to label all interactions at each stage and analyze the workflow of these labeled conversations. At a first pass of global analysis of these intrinsically organic human-to-human conversations, the workflow results are interesting but not surprising — the bulk of the conversational flow focuses on question and response style interaction as the two humans come to a mutual understanding of the recommendation context and the user's constraints and personal preferences.

However, leveraging our informal observations from the user study that suggested the nature of most conversations shift substantially as the first recommendation is made and then slightly more after subsequent recommendations, we are motivated to undertake a more nuanced stage-wise analysis of the conversation with stages defined according to 0, 1, and 2 (or more) recommendations. Once we make this distinction, we start to observe very distinct patterns of behavior that are also supported by a variety of additional quantitative and qualitative analysis. Among numerous insights discussed throughout the paper, we observed the following:

- Over time, the conversation shifts from preference elicitation and refinement in the first stage to inquiry and critiquing in subsequent stages after the first recommendation.
- Roughly a quarter of the total interaction can be attributed to conversation management in the form of mutual acknowledgement and progress updates.
- We observe a mixed-initiative approach in the conversation choreography, whereby the user leads the conversation at some points and the recommender leads at other points.
- Informational queries intended to clarify features of recommended items often implicitly serve as expressions of soft preference for subsequent recommendations.
- Explanation seems to drive a substantial fraction of the interaction and follow-up inquiries and critiques by the user. While explanation has been noted as being important for transparent recommendation [27, 28], this critical role of explanation in actually *directing* the conversation seems to be under-emphasized in existing work on dialog-based conversational recommendation.

In summary, this work presents a novel qualitative and quantitative analysis of workflow in natural dialog-based context-driven conversational recommendation. In terms of scope, our focus in this work is on a descriptive analysis of the high-level workflow of natural human-to-human conversations; i.e., we do not aim to prescribe an actual system to embody this complex, natural style of interaction, nor to solve the low-level natural language processing challenges that such automation would entail. Rather, we simply aim that the insights obtained through this study should help us better understand the structure of organic context-driven recommendation. To this end, we believe this work provides important design guidance for the next generation of fully automated and naturally interactive conversational *recommendation* assistants.

2 RELATED WORK

Conversational recommender systems (CRSs) help users navigate through a complex item space to find an item of interest. During the dialog, the CRS must select items for recommendation and elicit feedback from the user, which it must take into account before making the next set of recommendations [12]. Many CRSs use **GUI-based interactive recommendation** [1, 22], which allows users to provide their feedback in one of four ways: (i) by asking questions for a value of a specific item feature [21]; (ii) by collecting user ratings on the proposed recommendations [25], (iii) by inviting users to select one of many recommendations [20], or, (iv) by allowing users to provide critiques on item features [31]. These feedback forms differ in their level of ambiguity and the efforts they demand from the user [23]. However, these systems limit the way users can provide their feedback since their interactions are restricted to GUI elements (e.g., menu, form, button).

More recently, the word "conversational" has implied not just a sequential interaction, as before, but a dialog in natural language, e.g., [7]. **Dialog-based CRSs** allow users to provide their feedback in a more natural and free-form manner [14]. The system identifies users' intents from their utterances and incrementally refines users' preference models to adjust its dialog policy [26]. Most work on dialog-based CRSs is focused on question-answering (QA), where they aim to equip the system with a strategy for selecting an optimal set of features to request from the user and a logical order in which to ask about them [5, 34]. Item features to query can be chosen for example by deep reinforcement learning [24], deep recurrent networks [4], or multi-memory networks [34].

Understanding and disambiguating user intent is a critical task for dialog-based CRSs [11]. Despite the substantial research in this area, identifying user intents from their utterances as to what kinds of preferences a user wants to convey (and how they express them) has gained some recent attention. In 2017, Kang et al. [14] looked into how users initiated natural language queries with a recommender system. They qualitatively derived a categorization of user queries, but did not analyze full dialogs to study user behavior or workflow. In 2018, Chen et al. [19] taxonomized user intents into 12 classes by analyzing the MSDialogue dataset for dialog-based QA systems with no focus on recommendations. Recently, Cai & Chen [2] pointed out that user behavour towards dialog-based QA systems and towards recommender systems are different, and proposed taxonomies for user intents and recommender actions using the REDial dataset [15] centered around movie recommendations. They further utilized the labeled dataset for user intent prediction and combined both taxonomies for user satisfaction prediction [3]. Pecune et al. [18] developed a computational model of explanations for movie recommenders by summarizing the types of explanations humans used in the Switchboard dataset [10] and incorporated it into actions of their movie recommender system with an aim of improving the quality of the recommendation and interaction perception. These methods, although close in principle, have significant differences from our work. None of these studies cover *context-driven* conversational recommendation [17], where the user has explicit contextual constraints driving their recommender interaction. The incorporation of context in our study naturally leads to a strong formulation of preferences and constraints in the dialog, which is a critical aspect of conversational workflow relevant to the typical goal-directed use of a conversational assistant like Amazon Alexa, Apple Siri, or Google Assistant that we wish to study in this work.

Furthermore, while we build on the foundational coding protocol and analysis of Cai and Chen [3], prior work has not sought to use these codes to analyze the generic *sequential workflow* of these recommendation-based conversations as we focus on in this work.

3 DATA COLLECTION

To analyze user intents and recommender actions for context-driven recommendation, we conducted a user study on restaurant recommendation. We have selected this domain since most study participants would be expected to have extensive experience selecting restaurants and because the context we provide (e.g., family outing with young children) should have a strong influence on intrinsic user preferences as well as context-specific preferences (e.g., has a kids menu, special diet, etc.) to drive the conversational interaction. Furthermore, there is often a higher level of commitment required for negotiating a restaurant recommendation than, for example, a movie recommendation; i.e., in an online movie recommendation setting, a user could simply stop a movie after a few minutes and find another if she dislikes the recommendation; in the restaurant domain, users are less tolerant of making a mistake in their selection. Hence, we believe users are more likely to make a concerted effort in a restaurant conversation to ensure the recommendation is appropriate to the given scenario.

In the remainder of this section, we will discuss limitations of existing datasets, our experimental protocol to collect data, and the coding taxonomy we extend and expand on in our analysis.

3.1 Existing Datasets

For *restaurant recommendation*, we are aware of three readily available datasets: MultiWOZ [33], ParlAI [13], and the CRM [24]. While portions of these existing datasets are synthetic or partially synthetic (machine-generated) conversations rather than the fully natural language dialogs we aim to study, there are additional shortcomings of existing data. In general, we note that none of the existing studies and datasets cover fully human-to-human context-driven conversational recommendation between a single user and a single assistant that we believe reflects the most common use case of a CRS. Additionally, these existing restaurant recommendation datasets were collected via text-to-text interaction; we conjecture that spoken natural language interaction may lead to richer dialogs than text interaction and thus better reflect interactions we would expect in deployed CRS systems. Hence, we decided to collect our own data for this research. We believe our contextual goal-oriented setting with a single human information seeker and a single human assistant reflects the use of CRSs in many deployed conversational assistant settings, which is not represented in the existing datasets of which we are presently aware.

3.2 Subject Recruitment and Task Assignment

We recruited graduates and alumni from University of Toronto to participate in our study through departmental email lists. People who agreed to participate in the study were shown an experiment consent form; they were given the chance to opt-out or stop at any stage of the experiment. This experiment was approved by our institution's research ethics board (REB).[1] In total, 24 people completed the study and have their results reported in this work.

