AALBORG UNIVERSITY
COPENHAGEN

Department of Communication & Psychology
Information Studies Programme
Aalborg University, A.C. Meyers Vænge 15
2450 Copenhagen SV, Denmark

Engagement and Usability of Conversational Search – a Study of a Medical Resource Center Chatbot

Author:
Tamás Fergencs

Semester:
Spring 2020

Supervisor:
Dr. Florian Meier

Length: 134 019 characters with spaces / 56 pages

Table of contents

List of figures

List of tables

Engagement and Usability of Conversational Search – a Study of a Medical Resource Center Chatbot

Tamás Fergencs
tferge18@student.aau.dk
Aalborg University, Copenhagen, Denmark

Abstract. Due to advances in natural language understanding, chatbots have become popular for assisting users in various tasks, for example, searching. Chatbots allow natural-language queries, which can be useful in case of complex information needs, and they provide a higher level of interactivity by displaying information in a dialog-like format. However, chatbots are often only used as auxiliaries for a graphical search user interface. Thus, they must be engaging and usable so that users both *want to* and *able to* use them. In this study, a chatbot-based and a website-based search interface were compared in terms of engagement and usability. Engagement was measured using the User Engagement Scale; think-aloud protocol and a questionnaire were used to assess usability. Behavioral measures were used to triangulate data. Findings indicate that the usage of the chatbot did not lead to a higher level of engagement, moreover, its usability was lower compared to the website-based search interface.

Keywords. Conversational search, search user interface, user engagement, chatbot

1 Introduction

Conversational interfaces are becoming increasingly popular due to the advancement in natural language understanding technology. They enable human-computer interaction via natural speech or text instead of using buttons and menus. Text-based conversational interfaces, the so-called chatbots, have been around for quite some time, but they gained commercial interest only recently, due to digital communication becoming a standard (Dale, 2016). Customer service chatbots are proliferating as businesses explore the possibilities of conversational commerce to interact with and provide support for consumers (Chung et al., 2018; Exalto et al., 2018; Zhu et al., 2018). Mobile health solutions are starting to utilize conversational agents to promote health or facilitate recovery (Denecke et al., 2018; Gratzer & Goldbloom, 2019; Perski et al., 2019). Chatbots also find their way into the field of education, where they aid university students to learn about school facilities or act as teaching agents to supplement classroom learning (Graesser, Li, & Forsyth, 2014; Reed & Meiselwitz, 2011).

Regardless of the field of application, conversational agents are seen as a useful tool to facilitate user engagement – they can motivate increased usage of an application, enrich business-to-consumer interactions, or simply serve as "wow factor" for marketing purposes. One specific use case of chatbots is assisting with searching and retrieval of web content – a concept denoted as conversational search (Radlinski & Craswell, 2017). Instead of scrutinizing a lengthy FAQ page, a user can simply submit their question to a chatbot, which queries the database and returns a relevant answer (Lee et al., 2019). Or, a library chatbot can help in promptly retrieving reading material based on the user's preferences (Allison, 2012; Ward, 2005). Thanks to natural language understanding, users can submit complex search queries, which can help in cases where the information need is difficult to formulate. This can be especially useful for non-targeted searching, where exploration of the collection is the main activity (Vakulenko, Markov, & de Rijke, 2017).

However, a chatbot is often used as an auxiliary to a website search interface, and not as a standalone search system. If the chatbot is not engaging enough, the initial interest can quickly fade, and users will return to using the website search. Chatbots can increase user engagement by enhancing interactivity, that is, by delivering information in a dialog-like manner (Sundar et al., 2016). However, it is uncertain whether a higher level of interactivity is enough for users to prefer using the chatbot if there is an alternative. Besides, implementing search functionalities to a conversational interface is not a straightforward process and, even if it's successful, users may have trouble transitioning from a traditional graphical search user interface to a conversational interface (White, 2018). This is due to the inherently complex nature of search behaviors, which generally do not adhere to a simple query-answer model, but are rather characterized by constantly evolving information needs (Bates, 1989). A search chatbot, therefore, should satisfy both the need of enhancing user engagement and serve as a user-friendly supplement (or even substitute) of a graphical search user interface. If the chatbot has poor usability, people may not be *able* to use it. If the chatbot does not motivate engagement, people may not *want* to use it.

