This is a repository copy of *Conversational agents for recipe recommendation*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/156597/

Version: Accepted Version

Conversational Agents for Recipe Recommendation

Sabrina Barko-Sherif
University of Regensburg
Regensburg, Germany
sabrina.barko-sherif@stud.uni-
regensburg.de

David Elsweiler
University of Regensburg
Regensburg, Germany
david@elsweiler.co.uk

Morgan Harvey
Northumbria University
Newcastle upon Tyne, United
Kingdom
morgan.harvey@northumbria.ac.uk

ABSTRACT

As technology improves, the use of conversational agents to help users solve information seeking tasks is becoming ever more prevalent. To date we know little about how people behave with such systems, particularly in diverse contexts and for different tasks, their specific needs or how best to support these. By employing a Wizard of Oz (WoZ) methodology and developing a conversational framework, in this work we study how participants (n=28) interact with such a system in an attempt to solve recipe recommendation tasks. Our results are mostly encouraging for the future development of conversational agents in this context, however, they also provide insights into the complexities of building such a system that could convincingly engage with users in productive, human-like conversations.

CCS CONCEPTS

• **Information systems** → *Users and interactive retrieval*; • **Human-centered computing** → *Empirical studies in HCI*.

KEYWORDS

conversational agents, recipe recommendation

1 INTRODUCTION

With recent improvements in Machine Learning allowing computers to better interpret human language, search and assistance systems are becoming increasingly conversational in nature. In this paradigm, the user and machine communicate via audio, or via short text/chat messages, where simply reading out a list of potentially relevant items (i.e. a traditional SERP) is not an appropriate response - the system must become a more *active* partner in the task [37]. This active, conversational element means that a more nuanced understanding of how human beings engage in dialogues to resolve problems and complete tasks, and the stages of those

conversations, is necessary to obtain good performance from the user's perspective.

Examples of such systems include: digital help services, such as those provided by companies via social media[1]; chat-bots, which automate such services [5, 40]; and voice-based interaction devices, such as Amazon's Echo, Microsoft's Cortana and Apple's Siri, which are becoming an increasingly common feature of modern households and automobiles and have propelled conversational search research in our field. As a result, information seeking conversations now take place in diverse situations, embedded in users' everyday lives, providing a very different mode of interaction to traditional information access and retrieval systems.

An audio dialogue-based interface is not always appropriate but is particularly useful in certain situations, such as when driving or cooking, where interaction with a screen, keyboard and mouse may be difficult and, as such, traditional interaction modes are unsuitable [15]. Despite the recent, increased research focus in this area (e.g. [28, 33, 38]), existing systems are quite limited in the level of complexity of conversation in which they can participate. We still know little about how users can be supported to complete tasks conversationally in correspondence with a machine, how requests and information are communicated by the user and how appropriate such an interaction paradigm is for, for example, services such as recommendation [29].

In this work we present the results of a Wizard of Oz (WoZ) study designed to investigate how aspects of conversational interaction vary for a recommendation task in different conditions within the specific setting of home cooking. The study was completed by 28 participants, who were randomly assigned to one of two conditions: text-based and audio-based interaction, and were asked to interact with the Wizard, which they thought was a bot, to complete three recipe finding tasks.

We believed, based on the literature, that the kitchen domain would be a fertile context for the kinds of complex needs suited to conversational search [8, 38], offering situations where users simultaneously perform practical, sometimes cognitively challenging tasks, which make searching in the traditional sense problematic. People increasingly locate conversational agents in their kitchens [25] and use them for simple cooking-related tasks, such as setting timers [16]. More challenging tasks, like discovering recipes to cook, are not well supported. Here, we focus on recommendation because this is an important, understudied conversational IR task [43] and one of the key needs conversational agents should support in the kitchen domain [14]. A final motivation for studying this domain is, as has been argued in the past, that assistance in the kitchen, including health-oriented food recommendation, could

---

[1]e.g. https://twitter.com/KLM

offer societal benefit [10]. Getting people to cook (more healthily) has been suggested by several government health agencies as a way to improve nutrition [24, 35] and providing assistance may lower the barriers to healthier cooking.

We analyse the collected data to understand how users communicate their needs and negotiate with the system to receive appropriate meal recommendations. 1) we analyse the flow of interaction within sessions to understand the make up of conversations and how this can vary. 2) we examine aspects of the language used to determine whether observations from previous conversational search behaviour studies transfer to this specific task and context. 3) we compare spoken interaction, as would be employed with devices such as Google home, to typed input, which a user might utilise on a mobile device, such as a smartphone. The results of our analyses provide clear insights into the feasibility of implementing a conversational agent for this task, as well as the challenges that need to be addressed before this can be fully realised.

2 LITERATURE REVIEW

To set the context for our work and motivate the precise research questions and methodological decisions taken, we summarise three bodies of related work: Section 2.1 details the tradition of studying conversations in information science; Section 2.2 reviews recent work on conversational search in our field; and finally, Section 2.3 relates this work to the kitchen domain by detailing research on information behaviour in the kitchen, including literature on food recommendation research.

2.1 Conversations in Information Science

Information science has a strong tradition of studying conversations, particularly between librarians and library patrons, where the aim is to establish user needs and how these are communicated. Many of the key models, which have guided information seeking research for decades, were conceived by analysing such conversations (e.g. [3, 21, 31]). Such studies illustrate the complexity of conversations [26], the difficulties of eliciting user needs and the fact that, particularly early on in the process, users are often unable to clearly articulate what they want [31]. Belkin's research used dialogues to model information seeking behaviour and later dialogue structure inspired the design of interactive IR systems [4, 42]. Pejtersen [26] analysed 134 user-librarian conversations about fiction in Danish public libraries, uncovering diverse user strategies, as well as the need for technology to support librarians assist library users. Prekop [27] studied information seeking using a complex, real world example of collaborative information seeking activity, drawn from the military domain. In doing so patterns were identified describing prototypical behaviours, actions and interactions of participants in collaborative tasks.

2.2 Conversational Search in IR

Recently, the IR community has tried to move toward conversational interfaces with non-human agents, enabling a similar kind of interaction to human collaboration, rather than via a traditional search interface. Radlinski and Craswell [28] describe a conversational search system as "...a system for retrieving information that permits a mixed initiative back and forth between a user and agent, where the agent's actions are chosen in response to a model of current user needs within the current conversation, using both short- and long-term knowledge of the user". They stress the importance of supporting a dialogue between the user and the system and that, to build such systems, it is necessary to be able to clearly define the permissible steps within these dialogues. Luger and Sellen [23] interviewed 14 users of Conversational Agents about their everyday use and found their expectations to be "dramatically out of step with the operation of the systems, particularly in terms of [...] system capability and goals."