In our experiment, we setup a task of restaurant recommendations for our local city and designed three scenarios to represent real-life situations: (i) *A business dinner with a colleague and two clients, one of whom is a vegetarian*; (ii) *A Valentine's Day dinner with your date*; and, (iii) *A family brunch on the weekend with your young child, parents and spouse*. These scenarios aim to support participants to drive the conversation and were developed under the guidance of the five contextual dimensions proposed in [32]: individual, location, time, activity, and relational.

We ran the experiment in two phases. In the first phase, each participant is asked to provide details related to their dine-out frequency (e.g., once a week, once a month, etc.) and their experience level with each of the three scenarios (e.g., not at all, somewhat, and highly experienced). Based on their experience level for each scenario, we divide our participants into two groups: (a) *expert group*; and (b) *novice group*. The experts are assigned the role of *recommenders* while the novices become the *users*. Each participant is assigned only one role throughout the experiment. Groups are created in such a way that all three scenarios would receive an equal number of *recommender–user* pairs and thus would result in an equal number of dialogs.

Once the groups are set, we create *recommender–user* pairs for each scenario by randomly selecting one *recommender* from the expert group and one *user* from the novice group. Each participant is selected only once in the whole experiment. We use randomization while creating pairs to avoid any selection bias. The reason for matching a user (a novice member) to a recommender (an expert member) is to better replicate a human-to-machine conversation where a recommender system is expected to have substantial knowledge about the conversation topic.

We then begin the second phase of our experiment where these *recommender–user* pairs converse with the goal of finding a suitable restaurant given a scenario. Each participant begins by going through the instructions as per her role. The scenario was disclosed to the user so that she can distill information from the context, drive the conversation according to her preferences and later decide which recommendation best suits her requirements. In contrast, the recommender is not aware of the user's scenario; she is supposed to recommend based on the preferences she identified through conversing with the user, which simulates a recommendation *cold-start* problem. Additionally, the recommender is advised to give recommendations primarily based on her prior experience. If she needs external assistance, she can find relevant information through online platforms e.g., Yelp. Both participants can ask any questions to the moderator before they start the conversation. Other than that, the moderator plays no role in the conversation. Once the dialog ends, the moderator informs both participants that the experiment is completed and confirms that they both agree to provide their experimental data for research analysis.

Using the above protocol, we obtained 12 dialogs (4 for each of the 3 scenarios) from 24 participants in our formal experiment. The dialogs were recorded during the experiment and then transcribed manually for the later identification and taxonomization of user intents and recommender actions. While the dataset may initially seem quite small, the conversational content is quite substantial as evidenced by the following statistics. Each conversation transcript has an average of 1,150 words. In total, we had 360 conversation turns, with each conversation averaging 30 turns. Since our workflow analysis is based on per-turn interactions, this leads to a dataset of 360 labeled interactions consisting of an average of approximately 38 words per turn (both user and recommender).

Due to the personal information revealed in conversations (about individuals, their families, health and religious concerns as they relate to food preferences, etc), our approved ethics protocol requires that full conversations must remain confidential. However, we do publish codes for the turn labels for research purposes and reproducibility of results.[2]

3.3 Coding Taxonomy Development

To understand and analyze natural human interactions during recommendation dialogs, we leveraged and extended two taxonomies of user intent and recommender actions from Cai and Chen [3].

[1] University of Toronto REB-approved Ethics Protocol #00039634 titled "Evaluation of Conversational Recommender Systems".

[2] https://github.com/D3Mlab/WWW21Paper

Table 1: Taxonomy for user intents, building on [3]. † indicates a new category that we have identified in our dialogs.

Category (Code)	Description	Example	%
Ask for Recommendation			**3.4%**
Initial Query (IQ) [3]	User asks for a recommendation in the first query	"Hi I am looking for a place to have a family brunch..."	2.4%
Continue (CON) [3]	User asks for another recommendation in a subsequent query	"Maybe you can give me one more choice so I can pick one..."	1.0%
Provide Preference			**19.5%**
Provide Context (PCT) †	User provides background information for the restaurant search	"I am looking for a restaurant for my Valentine's day dinner."	5.0%
Provide Preference (PP) [3]	User provides specific preference for the desired item	"I would prefer a place that has a very good scenic view."	11.5%
Refine Preference (RP) †	User improves over-constrained/under-constrained preferences	"It does not have to be chicken fingers."	3.0%
Answer (ANS) [3]	User answers the question issued by the recommender	"Yes that's correct."	**12.9%**
Acknowledgement (ACK) †	User shows understanding towards a previous recommender utterance	" I see."	**28.0%**
Recommendation Rating			**4.4%**
Been to (BT) (modified)	User has been to the restaurant before	"Oh I have been there before."	0.2%
Accept (ACT) [3]	User accepts the recommended item, either explicitly or implicitly	"Ok our final choice will be Eggspectation."	2.8%
Reject (RJT) [3]	User rejects the recommended item, either explicitly or implicitly	"Maybe there is a private room in the other three restaurants?"	1.0%
Neutral Response (NR) [3]	User does not indicate a decision with the current recommendations	"I will take a look in the menu and compare and maybe ask my partner."	0.4%
Inquire (INQ) [3]	User requires additional information regarding the recommendation	"So what about the interior design, the decorations and environment?"	**9.5%**
Critiquing			**3.6%**
Critique - Feature (CF) [3]	User critiques on a specific feature of the recommended item	"I am pretty sure it will be expensive so what is the price range?"	1.0%
Critique - Add (CA) [3]	User adds further constraints on top of the current recommendation	"I want sushi."	2.0%
Critique - Compare (CC) [3]	User requests comparison between recommended item with another item	"How about the price compared with Miku??"	0.6%
Others (OTH) [3]	Greetings, gratitude expression, chit-chat utterance	"Thank you so much for your recommendation."	**18.7%**

Table 2: Taxonomy for recommender actions, building on [3]. † indicates a new category that we have identified in our dialogs.

Category (Code)	Description	Example	%
Request			**14.2%**
Request Information (RI) [3]	Recommender requests the user's preference	"What kind of food do you like?"	4.4%
Clarify Question (CQ) [3]	Recommender asks for clarification on a previous requirement	"So you would like to reserve a private room?"	3.6%
Ask Opinion (AO) †	Recommender requests the user's opinion to a choice question (e.g., yes/no)	"So it is just open space but separated from others, is that ok?"	4.8%
Ensure fulfillment (EF) †	Recommender confirms task fulfillment during the conversation	"Anything else I can do for you today?"	1.4%
Inform progress (IP) †	Recommender discloses the current item being processed	"So let me just check the closest nearby parking."	**7.1%**
Acknowledgement (ACK) †	Recommender shows understanding towards a previous user utterance	"...you mentioned that one of the attendees is vegetarian..."	**16.7%**
Answer (ANS) [3]	Recommender answers the question issued by the user	"So for the Michael's on Simcoe, the price varies a lot..."	**8.3%**
Recommend			**5.6%**
Recommend - Show (RS) [3]	Recommender provides recommendation by showing it directly	"So I found a restaurant called paramount."	4.6%
Recommend - Explore (RE) [3]	Recommender provides recommendation and asks if the user has prior knowledge	"The first one that comes to mind is Miku, have you heard of it before?"	1.0%
Explain			**38.4%**
Preference (EP) [3]	Recommender explains recommendations based on the user's said preference	"Because it has vegetarian options, it has a full bar and a good view..."	11.6%
Additional Information (EAI) †	Recommender explains recommendations with features not previously discussed	"There are a couple different varieties (of food) that your guests might enjoy."	15.4%
Personal Opinion			**11.4%**
Comparison (PCM) †	Recommender compares recommended item with another item	"I would say the price for this place is a bit higher than HY steakhouse but..."	0.9%
Persuasion (PER) †	Recommender provides positive comment towards the recommended item	"It is on the pricier side but it is worth the experience..."	5.6%
Prior Experience (PEX) †	Recommender refers to past experience with the recommended item	"I have been there before during the summerlicious."	2.5%
Context (PCN) †	Recommender provides opinion considering the given context or current reality	"Since it's summer I don't think the weather will be that much of an issue..."	2.4%
Others (OTH) [3]	Greetings, gratitude expression, chit-chat utterance	"Yeah a lot of people recommended me to go there."	**9.7%**