The thesis aims to compare the conversational search user interface of Lundbeck's medical resource center with its graphical search user interface in terms of user engagement and usability. An experiment approach was used, where users completed various information retrieval tasks, throughout which interaction measures and self-report data was collected via the User Engagement Scale and an exit questionnaire. Quantitative data were statistically analyzed, and a thematic analysis was conducted on the qualitative data to see if there are any differences between the two interfaces in terms of engagement and usability.

Section 2 details the main concepts the thesis touches upon: conversational interfaces, user engagement and usability. Section 3 describes the Progress in Mind platform – both the website and the chatbot search interface. After reviewing the current state of the field in section 4, section 5, 6, and 7 describe experimental setup, the analysis of the data and the main results. Section 8 contains the interpretation of the findings and the limitations of the study, and section 9 summarizes the major conclusions.

1.1 Problem formulation

Lundbeck's chatbot is an experiment of trying to engage users by introducing conversational modality into the search system, but the company requires data to see whether the chatbot is successful in this term or not. Therefore, the encompassing research question for the thesis is as follows:

***RQ**: How does a conversational search interface of Progress in Mind compare to a graphical search user interface in terms of user engagement?*

It is hypothesized that the chatbot will achieve its goal, i.e. it will successfully enhance user engagement, in accordance with Lundbeck's intentions. Therefore, the pertaining hypothesis and the null hypothesis are:

H0: The usage of the chatbot for searching has no effect on user engagement.

H1: The usage of the chatbot for searching has a positive effect on user engagement.

While the main focus of the study will be comparing the overall engagement of users across the two interfaces, one aspect of user engagement will be discussed in greater detail: system usability. According to the user engagement model (Figure 2), usability is important in sustained user engagement, as poor usability can easily lead to disengagement – something which Lundbeck wants to avoid. Therefore, the following research sub-question can be added:

***RQ-s**: How does a conversational search interface of Progress in Mind compare to a graphical search user interface in terms of usability?*

Usability will be broken down into the constituents of effectiveness, efficiency and satisfaction (according to International Organization for Standardization (2018)) which will be measured separately.

2 Background

Lundbeck is an international pharmaceutical company established in Denmark, in 1915. The company engages in continuous research on brain diseases like depression, schizophrenia, Parkinson's disease, bipolar disorder, and Alzheimer's disease, and manufactures drugs for treating such disorders. In 1997 the company founded The Lundbeck Institute – an educational forum that provides medical education and seminars on psychiatric and neurological disorders. The institute recently created an online open-access database called "Progress in Mind", where articles and videos about current scientific trends, international news, and congress highlights are hosted. The publications on the website are written and curated by medical writers in a generally informal style, and the content is aimed at healthcare practitioners and academics in the field. Users can filter content by diseases or types of publications or use free-text queries to search across the database. The Progress in Mind resource center is accessible at https://progress.im/en.

Lundbeck Institute's goal is to transform the platform into the go-to resource center for healthcare professionals in psychology and neurology. Therefore, they are experimenting with new ways to make the content more accessible and interactive – which led to the development of a chatbot. This chatbot interface is an auxiliary tool for the website search and uses a conversational modality to help users search the database, presenting search results in a chat window. This conversational style is aimed to improve interactivity, which, as Lundbeck anticipates, will lead to greater user engagement and promotes the usage of the platform. As Sundar et al. (2016) have shown, delivering online content in a dialog-like manner can lead to improved interactivity and, in turn, a greater level of engagement. However, recent research points towards the dilemma of utility, as conversational interfaces are still not developed enough to satisfy complex user needs (which is especially relevant in search tasks) and thus their usage often leads to disappointment (Luger & Sellen, 2016). Research on how to facilitate user engagement using a conversational interface is still relatively scarce, which is the main motivation behind this thesis.