Conversational interfaces can vary enormously, with chat-bots that communicate with textual messages on one end of the scale (e.g. [5]), and fully flexible human dialogues on the other (e.g. [14]). Concentrating purely on spoken vs typed queries, Guy [17] discovered that the language of voice queries is closer to natural language than typed queries. Voice queries are not only longer than text queries, but also use richer language and are more likely to contain complex phrases and questions.

Many spoken dialogue systems employ clearly-bounded frameworks for problems within a closed and fixed domain [41]. This might be booking a flight or a concert ticket, where users' utterances can typically be trivially mapped to pre-defined schema (so-called "slot filling"). Although some approaches do permit more free-form user inputs, these are typically "resolved" by delivering the results of a normal web search with the user's input as a query [19]. Changing the level of freedom with which users can interact has consequences for how users behave with such systems. Dubiel et al. compared a slot-filling system against a fully conversational system, which used spontaneous, human-like dialogue, and found the conversational condition to be associated with fewer, shorter utterances and achieved better task outcomes [9].

Offering more freedom has been shown to reveal behaviours resembling those in human-to-human collaboration. Vtyurina [38] showed, for example, that people do speak naturally with agents, with interactions sharing many of the qualities of human-to-human speech. Examples include using short positive utterances, such as "yup" or "ok" to implicitly signal a move to the next step and "grounding behaviour" whereby participants, despite being unaware which parts of their speech the system can and cannot understand, still respond to system statements. The purpose of these responses is to let the other speaker – the agent – know that the information has been processed and accepted.

In studies designed to mimic conversational search situations, where pairs of users conversed during collaborative tasks, patterns have been identified resembling those of Prekop's work [37]. Trippas et al. [36] summarise their experiences from multiple studies of different types and describe utterances ranging from terse "query-like" utterances to "teleporting", where users describe their need in great detail in an attempt to move directly to the solution [32]. This contrasts with Vtyurina's results [38], which suggested that humans converse with digital agents much like they would with other humans.

In a theoretical contribution, Azzopardi et al. [2] list different kinds of interactions or interaction goals users can have with a conversational search system. *Reveal Actions* disclose details about their information need and including efforts to refine or expand information already provided; *Inquire Actions* ask about the options

presented; *Interrupt Actions* interject when the agent is providing long, detailed information; *interrogate actions* are attempts by the user to learn about what the agent knows or can do. Finally, *Closing Actions* act on information or suggestions provided by the agent.

There are also many variants regarding how agents may respond to user statements. Azzopardi et al. list different kinds of actions agents can perform, including: *Inquire Actions*, to elicit information or filtering criteria from the user; *Reveal Actions*, where information is provided to the user in response to questions or requests; *Suggest actions*, i.e. provide recommendations and *Explain actions*, where actions taken are justified.

The kinds of language used by the system is also important. Thomas et al. [9] distinguish several conversational styles and discovered that divergent styles between agents can be strenuous for users. One take-away from the work is that conversational systems should be capable of adapting to the user's style to reduce workload.

Thus, the evidence suggests the existence of a great variety of interaction patterns with such systems. The interactions can share properties of human-to-human conversation, but this is not a given and depends, to a large extent, on the configuration of the system.

Much of the research performed to date has been domain agnostic. That is, researchers have either attempted to provide models which will work in general settings, such as web search or make generalisations when studying tasks in one setting; e.g. planning a vacation [30] or booking a flight [9]. Here we wish to take a different approach. Building on the work of [38] and [14], we will study one particular task in a specific domain - conversational food recommendation in the kitchen - and, in doing so, determine whether (and how) the observations of other scholars transfer to this setting.

2.3 Information Behaviour in the Kitchen

Analysing questions posed on the Google Answers forum relating to cooking, Cunningham and Bainbridge established 17 varieties of cooking information needs [8]. This taxonomy was the starting point for a study of the kinds of information needs that occur in a conversational cooking setting [14]. Unsurprisingly, both efforts reveal the need to find recipes as one of the most common needs.

Recipe recommendation has been studied both algorithmically (e.g. [13, 18]) and in terms of how users respond to recommendations in user studies [11] (see [34] for a recent comprehensive review). Moreover, there is literature on conversational recommendation dating to the early noughties [7]. However, similar to the conversational search literature, there are limited studies of how human users interact with these systems and none in the context of food recommendation. In this work we present a study of this type, which investigates the conversational negotiation process involved in selecting (an) appropriate recipe(s). The aims being: to understand whether the features of conversations summarised above transfer to this specific information need type in the domain of the kitchen; and to investigate the feasibility of implementing a working system in the near future. The research questions are, therefore, as follows:

- RQ1: How do people communicate needs conversationally?
- RQ2: How do people communicate satisfaction / dissatisfaction with recommendations?

- RQ3: How do these aspects vary with interaction mode (typed vs spoken)?

3 METHODOLOGY
We employed a Wizard of Oz (WoZ) methodology to provide a means to obtain data about how humans might interact with a future intelligent agent in the context of recipe recommendation. In a WoZ study the test participants interact with a system that they believe to be automated but that is, unbeknown to them, actually remotely operated by a human. This technique is useful when one wishes to learn how humans behave with a system that is currently technologically infeasible.

When using such a methodology, there is a trade-off to be made between a "system" that can respond to anything, and therefore provides maximum conversational flexibility, and one that is entirely scripted and procedural, and therefore is more believable and implementable. We wished to investigate how conversations might work with an agent that would be plausible in the relatively short-term, whilst still allowing the Wizard some limited flexibility when replying to user utterances. While not a fully-scripted, slot-filling approach, utilising a stricter mode of interaction offers several advantages: believability in the setup (i.e. increased ecological validity); reliability (easier comparison across tasks and participants because of less variability in system behaviour); increased Wizard response speed; easier to move to next step, which would be to build such a system.