Specifically, we followed the grounded theory approach developed by Glaser and Strauss [8] to generate the specific codes and taxonomies required for our data. This involved adapting the user intents and recommender actions from Cai and Chen [3], while adding some new categories as observed in our data. The categories with corresponding description, examples, and percentage of their occurrences across all dialogs can be seen in Tables 1 & 2. We labeled categories as † for *new* to indicate that they were not covered in [3]. While we manually labeled these categories, it should be possible to automate this labeling process in the future for the purpose of designing automated CRSs.

3.3.1 Taxonomy for User Intents. We now turn to the design of our taxonomy for user intents to identify the types of user utterances. We came up with 8 top-level intents (see highlighted categories in Table 1), and 12 sub-intents described described in-depth as follows.

Ask for Recommendation: The user asks for a recommendation with an *Initial Query*, e.g., *"Hi, I am looking for a place to have a family brunch..."*. She sometimes needs more suggestions to compare with the current recommended item, so she asks for another recommendation in a subsequent query (*Continue*), e.g., *"Maybe you can give me one more choice so I can pick one..."*.

Provide Preference: We found that the user provides her preferences in nearly 20% of the turns. She mentions her context of visiting a restaurant as additional information for the recommender (*Provide Context*), e.g., *"I am looking for a restaurant for Valentine's day dinner..."*, specifies her requirements and constraints (*Provide Preference*), and sometimes refines her previously specified constraints (*Refine Preference*), e.g., *"It does not have to be chicken fingers."*.

Answer: The user answers questions issued by the recommender.

Acknowledgement: The user acknowledges the recommender that she understands what has been told to her.

Recommendation Rating: The user rates the recommendation either implicitly or explicitly. She may *Accept* (e.g., *"Ok our final choice will be Eggspectation."*) or *Reject* (*"Maybe there is a private room in the other three restaurants?"*) the recommended restaurant. She sometimes conveys that she has already *Been To* the suggested restaurant or sometimes she remains neutral and does not clearly take any decision (*Neutral Response*), e.g., *"I will take a look in the menu and compare and maybe ask my partner."*.

Inquire: The user asks for information regarding certain features of the current recommendation.

Critiquing: The user provides her feedback through critiquing an attribute of the recommended restaurant (*Critique – Feature*), adding more constraints on top of the current recommendations (*Critique – Add*), and sometimes asks the recommender to compare the current recommendation with another item on a specific attribute (*Critique – Compare*), e.g., *"How about the price compared with Miku?"*.

3.3.2 Taxonomy for Recommender Action. Now we turn to the recommender's actions, which we have classified into 7 top-level actions and 12 sub-actions (see Table 2).

Request: The recommender requests knowledge from the user to fulfill one of the four purposes. She requests the user to provide her preferences (*Request Information*), e.g., *"What kind of food do you like?"*. Sometimes she needs clarification from the user when she is not perfectly clear about a certain feature (*Clarify Question*), e.g., *"So you would like to reserve a private room?"*. She also asks for user's opinion on a choice question (*Ask Opinion*), e.g., *"So it is just open space but separated from others, is that ok?"*). She confirms the task fulfillment from the user (*Ensure fulfillment*), e.g., *"Anything else I can do for you today?"*).

Inform Progress: The recommender informs the user of the task in progress and a potential wait time.

Acknowledgement: The recommender acknowledges the user that she understands what has been told to her.

Answer: The recommender answers questions issued by the user.

Recommend: The recommender provides a recommendation either by directly presenting the item (*Recommend - Show*) or asking the user if she knows of the place (*Recommend - Explore*).

Explain: The recommender spends most of her time (≈ 38%) explaining the recommendations to the user. She employs various methods of explanation, such as, highlighting the user's mentioned preferences with *Explain - Preference* (e.g., *"Because it has vegetarian options, it has a full bar and a good view."*), providing *Explain - Additional Information* about the features that the user has not mentioned before (e.g., *"There are a couple different varieties (of food) that your guests might enjoy."*). The recommender also shares *Personal Opinion* as a way to explain the recommendations and support the decision of the user. The recommender utilizes *Persuasion* and supports the recommendation with reasons from *Prior Experience* and given *Context*. Moreover, the recommender also uses *Comparison* to give personal insights when more than one recommendation is present.

4 INITIAL CONVERSATION OBSERVATIONS

4.1 Context-related Categories

Since our experimental design intentionally infuses our dialog data with contextual information, we often observe users and recommenders exchanging this information and applying it in the recommendation process. For user intents, we observe a common occurrence of *Provide Context* (5.0%), which is passing on context information to the recommender (e.g., "I am looking for a restaurant for my Valentine's Day dinner."). For recommender actions, on the other hand, we observe contextual knowledge provided in the recommendation explanation through *Personal Opinion - Context* (2.4%). The recommender may explain based on what the user said about the context (e.g., "The atmosphere of the restaurant is casual but it's still presentable since you're looking for an investor for your company."), or provide additional contextual advice (e.g.,"Since it's summer I don't think the weather will be that much of an issue..."). The subcategory *Personal Opinion* was included since we observed that recommenders tend to combine their own personal perspective into the recommendation — something that could potentially be automated in a CRS by mining user review data.

4.2 Grounding and Conversation Management

Grounding is the process of showing understanding towards what someone has said and signaling that a common ground has been established [6]. This is very crucial since speakers tend to ensure that what has been said is understood by the counterpart before proceeding to the next phase of the conversation. The listener, on the other hand, generally relies on the speaker to maintain the conversation flow. In voice-based settings, where there is no face-to-face grounding [16], the means for grounding is solely verbal. We heavily observed this phenomenon in our data since it is transcribed from audio recordings. In our dataset, we observed a variety of commonly seen conversational strategies that humans use in showing understanding, such as backchannel responses (in our dataset, e.g., "Umm hmm.", "Ok."), repeating (e.g.,"University and Queen, okay"), completing the counterpart's sentence (e.g., (Recommender: "So maybe just go to the Eggs...") - User: "Eggspectation."), and paraphrasing. We labeled all of those utterances as **Acknowledgement** [29]. We observe that *Acknowledgement* takes up to 28.0% of the user intent, and 16.7% of the recommender actions in our data, which means that our participants spent a substantial portion of their interactions establishing common ground.

Also related to general conversation management, we observe the recommender action *Inform Progress* taking up 7.1% of the total observed recommender actions. In our data, recommenders either provided details informing the current item being processed (e.g., "So I am searching for the best Indian restaurants in Toronto."), or simply notified the user that there might be some wait time (e.g., "One second."). While fully automated recommenders will obviously not face the same delay issues as human recommenders, it is important to keep in mind that system latencies do affect users' perceptions regarding recommendation quality and relevancy [30].