2.1 Conversational interfaces

2.1.1 Technology of conversational interfaces

Though often intertwined with science fiction, the notion of using human language to interact with inert machines has been of interest in the human-computer interaction field for a long time. Human-human conversational patterns are such fundamental constituents of our lives that sometimes even IT experts anthropomorphize the way they communicate with computers (Cassell, 2000). Besides the form of communication, humans possess personal skills, experiences, and perceptions, which they bring into play during their interaction with computers (Norman & Thomas, 1990). The motivation for bringing a conversational modality into human-computer interactions, therefore, stems from the observation that people tend to make sense of the behavior of computer systems through anthropomorphic interpretation. A great example of designing for the anthropomorphic sensemaking process of users is the usage of metaphors in a graphical user interface (Flach, 2011). Microsoft used for the design of their operating system a visual language that is analogous to an office filing system with accessories (wastebin, folders, documents etc.) to mirror a real-life context, thus making the graphical system less abstract and its functions easier to interpret. Conversational interfaces can enable the same kind of natural communication with the computer.

Natural-language understanding

Although natural language processing (NLP) has been the subject of study since the 1950s, the late 1960s and 1970s brought about a surge of technological advancement in the field (Jurafsky & Martin, 2000b). One of the first notable systems that was capable of processing and replying to natural-language inputs was ELIZA (Weizenbaum, 1966), a computer program that simulated a psychologist and used rudimentary logic to carry on a conversation. The software operated by substituting certain parts of the input text and returning it in the form of a question or reassurance, essentially mirroring the inputs of the participant – a conversational technique fundamental to Rogerian therapy (Jurafsky & Martin, 2000d). Though the system was not capable of producing sophisticated sentences, users of ELIZA generally perceived it to be a real psychologist and thus the program claimed public success (Bassett, 2018). ELIZA served as a convincing demonstration that attributing natural-language communication skills to machines does enable personification and embodiment and that it can elicit anthropomorphic impulses from people.

Another early attempt to make a computer understand natural-language commands is the SHRDLU interface developed by Winograd (1972). Users of SHRDLU were able to move digital three-dimensional objects by submitting commands to the system in English (e.g. "Pick up a big red block."). This project paved new ways for the discipline of *natural language understanding* – a field that is concerned with specifying the "theory of language comprehension and production to such a level of detail that a person could write a computer program that can understand and produce natural language" (Allen, 1987).

Around the same time the first large-scale, natural-language data management system was developed by Woods (1973), named LUNAR (short for The Lunar Sciences Natural Language Information System). The software was used to retrieve information about the moon rock samples brought back from the Apollo 11 mission. Users would submit natural English commands, e.g. "Which samples contain chromite?" and the system would return the appropriate answer or list of results.

Though rudimentary in logic, these first *natural-language interfaces* (NLI) count as significant achievements for their time. Their main drawback was the difficulty of generalization – while these interfaces were able to serve particular use cases, adapting them to other fields proved to be burdensome (Liang, 2014). The difficulty lies in the intricacies of the natural language, which often baffles the computer. To reduce linguistic complexity, NLIs, especially the earlier iterations, are not trained to understand the entirety of a language. Instead, they use a specific subset of the language to interpret user queries. This "habitable" language (Watt, 1968) serves as the vocabulary of the system.

This vocabulary is defined to the system as *formal language*. A formal language is a set of strings generated from an alphabet (composed of characters or symbols) and their formation rules (the logic of how the characters or symbols are arranged to form the strings) (Jurafsky & Martin, 2000d). By defining the rules of formation, the system can generate all instances of strings which it accepts, without the need to manually define each instance. This is called *generative grammar* that enables the system to model the natural language by generating strings that adhere to grammatical rules.

This enables the software program to analyze an input word and produce a certain structure of it, e.g. if the system receives the word "going", it can break that word down into verb root (go) + gerund (-ing). This process is called *morphological parsing* (Jurafsky & Martin, 2000c). When words are strung together into sentences, the software must also understand the structure of the sentence and how the separate words contribute to the overall meaning of it. So, the system must conduct *syntactic analysis* or *parsing*, through which constituents of the sentences and their syntactic relation are analyzed (Rich, 1984). The last step to understand human

language is attributing meaning to user statements by linking the parsed linguistic elements to the "non-linguistic knowledge of the world" via *semantic analysis* (Jurafsky & Martin, 2000e). For example, if the user says "List all employees of Lundbeck subsidiaries with more than 1 million DKK in yearly revenue." could either be interpreted as:

1. *List all [employees [of Lundbeck subsidiaries] [with more than 1 million DKK in yearly revenue]].* That is, list all employees who have more than 1 million DKK in yearly revenue and work at a Lundbeck subsidiary.
2. *List all [employees [of Lundbeck subsidiaries [with more than 1 million DKK in yearly revenue]]].* That is, list all employees who work at a Lundbeck subsidiary which has more than 1 million DKK in yearly revenue.