3.1 WoZ Conversation Framework
We developed a *conversational* framework (see Figure 1) describing how an idealised dialogue between a human user and the virtual agent (named *Telefood*) should proceed in this context - i.e. recipe recommendation. The framework is depicted as a flow chart with nodes for user utterances (preceded by *U*) and system responses and queries (preceded by *S*). While we acknowledge the potential for ethical issues to arise in such settings [12], we attempt to mitigate these by ensuring consistent and repeatable responses through the framework. As such, the wizard adhered to the following conditions:

(1) Conversation should follow flow described by the framework
(2) However, if the user asks a question or makes a statement that does not adhere to the framework at the point in the dialogue that had been reached, the system should either:
   (a) Generate an error message of the form "I didn't understand that" and request that the user try again
   (b) Or, if the request was something that would be relevant to another part of the framework, jump to the appropriate framework node and proceed
(3) For each node in the framework, the system has access to pre-defined sets of stock questions and response phrases for each part of the framework, which are drawn stochastically
(4) The system first greets the user by responding to the utterance "Hey Telefood!" (framework nodes U1 -> S1)
(5) The system should then attempt to obtain general information (i.e. not specific ingredients) about the user's requirements (nodes U2.1 and S2.1)
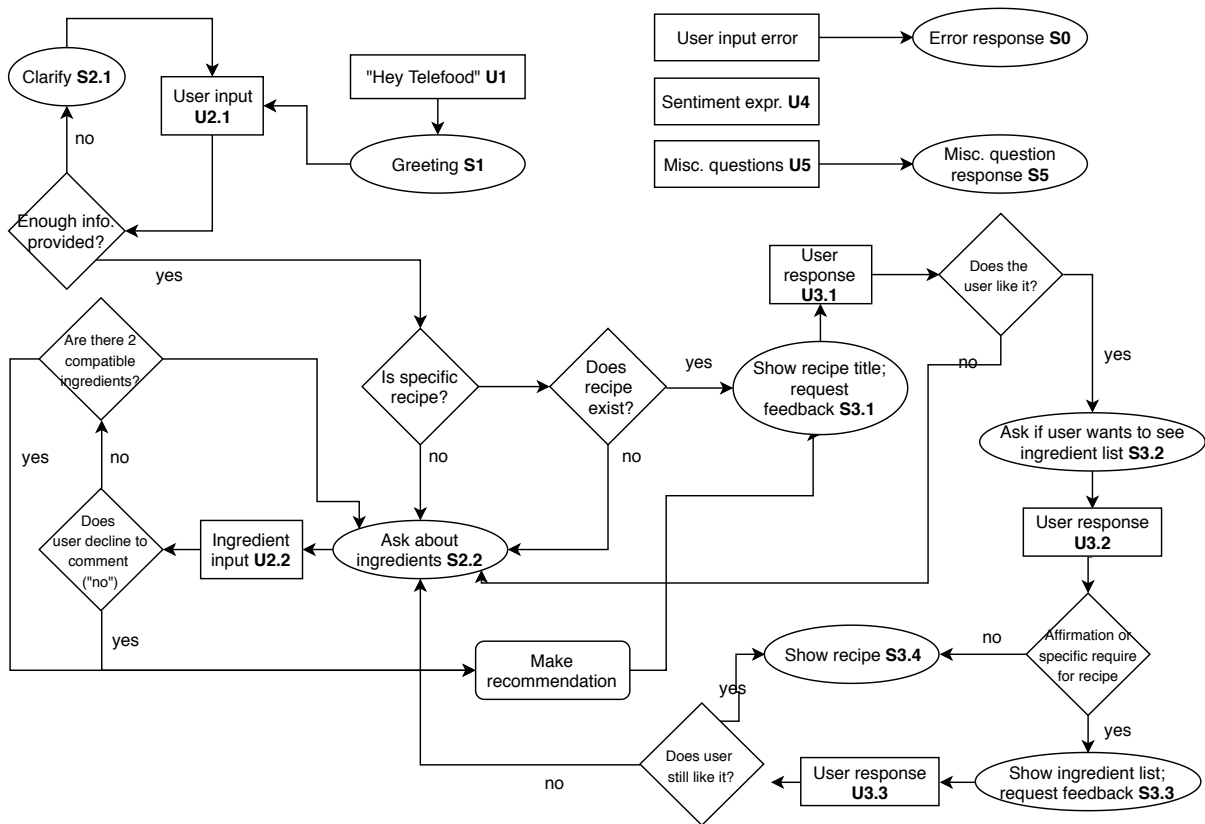(6) The system then proceeds to ask questions regarding preferred or disliked ingredients (U2.2, S2.2)

Figure 1: Conversation framework.

(7) Once the user has provided information through nodes U2.1 and U2.2, which may be repeated, where necessary, the system provides a single recommendation (S3.1), which the user responds to (U3.1)

(8) Any spelling or speech errors made by users should be replied to with an error message (as for 1a above)

We designed our study to mimic a plausible system whereby the user interacts conversationally, either by typing or speaking messages, with the aim of gaining a suitable recipe recommendation. The Wizard Framework provides explicit support for *Inquire, Reveal* and *Suggest* actions via the *S-states*. *Explain* actions are, though, restricted to the communication of suggestions in *S3.1*. We also note that while the framework does not allow all user input to be handled as an actual human would, there is no restriction with respect to what participants can say. Therefore, all of the user actions proposed in [2] could appear in the transcripts. The interaction was performed via the mobile phone app *telegram*, which offers users the possibility to send and receive text or voice messages. To reduce response latency [1] all responses were pre-prepared, copy/pasted as needed and sent as text (not as audio).

3.2 Experimental Design and Participants

We employed a between-groups design, where participants were randomly assigned to one of the two interaction conditions: text or audio. Participants completed 3 out of a pool of 6 plausible tasks, selected at random, each of which required a recipe of some sort to be found. The tasks were derived such that the complexity of the task varied. The literature on task complexity states that complexity can mean different things [6]; complexity in our case depends on the specificity of the problem (open ended vs specific), for example whether the user should look for a birthday cake (specific; e.g. task 2) or just a quick main meal (open ended; e.g. task 1). We will predominantly use tasks 1 and 2 as exemplars:

> Tonight you would like to cook for you and two of your friends, Mark and Lea. You know that Mark has been vegan for several years and, of course, must take that into account. Your friends will arrive in 4 hours at the earliest, so there is no time pressure. *(task 1)*

> You would like to bake a birthday cake for your friend Jasmin. She has a lactose intolerance and, therefore, you would like to avoid or substitute for milk products, where possible. *(task 2)*

Note that all task descriptions ended with the same two sentences: "With the help of the assistance system, search for an appropriate recipe. You can wake the system up with 'Hey Telefood'" to instruct the participant how to start interacting with the system.

28 participants were recruited via advertisements on social-media; internal advertising using within-university communication systems and via word of mouth. The experiments were conducted

in Germany and, as such, all interactions were done in the German language, the native language of all the participants and of the Wizard. Throughout the following, all text has been translated into English in such a way that it most closely maintains the feeling of the original expression. 13 participants identified as male, 15 as female; ages ranged from 18 to 33 with a median of 22; participants study 11 different majors, although the vast majority (18) were digital media and information technology students.