4.3 Explanations and Personal Opinions

As previously observed (and as will subsequently be explored) in workflow analysis, recommender explanation plays a major role in our observed conversations (38.4% of observed recommender actions to be precise). Specifically, we observe two direct explanation strategies from the recommender, *Explain - Preference* (11.6%), *Explain - Additional Information* (15.4%). *Explain - Preference* is the act when the recommender mentions features of the recommended item that were requested by the user specifically, e.g., "Because it has vegetarian options, it has a full bar and a good view...". *Explain - Additional Information* means that the recommender presents information about the recommendation that was not discussed so far in the conversation (e.g., "There are a couple different varieties [of food] that your guests might enjoy."). We observe a higher occurrence percentage of *Explain - Additional Information* than *Explain - Preference*, which means recommenders tend to provide more information — though not required by the user — to paint a fuller picture of the recommended item. Considering that recommenders have access to the Internet to search for information related to the recommended item in our experiment, we conjecture the ready availability of this wealth of information might lead the human recommender to explain more than required.

An alternative explanation method employed by the recommender comes in the form of personal opinion. In previous work, researchers observed participants mentioning personal opinions when discussing favorite movies, and recognized this as a social explanation method [18]. While personal opinion in explanations may initially appear difficult to automate, as noted previously, this can potentially be supported in a CRS by mining user experience and opinions from review data.

Among our restaurant recommendation dialogs, utilizing personal opinions to explain the reason behind a recommendation or make suggestions is also very commonly observed. We classify four subcategories for *Personal Opinion*: *Personal Opinion - Comparison* (0.9%), *Personal Opinion - Persuasion* (5.6%), *Personal Opinion - Prior experience* (2.5%), *Personal Opinion - Context* (2.4%). *Personal Opinion - Persuasion* is when the recommender tries to "sell" the item by subjectively sharing positive perspectives (e.g., "It is on the pricier side but it is worth the experience..."). *Personal Opinion - Prior experience* is a strategy where the recommender refers to her prior experience with the recommended item to support the reason why it is suitable for the user (e.g., "I have been there during the summerlicious"). *Personal Opinion - Context*, as mentioned previously, supports reasoning either based on a user's background information or personal knowledge of the reality (e.g., "The atmosphere of the restaurant is casual but it's still presentable since you're looking for an investor for your company."). *Personal Opinion - Comparison* occurs when the recommender compares two candidate choices regarding a certain aspect, and potentially provides reasoning why they think one is more of a fit than the other for the user's situation (e.g., "So I would say the price for this place is a little bit higher than HY steakhouse but this restaurant has been there for a long time."). Overall, while not insignificant, we remark that opinion-based explanation strategies were employed by the recommender less frequently than the direct explanation strategies described earlier.

4.4 Analyzing User Preferences

Compared to previous work [13, 24], we observe a larger variety of preference types for restaurant features. We labeled these preference types in our data and enumerate them in Table 3, where they are sorted by the frequency in which they appeared in a dialog.

Specifically, for the ParlAI [13] and CRM [24] datasets, researchers only defined features such as cuisine (e.g., Italian, Indian), location (or state and city information), price range, rating and party size. In contrast, we summarize 19 mentioned preference types in our dataset in Table 3, which underscores the importance of mining natural dialog for the range of user preferences that a domain-specific conversational recommendation system should support.

Table 3: Types of restaurant features mentioned by users.

Feature (Code)	Examples	Freq. (max 12)
Cuisine Type (CUT)	Indian, Italian	10
Price Range (PR)	Above average, $$	10
Location (LOC)	"...near Queen and University (intersection)."	10
Type of Meal (TM)	Dinner, Brunch, Lunch	8
Menu (MNU)	"I want a diverse menu."	8
Reservation (RVT)	"A reservation for two."	6
Parking (PKG)	Private Parking/Nearby Parking	5
Type (TYP)	"good for family"	5
Ambience (AMB)	Classy, Trendy	4
Noise level (NOL)	Quiet, Average, Loud	3
Outdoor Seating (OSN)	Patio, Balcony, Rooftop	3
Private room (PRO)	"Do they have a private room?"	3
Scenic View (SV)	"I would prefer a place that has a nice view."	3
Alcoholic Drinks (DRN)	Fullbar	2
Popular times/Wait times (WT)	"Average wait time is 15 minutes."	2
Rating (RTN)	Highly rated, Four stars	2
Reviews (RVW)	"...show me a strongly negative review."	1
Picture (PIC)	"Do you have pictures of the restaurant?"	1
Dine Time (DT)	"...around 10:30 would be good for us."	1

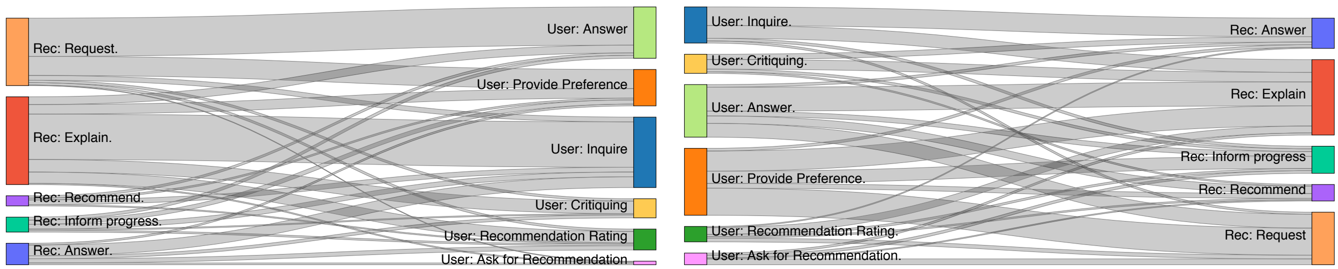
5 WORKFLOW ANALYSIS

5.1 Global Analysis

We now proceed to our workflow analysis that is a key contribution of this work. We start by providing a global analysis of the interactive conversational workflow in our experiment. To achieve this, we present a Sankey diagram[3] in Figure 1 that shows the workflow of the user intents and recommender actions during the conversation, while splitting the Sankey diagram into recommender actions–to–user intents and user intents–to–recommender actions for more clarity. Here, we have collapsed user intents and recommender actions into higher-level categories as noted in Tables 1 and 2. For example, Figure 1a shows recommender actions to request information, explain a recommendation, or make a recommendation, for which the user could answer, provide preferences or make critiques.

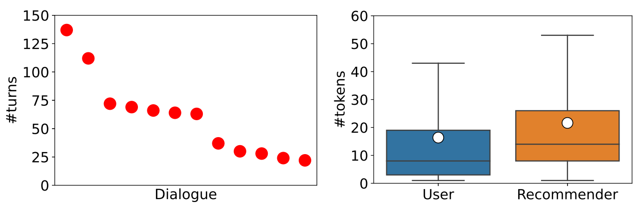
Overall, we observe that the flow in Figure 1a is dominated by (i) the recommender requesting information from the user who provides answers or preferences, and (ii) the recommender explaining recommendations leading to user inquiries that obtain additional information. In this direction of the flow, the recommender tends to mainly collect information and user preferences to provide relevant information. However, in Figure 1b, the flow is more diverse, but we still see dominant flows for question-response style interaction. In order to analyze and understand global conversational statistics, we

[3]A Sankey diagram emphasizes the major transfers or flows within a process with the overall height of the flow proportional to the fraction that it represents in the data. Hence, it can help identify the most important sources and destinations for a flow.