A human would unmistakably understand that "with more than 1 million DKK in yearly revenue" refers to "Lundbeck subsidiaries", but a computer might associate it with "employees". Humans do not naturally associate revenue with individual people, but rather with corporations, but this external knowledge is not apparent for software systems. The computer must be taught to pick the right semantic interpretation via semantic analysis of the sentences.

The goal of analyzing the input sentence is to get the information needed to decide on the response or follow-up action. The system needs to, first, recognize the *domain* of the user's goal (e.g. booking flight tickets, table reservation in a restaurant, etc.), then the *intent* of the user, that is, what kind of goal do they want to achieve (e.g. transferring flight tickets to someone, canceling a table reservation). In the case of a multi-domain chatbot, dividing to domain and intention might be useful, but in a more specific context, these two might be merged to just identifying the user intent (McTear, Callejas, & Griol, 2016c). To further specify the need of the user, the system must extract relevant information from the utterance called *slots* or *entities*. For example, in the case of booking flight tickets, the relevant entities can be the departure date, place of departure, the number of passengers, etc. Conversational systems often prompt the user to provide the entities needed to decide on a response or complete an action – a process known as *slot filling* (Jurafsky & Martin, 2000e). Though this approach usually leads to a fairly linear question-response interaction, sophisticated conversational interfaces should be able to handle more lifelike discourses.

Dialog management

When we talk to computers, communication does not consist of isolated sentences uttered one after another. They all form a stream of connected thoughts that form a discourse (Jurafsky & Martin, 2000a). During a conversation, we omit pieces of information from our utterances for conciseness, but it causes little trouble for humans to fill in these gaps of information via the context. A particular example is the *reference resolution*, when we denote an already introduced object ("I saw *a blue car* on the street") by e.g. a pronoun ("I saw *it* too").

Although a satisfactory morphological, syntactic, and semantic parsing is essential for NLIs to imitate basic human communication, they are not adequate for imitating a conversation. For a system to recognize the structure of the conversation and identify how the language is used to refer to things, it must have the ability for *pragmatic parsing*. Sophisticated conversational agents specialized for executing complex tasks (e.g. ordering a plane ticket or manipulating a data repository) need to be able to interpret the entirety of the conversation to be efficient. This ability is called dialog management (McTear, Callejas, & Griol, 2016b) and consists of four main modules according to Traum & Larsson (2003):

- Updating the dialog context based on the state of communication;

- Provide the context for interpretation;
- Coordinating with other modules (e.g. database manager, calendar editor, etc.);
- Deciding what information to convey and when to convey it.

Designing a good dialog management strategy is a complex task requiring many design decisions to be made. One of these decisions, which depends on the context of use for the agent, is whether the system is user-directed, system-directed, or has a mixed initiation (McTear et al., 2016b). In a user-directed system, the system only reacts to the commands of the user and does not ever prompt an action. Using a system-directed strategy, the system asks the user for information or prompts them for an action, and the user is only reacting to these prompts. The mixed-initiative systems allow both the user and the computer to take the initiative and enable the user to "derail" the conversation and introduce a new topic or question during the conversation. The latter approach, though, requires sophisticated natural language understanding, as the user might introduce so many new topics that the system loses track of its conversation agenda.

Since the information the computer receives can be ambiguous or incomprehensible for the NLP, the system can employ multiple strategies to keep the conversation within the limits of its functionality. Confirmation strategies and error handling are two main methods of handling ambiguous input (McTear et al., 2016b). Confirmations are needed if the system is not certain enough about whether it interpreted the input correctly or not. An explicit confirmation generates a direct inquiry about the ambiguous information (e.g. "Do you only want to search for restaurants in your area?"). An implicit confirmation includes some parts of the user's previous input, and the user can decide whether the system misrecognized their intention or not (e.g. "Do you want to search only for currently open restaurants in Copenhagen Center?").