The Wizard used the German recipe website *ChefKoch.de* [2] as a source of recipes for the experiment. The site offers a variety of recipes - over 330,000 at time of writing - which can be filtered by effort, time and/or by keywords. The Wizard could also use Google to search for any (reasonable) information requested by participants, for example descriptions of what a particular ingredient is, allowing the illusion of interacting with a real assistance system to be maintained.

Participants were greeted by another researcher (i.e. not the Wizard) and taken to a quiet lab where they were introduced to the Telegram application, which had already been downloaded and installed on a smartphone. Participants were instructed to use the app to search for, and start a conversation with, a specific user, which they were told was a bot. Depending on the experimental condition assigned, participants were asked to either communicate with the bot via text input or by recording and sending audio messages. Those in the audio condition were requested to speak clearly and distinctly so that the "bot" could understand them.

A task was marked as completed when the participant received a recipe or the ingredient list without asking for more information or objecting, or if the participant explicitly addressed the system again with "Hey TeleFood" to begin the next task. Participants had free choice when it came to the ingredients and there were no seasonal, price-dependent or topical restrictions given to ensure the wizard could retrieve the information fast enough and did not require too much time to filter the recipes. Tasks took between 4 and 20 minutes to complete with the median time to complete being 6 minutes.

4 DATA ANALYSIS

After acquiring the data, the spoken conversations were transcribed for further analysis. We do not analyse phonetic or phonological aspects in this work and, therefore, transcription was only performed at a word level. The audio data was transcribed to standard German and any dialectic expressions and instances of slurring were transcribed in their literal sense, also to standard German. Typed input over multiple lines without the Wizard responding was treated as a single utterance.

We consider each user query and system response individually and coded them using the conversational framework described in the above section. Coding was performed by two researchers, who first worked together to code 25% data set, resolving any disagreements through discussion. They then proceeded to work separately, each coding half of the remaining data. A random sample of 100 rows of these was selected and re-coded by the other researcher in order to check that there was a reasonable level of inter-rater reliability. Cohen's kappa values were obtained for the user frames

and the system frames, both of which indicated almost perfect agreement [22] (κ = 0.836, z = 20.7, p-value $\ll$ 0.01; and κ = 0.921, z = 25, p-value $\ll$ 0.01).

To answer RQ1, we needed to understand how users expressed their needs through their initial interaction with the system (i.e. the first instance of *U2.1*). To do this we selected all such queries and manually coded them based on three characteristics: *binary politeness*, based on presence or absence of politeness markers (e.g. "please" and use of the subjunctive); *how human-like the utterance is*, from 0 = query-like (e.g. "lactose-free cake recipe") to 2 = as one would communicate with another human (e.g. "I'm looking for a vegan recipe for three people, the cooking time doesn't matter"); and the *level of detail*, in terms of how many information-bearing terms were present, where 0 = 1 piece, 1 = 2 pieces, 2 = 3 or more. These codes were assigned to all *U2.1* queries by both researchers and very high levels of inter-rater agreement were attained - Cohen's Kappa was 1 for politeness and level of detail and was 0.837 (z = 10.1, p-value = 0) for humanness.

To help us answer RQ2; we performed a similar annotation process for user responses to recommendations (i.e. *U3.1*). One annotation was again politeness, which followed the same procedure as described above. Additionally, two researchers coded these responses in terms of *clear sentiment towards the suggestion (positive or negative)* (e.g. "That sounds better"), *Requests for the recipe* (e.g. "…send me the recipe please"), *Further Criteria*, where the need was expanded or refined (e.g. "I don't like strawberries"), *Requests for the ingredients*, which was a request to see the ingredient list; or, finally, *More Information*, which was a request for other information about the recipe, such as the cooking time or steps. As the utterances sometimes contained more than one of these aspects, multiple codes could be applied. Again a very high inter-rater agreement was achieved 0.856 (z = 10.3, p-value = 0). We will make the full transcripts and annotations available with publication.

5 RESULTS

The experiment provided a rich data basis from which to answer our research questions. In sum, 999 participant utterances were collected, 514 of which were from the audio condition. We provide quantitative insights and complementary qualitative examples from the transcripts in three sections: Sections 5.1 and 5.2 provide insight relating to how conversations progressed and language related aspects, respectively, offering evidence relating to RQs 1 and 2. Section 5.3 examines similarities and differences observed between conversations under the spoken and typed conditions, providing insight relating to RQ3. Note that we use the notation $\bar{x}$ to denote the mean of x and $\tilde{x}$ to denote its median.

5.1 Conversation Flow

We analysed how the system and the users navigated around the conversation framework by calculating the Markov transition probabilities between framework nodes (states). Figures 2 and 3 show these transition probabilities in graphical form for the user-to-user state transitions and system-to-system state transitions, respectively. The numbers represent the probabilities of transitioning from one state to another; the thickness of the lines are proportional to the same probabilities. Note that extremely infrequently-occurring

transitions (those with fewer than 5 occurrences over all of the data) have been pruned from this data.
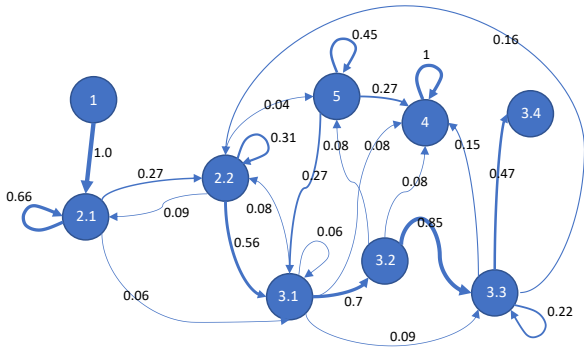
Figure 2: Markov transition diagram between user framework states.
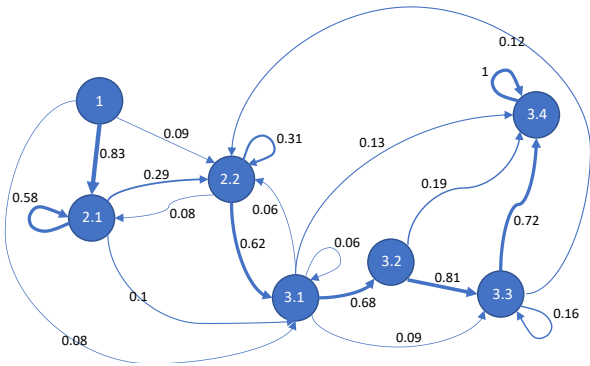
Figure 3: Markov transition diagram between system framework states.