(a) Recommender to user. **(b) User to recommender.**

Figure 1: Sankey diagram for global conversational flow.



(a) Turn distribution. **(b) Token count per turn.**

Figure 2: Global statistics.

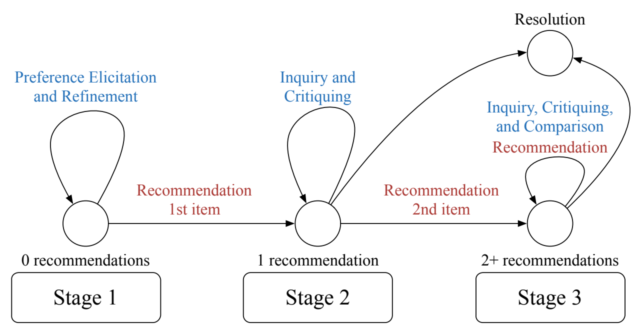
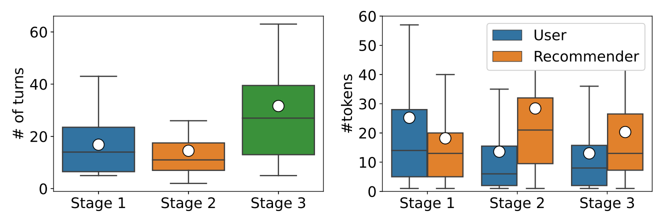


Figure 3: Stages of conversational flow used in our analysis separated according to number of recommendations made. Abandonment could happen at any stage, but is not shown as it did not occur in our data.

refer to Figure 2. Here, we observe that the length of conversations roughly follows a power law distribution, where the longest conversation contains 132 turns and the shortest conversation contains only 24 turns. Also, we observe that in terms of the total token (word) count, the recommender tends to converse longer with the goal of collecting preferences, explaining recommendations, and answering questions as observed in Figure 1a.

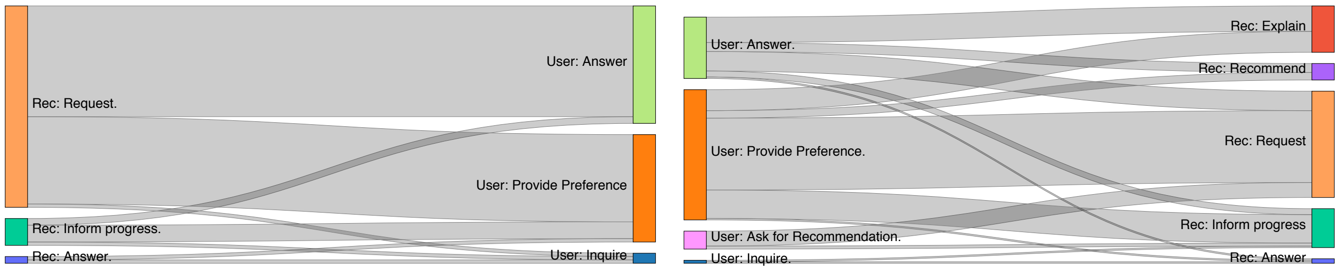
The observations so far based on the global perspective reflect basic intuitions but are somewhat limited in providing novel insight. To delve deeper, we leverage our understanding of the transcripts, where we observed distinct conversational behavioral patterns before and after the first recommendation was made and after subsequent recommendations. This suggests that we can break the conversation down into three distinct stages as illustrated in Figure 3. In summary, we observe that we can cleanly and unambiguously break the conversation for each dialog transcript into the stage with 0 recommendations (Stage 1), 1 recommendation (Stage 2), and 2 or more recommendations with comparison (Stage 3).

With this subdivision of turns, we can now revisit Figure 2 to see if there are distinct characteristics of each stage. Indeed, Figure 4 shows that there are different statistics for each stage that support our split. In particular, Figure 4a shows the #turns in Stage 1 is higher than #turns in Stage 2, which indicates that the preference elicitation stage takes more time. Stage 3 is longer as it takes more turns for clarifying and refining the recommendation before making a final choice. Figure 4b shows the distribution of the number of tokens for each turn, per stage, and per role. We observe that for Stage 1, the user generally uses more words (tokens) to express her preferences, then in Stage 2 and Stage 3 we observe different patterns where the number of words drops for the user and it increases for the recommender. This indicates that during the recommendation refinement and clarification phase, the recommender is more expressive — this is mainly due to the explanation action, which dominates recommender actions in this stage. This analysis further motivates a fine-grained stage-wise analysis of the global workflow as discussed in the next section.



(a) Turn distribution. **(b) Token count per turn.**

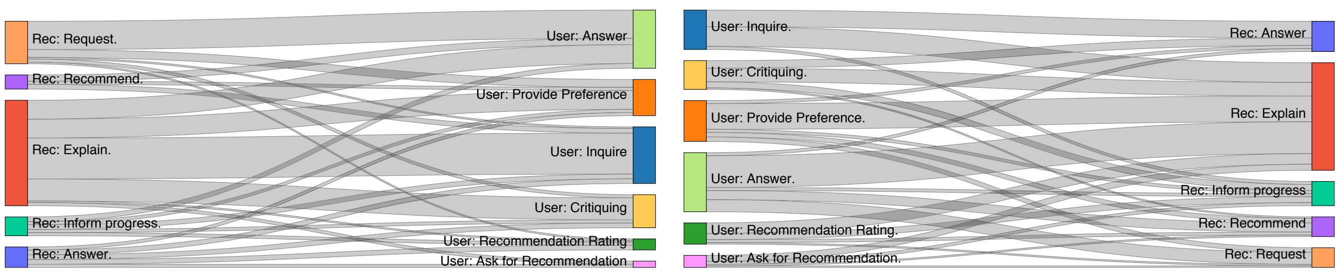
Figure 4: Stage-wise statistics.



(a) Recommender to user.

(b) User to recommender.

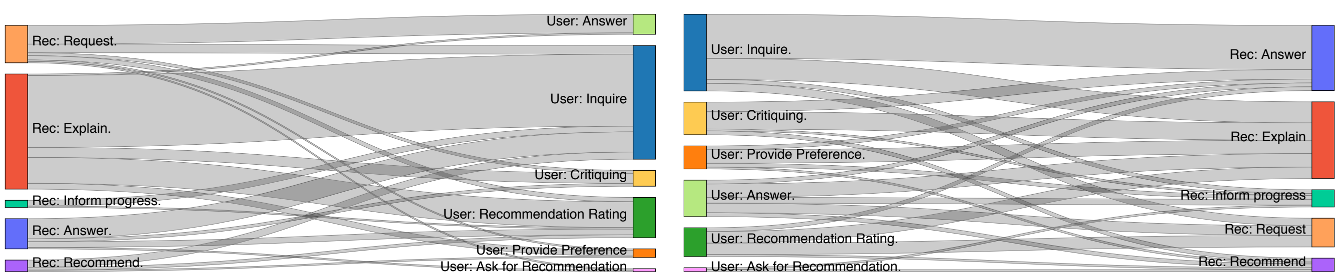
Figure 5: Detailed conversation flow for 1st stage (Preference Elicitation).