Confirmation strategies are useful for avoiding misunderstandings if the system is uncertain about the information received. The "certainty" of the system is represented by a so-called rejection threshold (Bohus & Rudnicky, 2005): if the system can decide to interpret the user input as a certain user intent with a confidence score above the rejection threshold, then the corresponding follow-up action will happen. However, if not enough or no information is received, the system cannot decide on the user intent and cannot initiate a follow-up action. If this happens, the system must either prompt the user to repeat their input or rephrase their utterance. The former tactic can be used when, for example, a sound capturing error occurs and the user's voice input could not be interpreted e.g. because of background noise.

2.1.2 The Progress in Mind chatbot

A conversational interface can manifest in several ways, depending on the specific use case. McTear, Callejas, & Griol (2016a) distinguish three types of conversational interfaces: voice user interfaces, embodied conversational agents, and disembodied conversational agents or chatbots. Voice user interfaces, where the only modality of communication is speech, can be used for automating simple tasks (e.g. call routing or self-service through telephone) or as personal voice assistants (e.g. Siri or Google Assistant). Embodied conversation agents are computer-generated animated characters fitted with lifelike body language and facial expressions; these are increasingly used in the commercial sector for they are deemed more trustworthy than an inert computer interface (Bickmore & Cassell, 2001). Finally, chatbots (also called chatterbots) are created to emulate human-like conversations via text (or graphical elements) in a chat environment. The conversational interface of Progress in Mind belongs to the latter category.

The chatbot is designated as a supplement tool for searching the website and serves mainly as an interface for information retrieval – which will be discussed in detail in section 3. Therefore, the NLP architecture is organized around the *search* intent, which denotes the user's intention to search the database for content. To ensure that the conversation remains within the context of search, the chatbot uses a system-directed approach at the beginning. The first message of the bot clarifies its purpose and functionality: to search for "the latest news and interesting content" within the field of neurology and psychiatry. When the interaction begins, the bot automatically prompts the user to start searching the database by submitting a keyword to search or browse through some categories – thus ensuring that the domain of the conversation stays within the limits of the capabilities of the chatbot. In order to recognize which content to retrieve, the chatbot is trained with numerous entities that act as either browsing categories (e.g. *Videos*, *Podcasts,* etc.) or terms to search for within the publications (e.g. *alcohol dependence*, *dementia*, *patient management,* etc.). The bot can also handle contextuality – if the user first filters by a category, they are also able to search by a free-text query within the category. This is done by a so-called *contextual dataset*[1], which activates after choosing a category to browse. This prompts the bot to search the next submitted query within the currently selected category. This also means that even if the bot recognizes an entity that denotes a category, it will still handle it as a search query and not as a prompt to choose a new category. For example, if the user chooses Articles as a category to browse, but changes his or her mind and wants to search for videos instead, submitting the query "videos" will make the bot search for the term *videos* within the selected dataset instead of switching to another category.

There are special entities that, if detected, trigger a different intent from searching. One group of these entities are medicaments manufactured by Lundbeck, which the dataset does not contain any information about. If the name of a medical product is recognized, the bot offers a link to an external site of Lundbeck which lists all drugs against brain diseases the company produces (Figure 1, left). This feature is useful for informing users who might mistake the purpose of the bot as a generic product search tool.

Apart from product names, another category of entities is the "red flag" entities, which indicate that the user has searched for symptoms they experience themselves (e.g. *muscle pain*, *fever*, *diarrhea,* etc.). These are "everyday" symptoms, which are not likely to be searched by a psychiatrist in a medical resource center. When detected, it is assumed that the user tries to use the bot as a health information tool or medical helpline, which triggers the bot's "alert system" (Figure 1, right).

The search results in the chatbot do not contain the full content of the publication, but rather act as links to the content on the website. Therefore, the usage of the chatbot only covers the search process from the search tactic formulation until the selection of a search result. Lundbeck's goal was not to fit the entire website's content navigation system inside the chatbot, but rather to enhance the interactivity during the episode of content browsing.

[1] "Contextual dataset" is not part of the standard terminology, it is used as colloquial jargon to denote the logic where the NLP algorithm achieves reference resolution by inspecting the user's previous utterances.