We see that users mostly responded as expected to the Wizard's requests (i.e. a *U2.1* followed an *S2.1*, a *U2.2* followed an *S2.2*, etc) and the conversation flowed through the framework as planned. This can be seen in both the figures by noting the thicker lines between the states. There were 746 possible system-to-user transitions and in 580 cases the user responded as expected (i.e. by following the framework). Note, however, in the user-to-user transition diagram that there are instances where the conversation is forced out of its expected flow by the user's utterance. This is where the user either makes an unexpected request or query (i.e. a *U5*) or produces an expected positive or negative utterance (a *U4*). *U4* utterances were (typically positive) sentiment statements giving feedback, either to a recipe suggestion (e.g. "super, thank you", "I am really hungry now", "good recipe, except for changing the milk and margarine instead of butter") or task completion (e.g. "thank you, Telefood"). Although many of these appear similar to the *U3.1* utterances, the difference is that they came from the user at a point in the conversation where the system did not explicitly expect such a comment.

Utterances labelled with *U5* were typically questions relating to either the recipe, e.g. "how long does it take?" or ingredients (typically preceded by *2.2*) e.g. "how many calories does Quinoa

State	# of times visited								
	1	2	3	4	5	6	7	8	9
2.1	12	17	19	16	7	2	5	3	1
2.2	32	24	9	6	2	0	0	0	0
3.1	56	21	3	1	0	0	0	0	0

Table 1: Number of times each state in the conversational frame work was visited.

have and what is its nutritional profile?". Other examples included questions relating to potential changes in the recipe, e.g. "can I leave out the peas?". While our planned setup did not explicitly account for such questions, the Wizard provided appropriate information as one could imagine a system being able to satisfy these in practice, although this would require that the system be able to accurately determine when an utterance contains such a query. Much more challenging to address would be other examples of *U5* that contain more complex requests, such as "can you tell me a place or supermarket where I can obtain these ingredients?". In this case the Wizard responded with one of the standard error messages.

In only 6 instances did a conversation proceed directly from the welcome greeting (*S1*) to a user giving a single initial response (*UI2.1*) to the Wizard giving a recommendation (i.e. *1 -> 2.1 -> 3.1*) and these were all under the text condition. Table 1 outlines the number of times that each framework state was visited. Note that there was considerable variation in the number of times that the initial information-gathering states (i.e. *U2.1* and *S2.1*) had to be visited, while the latter ingredient gathering and recommendation states were generally visited less frequently. For example, out of a total of 81 successfully completed tasks, in 32 cases users only needed to be prompted to provide ingredient preferences once but, in 2 cases, state *U2.2* was visited 5 times. In 9 instances users never needed to be prompted to give ingredient preferences at all, as they had already provided this information in state *U2.1*. In most cases (56 or 69.1%), the conversation only visited the recommendation and user response states (i.e. *S3.1* and *U3.1*) once, meaning that the first recommendation was accepted. This suggests either that the wizard was able to interpret user needs well and provide high-quality suggestions or users were simply easy to satisfy.

If we consider the average number of times states are revisited on a per-task basis, we find that for the open-ended task 1, *U2.1* is on average visited 3.8 times, while, for the much more clearly-defined task 2, the same state is only visited 1.4 times on average. We observe a similar pattern for the state in which ingredient preferences are sought (*U2.2*), which is visited twice on average for task #1 but only 1.25 times on average in task 2. Regarding the recommendation state (*3.1*) we observe very little difference between the average visits by task, varying from 1.56 for task 3 to 1.18 times for task 6.

The fact that only few conversations proceeded directly from stage 2.1 to a recommendation suggests that the teleporting behaviour reported by Trippas et al. [36] was rare. To explore this further we analyse our annotations relating to level of detail in the initial queries (i.e. *U2.1* preceded immediately by *U1*). 11 out of 82 of these initial utterances were labelled as teleporting, where participants tried to describe the full (or close to full) task criteria in a single utterance. Good examples include:

"I would like to cook for two, that is for me and two of my friends. I do not have any time pressure, but the recipe should be vegan" (user 18)

"I would like a recipe that I can cook in a maximum of 10 minutes and that doesn't have any further restrictions. It can have meat and dairy, it can also have gluten in" (user 26).

This second example is interesting because the user specifically emphasises the lack of restrictions on the recipe. An utterance with this kind of formulation would be difficult for a system to reliably deal with in practice and would likely lead to recipe suggestions with exactly those restrictions the participant did not want.

Such teleporting examples contrast with 30 examples that featured only a single criteria, e.g. "Name me a vegan recipe" (user 5) or "I would like to cook an evening meal" (user 20). The majority (just over half) of initial user inputs, however, contained 2 criteria which, in the main, did not constitute enough information to allow a suitable recommendation and prompted further questioning by the Wizard. Only a single participant used a very detailed query for all three assigned tasks and a further 4 participants never supplied an initial query with only a single piece of information. One participant provided an initial query with 2 pieces of information for all three tasks. However, in the main, the impression gained is that the task explained more variance in terms of query detail than the participant involved.

Examining how the level of detail of initial input varied over tasks reveals differences. In more specific tasks (e.g. task 2), users tended to mention criteria in their initial need description utterance (i.e. these descriptions were more likely to be annotated as detailed, mean level of detail was 1.06 for task 2, compared to 0.73 for task 1). This means when criteria were not explicitly mentioned in the task description, participants did not tend to add their own personal criteria, at least not in the initial descriptions.

5.2 Language Use

Given that previous work has revealed conversational search interactions to differ from search queries in traditional search systems in terms of the language used, it made sense to analyse this aspect here too. In addition to basic properties that have been examined previously, such as length [9, 17], which have been shown to differ in conversational settings, we analyse other aspects associated with human-to-human speech, including politeness and aspects such as sarcasm to establish if previous observations of more human-like interaction hold in our specific setting.

5.2.1 Length of utterances. When considering all utterances, the average length in terms is low ($\bar{x} = 3.819$, $\tilde{x}=2$ terms), but the distribution is very skewed (max =50). This varies to a large extent over user states with error states U5 ($\bar{x}= 6.93$ and $U2.1$ ($\bar{x} = 5.06$) being associated with the longest utterances and $U3.2$ ($\bar{x}=1.72$) and $U3.4$ ($\bar{x}=2.00$) the shortest. This makes sense given the open and free-form nature of the responses for $U2.1$ and $U5$.