(a) Recommender to user.

(b) User to recommender.

Figure 6: Detailed conversation flow for 2nd stage (Inquiry and Critiquing).



(a) Recommender to user.

(b) User to recommender.

Figure 7: Detailed conversation flow for 3rd stage (Inquiry, Critiquing, and Comparison).

5.2 Dialog Stages

Based on Figure 3, we now proceed to a workflow analysis of the different stages in this diagram: **Preference Elicitation** (Stage 1) and **Inquiry and Critiquing Stage** (Stages 2 and 3). Due to the cooperative nature of our study participants, we did not observe user abandonment; however, this would be an important user action to model for most deployed CRSs.

5.2.1 Preference Elicitation (Stage 1). In this phase that is illustrated in Figure 5, a user and a recommender reach an initial level of agreement for user preferences, which lays the groundwork for the subsequent recommendation dialog. In particular, we observe in Figure 5a that the recommender-to-user interactions are dominated by the recommender requesting information, where for half of the interactions, the user provides preferences and the user answers requests to provide more context. On the other hand, for the user-to-recommender interactions in Figure 5b, they are mainly dominated by the user providing preferences for which the recommender asks clarifications or provide explanations. These forward and backward interactions between the user and the recommender illustrate the flow of actions and messages exchanged during this stage that are focused primarily on preference elicitation, refinement, and clarification (which also correspond to the cold-start nature of the recommendation dialogs). This stage naturally ends with the recommender providing a first recommendation. In our study, user acceptance of the recommendation is not required to proceed to Stage 2. In fact, once the user accepts a recommendation in Stage 2 or 3, the dialog proceeds directly to the *Resolution* stage.

5.2.2 Inquiry and Critiquing (Stages and 3). This stage starts with dialogs centered on the initial recommendation made in Stage 1. Due to the complexity of interactions in this stage, we have broken it down to 1 recommendation (Stage 2) and 2+ recommendations (Stage 3) for more clarity as illustrated in Figure 6 and 7 respectively. Unlike Stage 2, Stage 3 allows *comparison* of recommendations since there are 2+ recommendations to compare.

Overall, in Stage 2, a typical user *Inquires* more than she *Provides Preferences* by making information requests as well as critiques on the items. In particular, in Figure 6, we observe that a recommender provides more explanation compared to the previous stage, and often this leads to the user providing additional information (presumably in response to the explanation), and otherwise leads to a non-informational user response. Also, we observe that recommendation tends to lead the user to provide more information, potentially due to the accompanying explanation.

Regarding the last stage beyond the second recommendation, we observe in Figure 7a that the dominant flow is the recommender *Explain* to user *Inquire*. This is because explanation leads the user to further *Inquire* (or *Critiquing*) or often an acceptance or rejection. On the other hand, we observe in Figure 7b that the dominant flow is from a user *Inquire* to recommender *Answer*, which indicates that the user is simply clarifying details of the recommendation.

5.2.3 Summary Observations for all Stages. Overall, we observe a clear shift over time from heavy preference statements at Stage 1 to iteration at Stage 2 on inquiries and explanations, and a final clarification Stage 3 with a high rate of information inquiries, justification of explanations, and critiques as well as final choice resolution — leading often to a final user rating (e.g., accept, reject, etc.). The goal of these heavy iterations is to clarify, refine, and justify recommendations, which we remark is very different from conversational QA systems, where there is no need to clarify the precise meaning of a user's preferences (since there are no preferences in most QA).

Finally, we show in Figure 8 the top words used by a user and a recommender in our dialogs throughout the three stages that we obtained (using a *tf-idf* weighting). We can clearly observe that for Stage 1, the user uses a context word (e.g., they mention "startup") and the recommender seems to be negotiating the meaning of preferences (e.g., using words such as "like", "want", "downtown"). For Stage 2, the discussion seems fairly generic, where the user is expressing her preferences (e.g., "free parking", "like", "option", "layout", "drive", etc.), whereas the recommender uses explanation words (e.g., "good", "like", "street", "option"). Lastly, during Stage 3 we see more explanation, comparison, more specific details like "Indian" and also evidence of resolution/confirmation (e.g., "yea", "OK", "good", "like"). These word-level observations reflect what we have largely observed from the previous workflow analysis.

5.3 Informational Questions as Preferences

In our data, we observed a large number of information questions (labeled as *Inquire*) raised by users, which we conjecture to be an implicit expression of user preferences. For example, the question "Do they have any parking spots?" can serve as a question *and* an expression of preference.

Table 4: Initiator of a second recommendation (used for mixed-initiative analysis) and features mentioned in information questions and subsequent recommendations.

Dialog #	Rec. initiator	Inquired Feature
1	user	parking, picture, ambience (✓), reservation
4	rec	location (✓)
6	user	N/A
7	rec	noise level, price (✓), reservation (✓)
8	user	ambience (✓), price, scenic view
12	rec	location (✓)

Note: ✓denotes when a feature is mentioned in an information question, and later influenced how the recommender provides recommendation.

To verify this conjecture, we examined the dialogs with at least two recommendations present, since most information questions occurred after the first recommendation was made. We counted the number of times that the user asked an information question and then a later recommendation was explained based on the implied preference; we divided this by the total number of times that an information question was asked and a later recommendation was provided. Considering the data collected in Table 4 for dialogs with at least two recommendations, we observed that exactly 50% of the time, an information question was taken into account in a subsequent recommendation by the human recommender. Since human recommenders frequently considered the information questions as part of the preference requirements in their subsequent recommendations, we note that information questions do indeed appear to serve a dual role as implicit soft preferences. This appears to be a novel observation, and we consider it a crucial insight that should be leveraged in the consideration of user preference modeling in future dialog-based CRS designs.

5.4 Soft Preferences and Hard Constraints

Now we proceed to analyze the difference between **Hard Constraints** and **Soft Preferences** expressed by the user and how their expression differs over the stages of the conversation. We define Hard Constraints as essential pieces of information either explicitly provided by the user or requested from the recommender that are required in all recommendations. In contrast, Soft Preferences are not required to be satisfied.

We would claim that in our data, preferences gathered through *Provide Preference*, *Request* and *Critiquing* are hard constraints. Users tend to provide the most important features that they look for in a recommendation at the beginning of the dialog through *Provide Preference*. Subsequently, recommenders might recognize some crucial information that is missing and request those features, or ask questions for clarification. We consider information exchanged through the recommmender *Request* also to be hard constraints since they are either indispensable information for initializing a recommendation, or pivotal in narrowing down the scope of recommendation candidates. After the recommendations are made,

(a) Recommender Stage 1.

(b) Recommender Stage 2.

(c) Recommender Stage 3.

(d) User Stage 1.

(e) User Stage 2.

(f) User Stage 3.

Figure 8: Cloud of top words for each role and each stage.

the user might critique to reject the recommended item or add further constraints. *Critiquing* shows preference components from a user that are mandatory for accepting a recommendation, thus we consider *Critiquing* as hard constraints.

In contrast, we note that Soft Preferences often arise implicitly from users' information questions. As previously discussed, human recommenders take in those information questions and attempt to include the mentioned features in subsequent recommendations. Thus we consider information that users *Inquire* about to be soft preferences that are not necessarily required.