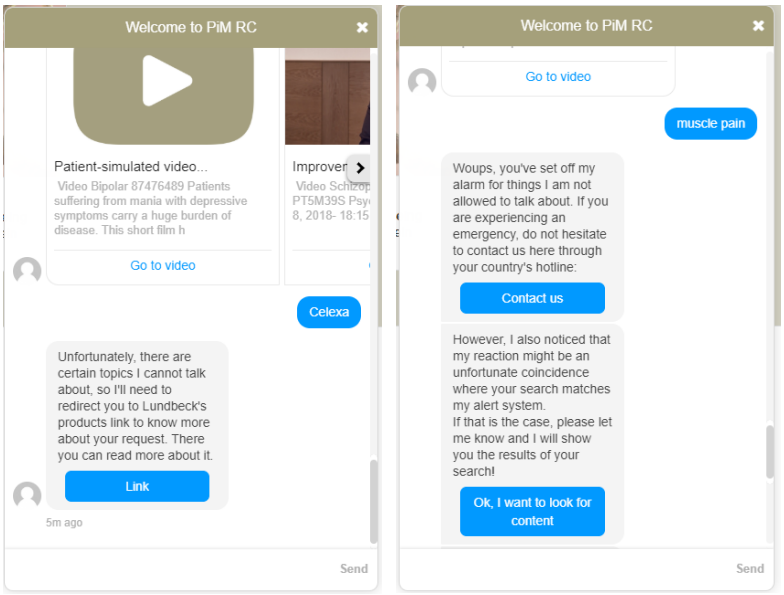

Figure 1: Special non-search-related responses of the chatbot. If the user searches for the name of a product manufactured by Lundbeck, it offers a link to a summary of medicaments Lundbeck produces (**left**). If the bot detects a possible medical emergency, it offers the user a link to contact the local hotline of Lundbeck (**right**).

2.2 User engagement

Lundbeck envisioned Progress in Mind to be the platform mental healthcare practitioners turn to when they want to keep up to date with the field and search for reliable resources. To achieve this growth in popularity, Lundbeck must not only draw in new users but also retain them by motivating long-term use. In that sense, the platform must not only be usable but also engaging. Lundbeck plans to use their chatbot as a novel interaction technique to enhance engagement. Considering the context of use, the Progress in Mind chatbot specifically enhances *message-based interactivity*, which is defined through the principle of contingency: "the idea that a given message is contingent upon the reception of the previous message and the ones preceding that" (Sundar, 2007). If the system does not only react to the immediately previous action of the user, but also considers the preceding actions, then the user will perceive the back-and-forth communication to be more interactive. During human-to-human communication this interactivity manifests naturally due to partners reacting to each other's utterances, which makes their information exchange a sequential flow. However, a web page already contains the message embedded within the site's content, therefore the only way to enhance message-based interactivity is to appropriately organize that static information (Sundar, 2007).

The Progress in Mind website already provides a certain level of interactivity through its search interface that responds to the user's queries and faceting actions (see details in section 3.1). The chatbot, however, imitates human-to-human communication, which enables even richer message interactivity. According to the user engagement-interactivity model by Sundar (2009), this enhancement in message interactivity should have a positive effect on user engagement if it does not lay additional navigational burdens on the user and does not impair usability. In order to assess how the chatbot performs in enhancing user engagement, the theory behind user engagement is to be discussed.

2.2.1 The psychology of user engagement

User engagement can be defined differently based on the field of context, whether it is organizational psychology, social platforms, or computer games – the common theme being the creation of positive experiences to motivate users during their activities (Edwards & Kelly, 2016). Therefore, user engagement is a subcategory of the holistic user experience (apart from utility, usability, and aesthetics (Sutcliffe, 2016)) which pertains to the phenomenon of how users are "drawn in" into the experience. Good user experience does not necessarily result in engagement, and vice versa: users can be engaged with a piece of technology despite that it provides a bad user experience (Lalmas, O'Brien, & Yom-Tov, 2014b). User engagement represents the way users gain value from the user experience, and it is the overall quality of that experience that leads to engagement. Researchers have been studying how user engagement manifests in various contexts like video games (Wiebe et al., 2014), online news (O'Brien & Cairns, 2015), information retrieval systems (O'Brien & Toms, 2010a) or healthcare (Sutcliffe et al., 2010). Aggregating the findings from their own and former studies, O'Brien & Toms (2008) proposed the generalized model of engagement, the stages it comprises, and the attributes of the experience that have the most influence in the specific stages. They identified four main stages: point of engagement, the period of sustained engagement, disengagement and possibly reengagement (Figure 2).