Length of utterance also varies over task and it is particularly interesting to examine the lengths over task for specific user states, e.g. initial queries (U2.1 following S1) and ingredient specification U2.2. The initial queries $\bar{x} = 8.6$ $\tilde{x} = 7$ are significantly longer than later U2.1s $\bar{x} = 3.6$ $\tilde{x} = 3$ (W= 2395, $p \ll 0.01$). Initial queries vary

over task with the clearly-defined task 2 having the lowest average ($\tilde{x} = 6$), whereas task 1, which is much more open in nature, had much longer queries ($\tilde{x} = 8.5$). We see the same pattern with the ingredient definitions (U2.2), where task 2 has a median length of 1.5 terms, whereas for task 1 the median was 4.

Next we present analyses relating to two states of particular interest: Initial user need descriptions (*U2.1*) and responses to recommendations (*U3.1*).

5.2.2 User need descriptions - U2.1. :

From the literature we know that dialogue with conversational search interfaces tends to be more "human-like". We explore this in the cooking context, specifically looking at the participants descriptions of their needs.

10 of the initial need descriptions provided by users were labelled as being designed for machine processing. Examples here included:"food, quick, easy" (user 17) and "cake without milk" (user 8). At the other end of the spectrum, 16 of these utterances were labelled as designed for humans e.g. "I would like to go jogging in an hour and eat something beforehand, what would you recommend?"(user 23). The majority (55), however, were somewhere in between these extremes, e.g. "I need a recipe quick before jogging" (user 18) and "I want something to eat quickly" (user 10). These do supplement queries with machine superfluous terms such as "I would like" or "I want" but would be easier for a system to process than the human-like examples listed above.

One feature of human speech is the use of politeness markers when requesting something. We marked 14 of the 82 initial user queries as having such markers. The utterances do not reflect how we would expect participants to behave if talking to a human in person. Interestingly, the examples came from 14 unique participants (i.e. they only ever did this once), which could mean that they did not perceive benefit and decided not to re-use the approach.

Unsurprisingly there are no examples of initial queries that were coded as designed for machine processing and also polite. However, there are several examples queries labelled as human-like that did not contain any politeness markers e.g. ("Find me a recipe that can be cooked in 5 minutes" (user 22)).

5.2.3 Response to recommendations - U3.1s. One key task for a conversational recommendation system is to gauge user reaction to recommendations based on the utterances that follow. 71 out of 113 (62.8%) of these utterances in our dataset were labelled as containing clear, positive sentiment, following which the wizard moved on to the next stage in the framework. Examples were often allow the lines of "Very good, thank you" (many users), "sounds perfect" (user 11) or "I want to cook that". In such cases determining user impression would be trivial. Negative responses seldom occurred. Indeed, only 1 *U3.1* utterance was labelled clearly negative. When the sentiment was not clear, we applied further codes, the frequencies of which are distributed as in 2. These include utterances containing feedback to the recommendation, which should be accounted for e.g. "Sounds good, can you look for recipes without milk?" (user 20) or "I don't have any spinach there" (user 11); requests to see the recipe e.g. "okay, show me the recipe please" (user 20) or the ingredient list or more information about the recipe, such as the cooking time.

Code	Count
Feedback	17
Recipe	14
Further Criteria	9
Ingredients	8
More Information	8

Table 2: Codes from analyses of recommendation responses where response was not merely indication of sentiment.

Examples that would be particularly challenging for a system to deal combined many of these aspects e.g. "that sounds very good, can you send me the ingredients and the processing time of the recipe?" (User 26) and "that doesn't sound bad, it can also be something more complex, I have about 3.5 hours time, please find out the second recipe" (user 15). Not only is this utterance of note for the multiple clauses, some of which would be difficult for a system to interpret, but it was one of the few navigate actions performed by our participants where there was an effort to compare or navigate between recommendations offered by the system.

As in the previous section, we examined the *3.1s* for politeness markers. Similarly to the initial queries, only a small number of responses to recommendations featured politeness markers (12 / 113 - 10.6%) and these came from 9 unique users. Taking the initial queries and responses together shows that just over half of participants (16) used politeness markers at all.

To conclude this section on language use, we wish to report two further observations made with respect to past literature. Unlike past studies, we find no sign of the short-positive utterances to offer feedback to the system or to proceed to the next step of interaction [14, 38]. This may be because of the restricted context of recommendation only or it may be a consequence of the structured dialogue framework bounding conversations.

One phenomenon not reported in the literature, that we did, however, observe in our study was the use of sarcasm. There were 6 cases where users included sarcasm in their input. Good examples include the response to the question, "which ingredients should be avoided?", User 5 replied "nothing poisonous". In response to the recommendation of "Caviar and Egg on Toast" and when requested for feedback, User 15 replied "Sadly, I am all out of caviar". Sarcasm would, of course, be challenging for any agent to deal with. Surprisingly, all of the examples of sarcasm were provided by participants in the text condition. In the following section we examine the differences between the two conditions in more detail.

5.3 Typed vs Spoken

To attain an understanding of the extent to which the conversation flow varies (deviates from the expected path) across conditions, we calculated the entropy over all of the Markovian transition probabilities, smoothing by addition of a small amount of extra probability to all zero probabilities ($\alpha = 0.000001$). For the user-to-user transitions (i.e. those in Figure 2) in the text condition, the entropy was 0.625, while for the audio condition, the entropy was 0.573. The entropies for the system-to-system probabilities (Figure 3) follow a similar pattern - the entropy ($H_{text} = 0.705$) is higher for the text condition than in the audio condition ($H_{audio} = $

0.642). This suggests that the audio condition caused users to more consistently follow a single, dominant conversational flow, while participants in the text condition were more likely to deviate from this.

In terms of utterance length, for all utterances across the two conditions there is no significant difference ($\bar{x}_{audio} = 3.92, \bar{x}_{typed} = 3.76\ \tilde{x}_{audio} = 2, \tilde{x}_{typed} = 2)W = 87,815, p = 0.14$). However, examining only utterances relating to ingredient preferences (*U2.2*) under the two condition reveals a significant difference in length with the typed input being longer ($\bar{x}_{audio} = 3.27, \bar{x}_{typed} = 3.82\ \tilde{x}_{audio} = 1, \tilde{x}_{typed} = 3, W = 3,013, p = 0.024$).