Considering how users stated their information questions, we generally observed that although users initiated these inquiries to learn more about the recommended item, most of the time the answers did not make users change their minds. Users often either claim that the inquiry is not a hard constraint (e.g., "I'm just curious, and this is not a hard requirement because my parents don't like a noisy place.") or simply did not reject the recommendation. In very few cases, we observed users *Inquire* and then *Critique*, e.g., User: "...do they have a patio?" → negative response from the recommender → User: "OK. So could you find me a place with a patio...", which represents a user's discovery of a hard constraint that stemmed initially from an information inquiry.

More quantitatively in Figure 9, we observe the distribution of hard constraints and soft preferences before (Stage 1) and after (Stages 2 & 3) an initial recommendation is made. Most elicited preferences are hard constraints (98.3%) and very few soft preferences (1.7%) are expressed before the initial recommendation. In later stages of dialog, soft preferences take a more dominant percentage (58.7%) over hard constraints (41.3%), since users can ask for additional information when there are concrete items to discuss.

5.5 Anecdotal Observations

In this section, we briefly comment on a number of non-quantitative anecdotal observations we made from the conversation transcripts that we believe are simply interesting to mention as we bring our workflow analysis to a conclusion.

5.5.1 Mixed Initiative. One aspect of a conversational system is that it will engage in *mixed-initiative interaction*. We observed a mixed initiative for subsequent recommendations after the initial recommendation. In our data shown in Table 4, the number of times a recommender proposed an alternative recommendation and the number of times a user explicitly requested another item are the same, indicating balanced initiative on behalf of the user and recommender.

Anecdotally, the user might request another recommendation to compare with the current item that already fulfills all the hard constraints (e.g., "Awesome. Yes, so far I think all my requirements are there so maybe you can give me more choices so I can pick one?"). The recommender might propose a new recommendation either by directly presenting the item (e.g., "So there is another good restaurant. It is called Sang's Great Seafood Restaurant.") or by asking the user if they would like to hear another recommendation (e.g., "Anything else I can do for you today? Or do you want me to recommend more just in case you want to try a different one?"). Hence, we found that sometimes the recommender takes the initiative with the interaction while the user works to assist it, contributing to the interaction as needed. At other times, the roles are reversed. This strategy leads to a more natural conversation, and we might infer that a good recommender system design should

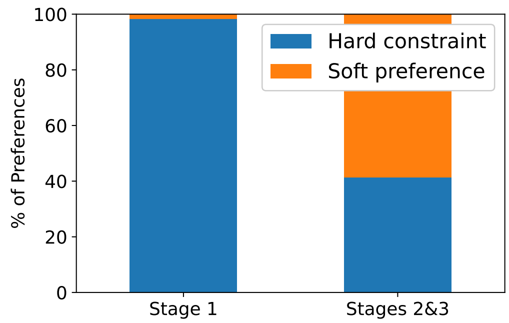


Figure 9: Distribution of Hard Constraints and Soft Preferences before (Stage 1) and after (Stages 2 & 3) the initial recommendation is made.

attempt to determine whether the user or recommender should take the initiative at each conversational stage.

5.5.2 Negotiation. When there exists a potential preference conflict, users and recommenders are observed to negotiate on this specific requirement. For example, the user might propose a relaxation on a previous preference under the impression that it is potentially unsatisfiable: "I think if you are looking for something negotiable, I am kind of fine with negotiating to sit inside the restaurant as well. Because it is not that easy to find seating on a rooftop or a balcony... Not having outside seating is fine. Just maintaining a certain noise level when we are sitting inside should be fine." The recommender, when facing a preference not available in the recommended item but which can be fulfilled with a substitute, would suggest it to the user, e.g., "So you would like to reserve a private room?" → "Yeah absolutely" → "If there is a room without the door. So it is just open space but separate from others, is it okay?". Hence, we see that the recommender is aware of the possible (or available) options and the user, sometimes only partially, knows her preferences. As the dialog proceeds and the user conveys her preferences, a good recommender needs to recognize what to retain and what to compromise for the user in preference conflicts.

5.5.3 Comparison of Recommendations. As previously mentioned in the dialog stages section, when there exists more than one suitable candidate, humans compare and contrast on the recommendations. Users would *Critique - Compare* and typically ask for a comparison of the items with regard to a specific feature in one turn. For example, the user might ask among the two recommendations "Which restaurant is less busier?", or ask for the other item's information with the same aspect: "Can you check if the other restaurants have similar comments or reviews?". Recommenders use *Personal Opinion - Comparison* to compare and contrast on the recommendations and make suggestions on a decision. As an example, aside from vanilla comparisons on prices and ratings, we observed one (human) recommender contrasting a multiple-location chain restaurant with a single location restaurant, suggesting the chain to be a more flexible choice with multiple possible locations.

6 CONCLUSION

In this paper, we have provided a workflow analysis of human-to-human dialogs for context-driven conversational recommendation. Specifically, we recorded the conversations between a human recommender and a human user who jointly attempt to find a restaurant that best fits the user's preference given a contextual scenario. On analyzing their conversational utterances, we have identified and labeled our data according to user intents and recommender actions that extend the existing literature. In terms of conversational workflow analysis, we have identified clear distinctions in conversational behavior as a function of the recommendation stage of the conversation (i.e., no recommendations, one recommendation, and two or more).

Overall, this conversational workflow analysis has allowed us to make a number of observations regarding the critical importance of preference refinement, the role of information queries as implicit soft preference critiques, and the importance of explanation in driving the overall conversational flow. Moreover, our studies point to several potential research directions that could assist in the design of future CRSs such as the following:

- Given the amount of the conversation devoted to iterating on preference understanding and the level of specificity (e.g., "you said chicken fingers, did you just mean a kid's menu?"), there is a need for research on disambiguation and clarification of user preferences and intents through natural language interaction to support actionable recommendation.
- Since users often expressed their preferences through informational questions, it is important to be able to accurately mine such preferences directly from natural language questions in parallel with the more explicit task of natural language question answering to respond to the user inquiry.
- Given the prevalence of personal experience and opinion explanations employed by the human recommenders, it may be useful to mine such explanations from user review data as they relate to a user's preferences as a means of building rapport between the CRS and user.
- It is critical to observe the dominance of recommender explanation in Stages 2 and 3 and the way it literally drives the bulk of the interaction with the user — leading the user to provide further inquiries, critiques, and statements of preference. Hence, it is important to consider the exploratory role of recommendation explanations for directing the conversation and providing prompts that implicitly elicit feedback. In this sense, beyond its informational and persuasion roles [28], explanation may be viewed as a way to further explore and reduce uncertainty in the user's preferences.

Altogether, these insights should help us better understand the structure and nature of organic context-driven recommendation and provide useful guidance in the design of the next generation of fully automated and naturally interactive conversational recommendation systems.

Acknowledgments: We thank the reviewers for their insightful comments that have helped us improve the discussion and clarify the paper presentation.

REFERENCES

[1] Derek Bridge, Mehmet H. Göker, Lorraine McGinty, and Barry Smyth. 2005. Case-based recommender systems. *Knowledge Engineering Review* 20, 3 (2005), 315–320.

[2] Wanling Cai and Li Chen. 2019. Towards a Taxonomy of User Feedback Intents for Conversational Recommendations.. In *RecSys (Late-Breaking Results)*. Association for Computing Machinery, New York, NY, USA, 51–55.

[3] Wanling Cai and Li Chen. 2020. Predicting User Intents and Satisfaction with Dialogue-Based Conversational Recommendations. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20)*. Association for Computing Machinery, New York, NY, USA, 33–42. https://doi.org/10.1145/3340631.3394856

[4] Konstantina Christakopoulou, Alex Beutel, Rui Li, Sagar Jain, and Ed H. Chi. 2018. Q&R: A Two-Stage Approach toward Interactive Recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 139–148.

[5] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards Conversational Recommender Systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 815–824.

[6] Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to Discourse. *Cognitive Science* 13, 2 (1989), 259–294. https://doi.org/10.1207/s15516709cog1302_7 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15516709cog1302_7

[7] Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural Approaches to Conversational AI. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 1371–1374. https://doi.org/10.1145/3209978.3210183

[8] Barney G. Glaser and Anselm L. Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research.* Aldine de Gruyter, New York, NY.

[9] Peter Grasch, Alexander Felfernig, and Florian Reinfrank. 2013. ReComment: Towards Critiquing-Based Recommendation with Speech Interaction. In *Proceedings of the 7th ACM Conference on Recommender Systems* (Hong Kong, China) *(RecSys '13)*. Association for Computing Machinery, New York, NY, USA, 157–164. https://doi.org/10.1145/2507157.2507161

[10] E. Holliman, J. Godfrey, and J. McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, Vol. 1. IEEE Computer Society, Los Alamitos, CA, USA, 517–520. https://doi.org/10.1109/ICASSP.1992.225858

[11] Andrea Iovine, Fedelucio Narducci, and Giovanni Semeraro. 2020. Conversational Recommender Systems and natural language:: A study through the ConveRSE framework. *Decision Support Systems* 131 (2020), 113250. https://doi.org/10.1016/j.dss.2020.113250

[12] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2020. A Survey on Conversational Recommender Systems. arXiv:cs.HC/2004.00646

[13] Chaitanya K. Joshi, Fei Mi, and Boi Faltings. 2017. Personalization in Goal-Oriented Dialog. arXiv:cs.CL/1706.07503

[14] Jie Kang, Kyle Condiff, Shuo Chang, Joseph A. Konstan, Loren Terveen, and F. Maxwell Harper. 2017. Understanding How People Use Natural Language to Ask for Recommendations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems* (Como, Italy) *(RecSys '17)*. Association for Computing Machinery, New York, NY, USA, 229–237. https://doi.org/10.1145/3109859.3109873

[15] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). Curran Associates, Inc., -, 9725–9735.

[16] Yukiko I. Nakano, Gabe Reinstein, Tom Stocky, and Justine Cassell. 2003. Towards a Model of Face-to-Face Grounding. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1* (Sapporo, Japan) *(ACL '03)*. Association for Computational Linguistics, USA, 553–561. https://doi.org/10.3115/1075096.1075166

[17] Roberto Pagano, Paolo Cremonesi, Martha Larson, Balázs Hidasi, Domonkos Tikk, Alexandros Karatzoglou, and Massimo Quadrana. 2016. The Contextual Turn: From Context-Aware to Context-Driven Recommender Systems. In *Proceedings of the 10th ACM Conference on Recommender Systems* (Boston, Massachusetts, USA) *(RecSys '16)*. Association for Computing Machinery, New York, NY, USA, 249–252. https://doi.org/10.1145/2959100.2959136

[18] Florian Pecune, Shruti Murali, Vivian Tsai, Yoichi Matsuyama, and Justine Cassell. 2019. A Model of Social Explanations for a Conversational Movie Recommendation System. In *Proceedings of the 7th International Conference on Human-Agent Interaction* (Kyoto, Japan) *(HAI '19)*. Association for Computing Machinery, New York, NY, USA, 135–143. https://doi.org/10.1145/3349537.3351899

[19] Chen Qu, Liu Yang, W. Bruce Croft, Johanne R. Trippas, Yongfeng Zhang, and Minghui Qiu. 2018. Analyzing and Characterizing User Intent in Information-Seeking Conversations. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 989–992. https://doi.org/10.1145/3209978.3210124

[20] Arpit Rana and Derek Bridge. 2020. Navigation-by-Preference: A New Conversational Recommender with Preference-Based Feedback. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 155–165. https://doi.org/10.1145/3377325.3377496

[21] Hideo Shimazu. 2002. ExpertClerk: A conversational case-cased reasoning tool for developing salesclerk agents in e-commerce webshops. *Artificial Intelligence Review* 18, 3-4 (2002), 223–244.

[22] Barry Smyth. 2007. *Case-Based Recommendation.* Springer Berlin Heidelberg, Berlin, Heidelberg, 342–376. https://doi.org/10.1007/978-3-540-72079-9_11

[23] Barry Smyth and Lorraine McGinty. 2003. An analysis of feedback strategies in conversational recommenders. In *the Fourteenth Irish Artificial Intelligence and Cognitive Science Conference (AICS 2003)*. Citeseer, -, -, 218–223.

[24] Yueming Sun and Yi Zhang. 2018. Conversational Recommender System. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 235–244. https://doi.org/10.1145/3209978.3210002

[25] Taavi T. Taijala, Martijn C. Willemsen, and Joseph A. Konstan. 2018. MovieExplorer: Building an Interactive Exploration Tool from Ratings and Latent Taste Spaces. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing* (Pau, France) *(SAC '18)*. Association for Computing Machinery, New York, NY, USA, 1383–1392. https://doi.org/10.1145/3167132.3167281

[26] Cynthia A Thompson, Mehmet H Goker, and Pat Langley. 2004. A personalized system for conversational recommendations. *Journal of Artificial Intelligence Research* 21 (2004), 393–428.

[27] N. Tintarev and J. Masthoff. 2007. A Survey of Explanations in Recommender Systems. In *2007 IEEE 23rd International Conference on Data Engineering Workshop*. -, -, 801–810.

[28] Nava Tintarev and Judith Masthoff. 2011. *Designing and Evaluating Explanations for Recommender Systems.* Springer US, Boston, MA, 479–510.

[29] D. R. TRAUM. 1992. A Speech Acts Approach to Grounding in Conversation. *Proceedings of International Conference on Spoken Language Processing (ICSLP'92)* - (1992), 137–140.

[30] Vivian Tsai, Timo Baumann, Florian Pecune, and Justine Cassell. 2019. Faster Responses Are Better Responses: Introducing Incrementality into Sociable Virtual Personal Assistants. In *9th International Workshop on Spoken Dialogue System Technology*, Luis Fernando D'Haro, Rafael E. Banchs, and Haizhou Li (Eds.). Springer Singapore, Singapore, 111–118.

[31] Jesse Vig, Shilad Sen, and John Riedl. 2011. Navigating the Tag Genome. In *Proceedings of the 16th International Conference on Intelligent User Interfaces* (Palo Alto, CA, USA) *(IUI '11)*. Association for Computing Machinery, New York, NY, USA, 93–102. https://doi.org/10.1145/1943403.1943418

[32] Norha M. Villegas and Hausi A. Müller. 2010. *Managing Dynamic Context to Optimize Smart Interactions and Services.* Springer-Verlag, Berlin, Heidelberg, 289–318.

[33] Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A Dialogue Dataset with Additional Annotation Corrections and State Tracking Baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*. Association for Computational Linguistics, Online, 109–117. https://doi.org/10.18653/v1/2020.nlp4convai-1.13

[34] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards Conversational Search and Recommendation: System Ask, User Respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (Torino, Italy) *(CIKM '18)*. Association for Computing Machinery, New York, NY, USA, 177–186. https://doi.org/10.1145/3269206.3271776