To identify words that are likely to appear in spoken (Audio) but not typed (Text) utterances and vice-versa we apply the log odds ratio [3]. Figure 4 shows the most distinctive words for these cases. Tokens were derived by splitting on whitespace and punctuation and then stemming using the German Snowball Stemming approach as implemented in the Corpus package for R[4].
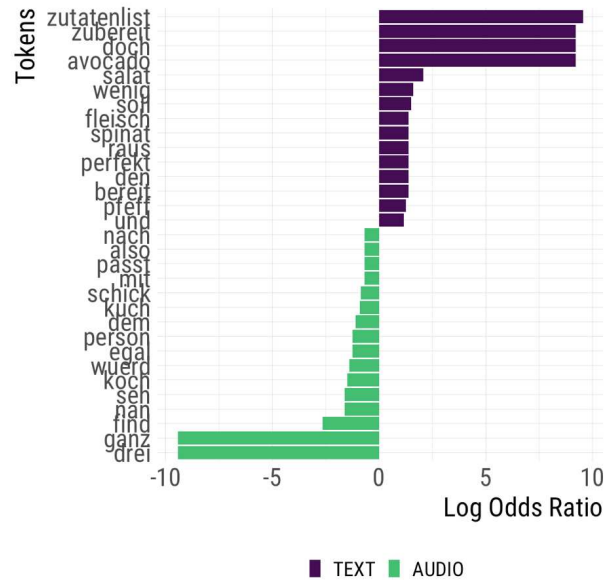
Figure 4: Most discriminative terms across conditions as determined by the log-odds-ratio.

Examining the output, it seems that several ingredients feature in set of words more likely to be typed. This set included the German words for salad, meat, spinach, pepper and avocado, as well as the word ingredient-list itself. On the other hand, no individual ingredient made the top set of terms for audio utterances. One might have expected the audio utterances to contain more stopwords given the previous findings of more human-like language for this condition in the literature. This was not really the case, however. The German words for "with" and "after" were in the audio set, but the word for 'ànd' featured in the typed set.

---

[3] To remove noise we only include terms which appear at least 5 times in the dataset.
[4] https://cran.r-project.org/web/packages/corpus/vignettes/stemmer.html

	2.1	2.2	3.1	3.2	5	All
Text	5	3	3	1	6	*18*
Audio	3	1	1	0	3	*8*
Both	*8*	*4*	*4*	*1*	*9*	

Table 3: Counts of errors origination from different conversational framework states.

We note that the most discriminative audio words included terms typically associated with politeness. Use of "[ich] würde ([I] would)" being a good example. The words also hint that typed utterances were more polite and request oriented ("sollte (should)"), for example:

> [I would like [polite] to know what I should eat before I go jogging] (user 12). 'Ìch **würde** gerne wissen was ich vor dem Joggen essen **soll**"

While these examples suggest greater levels of politeness in the audio condition, the manual annotations for the initial queries (U2.1s following S1) do not support this. We found no significant differences between the categories ($p = 0.27$)[5]. The manual annotations do, however, reveal audio initial queries to be significantly more human-like than typed utterances in this state ($p = 0.047$).

System error states, of which there were 26 in total, were more likely to be incurred in the text condition than when participants interacted via spoken utterances (18 vs 8). As Table 3 shows, errors were most common in the context in which general recipe context is sought (2.1) and, unsurprisingly, after the user has made an out-of-framework request (i.e. *U5*).

6 DISCUSSION

In the previous section we have reported the results of several analyses, which complement those reported in the literature and provide insights into how the wizard in our setup may be automated. This section highlights the key findings and provides interpretation with respect to the literature and the feasibility and challenges of realising a conversational recipe recommendation agent.

Many elements of our results are promising for the near-term future of human-machine interactions in this context (and beyond): many conversations followed the expected flow through the framework as anticipated and participants responded to queries and suggestions made by the Wizard with clear utterances that an automated system should be able to interpret. These kinds of interactions could feasibly be handled by existing technologies via the "slot filling" approach [41], provided a sensible conversational framework had been developed. Similarly encouraging is the fact that, in most cases, utterances in response to recommendations mostly contained clear expressions of sentiment, which could be used as either explicit or strong implicit ratings. However, there were also cases in which the user responded in an unexpected manner, likely presenting a machine with a much more difficult utterance to which to helpfully and "intelligently" respond.

There was considerable variation in conversational styles and in individual utterances between users, however, in contrast to much of the literature, we found surprisingly few clear differences between the spoken and typed conditions. Although spoken utterances were judged to be more "human-like", they certainly were not longer. We did observe instances of the "teleporting" behaviour described by Trippas et al. [36], particularly as initial queries to the system (i.e. *U2.1*, which contrasts with the idea of people interacting with systems in a human-like manner [38]). The fact that many of these teleporting queries were provided by users under the text condition suggests that the communication medium does have some impact on how people interact - it *feels* more like a traditional search system so people are more likely to interact as if it is.

Our data revealed that conversations did not feature the short positive utterances or „"grounding behaviour" reported in [38] and only featured a subset of the action types listed by [2]. This suggests that spoken dialogue systems will have to overcome the same obstacles as many other advanced interface features in getting users to make use of more advanced features and to explore the system's capabilities [39].

We found instances under both conditions of language use, such as employing politeness markers and even sarcasm, that does not serve to provide information and take the conversation forwards towards a positive conclusion. These suggest that, although the users know (or in this case think) they are interacting with a computer, which does not have feelings or sensibilities that can be offended, they still employ these non-information-bearing conversational elements. Some participants even employed sarcasm in their dialogues, which are very human-like and are particularly difficult for current machine learning approaches to deal with [20]. This ties into Luger and Sellen's work [23], suggesting that our participants' expectations of what a machine intelligence can interpret and respond to is out of step with their true capabilities.

7 CONCLUSIONS & FUTURE WORK

Our study is the first to investigate conversational agents in the context of recipe recommendation by means of a WoZ study. Our findings show that conversational interaction with a recipe recommender can be diverse – varying considerably by task and user and, to some extent, by interaction mode. There were many instances of human-like dialogue but this was not consistently the case. Many conversations followed the expected flow of our framework and, as such, the results provide us with optimism that the majority of the human-controlled behaviour of the Wizard could be automated by tracking the state of conversations through the framework.

Technological challenges exist, including the disentangling of complex teleporting queries, dealing with out of context responses and interpreting unclear sentiment responses to recommendations. These challenges represent our future work. We are currently using the collected data to develop and evaluate bots, which could replace the Wizard in future rounds of data collection and whose "performance" could be evaluated with respect to the data presented in this paper.

---

[5]Differences between manually applied codes under text and audio conditions were tested using Fisher's exact tests.

REFERENCES

[1] Sandeep Avula. 2018. Wizard of Oz: Protocols and Challenges in Studying Searchbots to Support Collaborative Search. In *SIGIR 2nd International Workshop on Conversational Approaches to Information Retrieval (CAIR'18)*.

[2] Leif Azzopardi, Mateusz Dubiel, Martin Halvey, and Jeffery Dalton. 2018. Conceptualizing agent-human interactions during the conversational search process. In *The Second International Workshop on Conversational Approaches to Information Retrieval*.

[3] Nicholas J. Belkin, Helen M Brooks, and Penny J. Daniels. 1987. Knowledge elicitation using discourse analysis. *International Journal of Man-Machine Studies* 27, 2 (1987), 127–144.

[4] Nicholas J. Belkin, Colleen Cool, Diane Kelly, S-J Lin, SY Park, J Perez-Carballo, and C Sikora. 2001. Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval. *Information Processing & Management* 37, 3 (2001), 403–434.

[5] Petter Bae Brandtzaeg and Asbjørn Følstad. 2017. Why people use chatbots. In *International Conference on Internet Science*. Springer, 377–392.

[6] Katriina Byström and Kalervo Järvelin. 1995. Task complexity affects information seeking and use. *Information processing & management* 31, 2 (1995), 191–213.

[7] Giuseppe Carenini, Jocelyin Smith, and David Poole. 2003. Towards more conversational and collaborative recommender systems. In *Proceedings of the 8th international conference on Intelligent user interfaces*. ACM, 12–18.

[8] Sally Jo Cunningham and David Bainbridge. 2013. An Analysis of Cooking Queries: Implications for Supporting Leisure Cooking Ethnographic Studies of Cooks and Cooking. In *iConference 2013 Proceedings*. 112–123. https://doi.org/10.9776/13160

[9] Mateusz Dubiel, Martin Halvey, Leif Azzopardi, and Sylvain Daronnat. 2018. Investigating how conversational search agents affect user's behaviour, performance and search experience. In *The Second International Workshop on Conversational Approaches to Information Retrieval*.

[10] David Elsweiler, Morgan Harvey, Bernd Ludwig, and Alan Said. 2015. Bringing the" healthy" into Food Recommenders.. In *DMRS*. 33–36.

[11] David Elsweiler, Christoph Trattner, and Morgan Harvey. 2017. Exploiting food choice biases for healthier recipe recommendation. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*. ACM, 575–584.

[12] Norman M Fraser and G Nigel Gilbert. 1991. Simulating speech systems. *Computer Speech & Language* 5, 1 (1991), 81–99.

[13] Jill Freyne and Shlomo Berkovsky. 2010. Intelligent food planning: personalized recipe recommendation. In *Proceedings of the 15th international conference on Intelligent user interfaces*. ACM, 321–324.

[14] Alexander Frummet, David Elsweiler, and Bernd Ludwig. 2019. Detecting domain-specific information needs in conversational search dialogues. In *Proceedings of the 3rd Workshop on Natural Language for Artificial Intelligence*.

[15] Souvick Ghosh. 2019. Informing the Design of Conversational IR Systems: Framework and Result Presentation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1454–1454.

[16] David Graus, Paul N Bennett, Ryen W White, and Eric Horvitz. 2016. Analyzing and predicting task reminders. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*. ACM, 7–15.

[17] Ido Guy. 2016. Searching by talking: Analysis of voice queries on mobile web search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 35–44.

[18] Morgan Harvey, Bernd Ludwig, and David Elsweiler. 2013. You are what you eat: Learning user tastes for rating prediction. In *International Symposium on String Processing and Information Retrieval*. Springer, 153–164.

[19] Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. 2015. Automatic online evaluation of intelligent assistants. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 506–516.

[20] Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)* 50, 5 (2017), 73.

[21] Carol C Kuhlthau. 1991. Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science* 42, 5 (1991), 361–371.

[22] J Richard Landis and Gary G Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* (1977), 363–374.

[23] Ewa Luger and Abigail Sellen. 2016. Like having a really bad PA: the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5286–5297.

[24] Department of Health. 2009. Change4Life marketing strategy: In support of healthy weight, healthy lives.

[25] Sheryl Ong and Aaron Suplizio. 2016. Unpacking the breakout success of the amazon echo. *Retrieved December* 12 (2016), 2018.

[26] AM Pejtersen. 1979. Investigation of search strategies in fiction based on an analysis of 134 user-librarian conversations, Third International Research Forum in Information Science, T. *Henriksen, ed., Oslo* (1979).

[27] Paul Prekop. 2002. A qualitative study of collaborative information seeking. *Journal of Documentation* 58, 5 (2002), 533–547. https://doi.org/10.1108/00220410210441000 arXiv:https://doi.org/10.1108/00220410210441000

[28] Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*. ACM, 117–126.

[29] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2011. Introduction to recommender systems handbook. In *Recommender systems handbook*. Springer, 1–35.

[30] Sosuke Shiga, Hideo Joho, Roi Blanco, Johanne R Trippas, and Mark Sanderson. 2017. Modelling information needs in collaborative search conversations. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 715–724.

[31] Robert S. Taylor. [n. d.]. The process of asking questions. *American Documentation* 13, 4 ([n. d.]), 391–396. https://doi.org/10.1002/asi.5090130405 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.5090130405

[32] Jaime Teevan, Christine Alvarado, Mark S Ackerman, and David R Karger. 2004. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 415–422.

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

[34] Christoph Trattner and David Elsweiler. 2017. Food recommender systems: important contributions, challenges and future research directions. *arXiv preprint arXiv:1711.02760* (2017).

[35] Christoph Trattner, David Elsweiler, and Simon Howard. 2017. Estimating the healthiness of internet recipes: a cross-sectional study. *Frontiers in public health* 5 (2017), 16.

[36] Johanne R Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the design of spoken conversational search: perspective paper. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. ACM, 32–41.

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

[38] Alexandra Vtyurina and Adam Fourney. 2018. Exploring the role of conversational cues in guided task support with virtual assistants. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 208.

[39] Ryen W White and Dan Morris. 2007. Investigating the querying and browsing behavior of advanced search engine users. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 255–262.

[40] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3506–3510.

[41] Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proc. IEEE* 101, 5 (2013), 1160–1179.

[42] Xiaojun Yuan and Nicholas J. Belkin. 2014. Applying an information-seeking dialogue model in an interactive information retrieval system. *Journal of Documentation* 70, 5 (2014), 829–855.

[43] 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*. ACM, 177–186.