Users engaging with a system are generally drawn in by appealing aesthetics, interest in the content or the technology, or having some kind of intrinsic or extrinsic motivation. These attributes can be used to prompt usage, but not enough to sustain it. An essential contributor to user engagement is the so-called "flow state" (O'Brien, 2016). Being in flow means that users hardly perceive the passage of time, and become deeply immersed in the experience (Csikszentmihalyi, 1985). This immersion, though, necessitates a certain level of challenge imposed by the system, otherwise, the interaction leads to boredom – or anxiety, if the user's skills are too low compared to the challenge imposed. Users need to be in control of what's happening, which couples with a need for prompt feedback and appropriate communication from the system. This communication is also desired from the user's side, as being "part of the story" and experiencing rich interactivity also enhances the feeling of being engaged (O'Brien & Toms, 2008). Though the user engagement process can comprise a wide range of emotions, if the formerly mentioned attributes lead to an overall pleasurable experience, then the user is more likely to remain engaged with the system. The influence of the contributing factors can vary in intensity throughout the interaction. For example, in a graphics-heavy module of an application the aesthetics might play a more prominent role, but a more text-focused part should keep the interest of the user on the content, rather than trying to be visually appealing.

Disengagement happens when the user makes the internal decision to stop the current activity – either due to external or internal factors (O'Brien & Toms, 2008). This is the point where users lose cognitive stimulation and any emotional stance towards the system. One of the internal factors for users' disengagement can be the frustration due to poor usability or due to them being overwhelmed by the challenge they experience. On the opposite side, boredom due to the lack of challenge or lack of novelty can also lead to disengagement. External factors can be, for example, interruptions or distractions in the environment. These same attributes that lead to disengagement can also be the reason for nonengagement, where actual user engagement does not happen in the first place – the user never "immerses" him- or herself in the experience. If the overall affect during disengagement is negative, the user might cease to use the system indefinitely, whereas a generally positive affect might prompt a reengagement at a later point in time. Reengagement is influenced by the same attributes which contribute to the initial point of engagement – intrinsic motivation, interest, aesthetic appeal, etc.

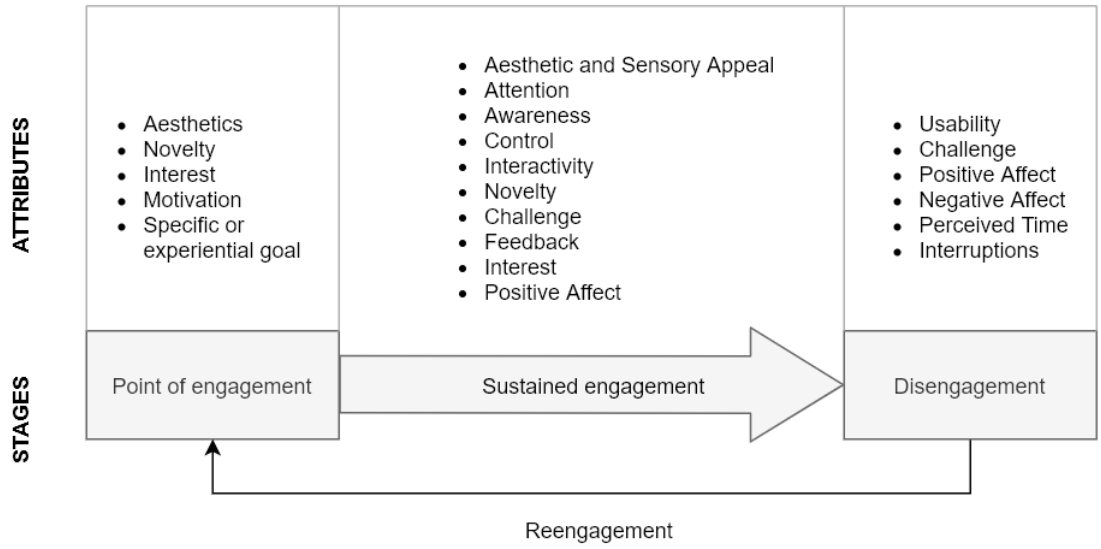

Figure 2: Model of user engagement with influencing attributes. Source: O'Brien & Toms (2008)

2.2.2 Measuring user engagement

As user engagement is a complex process, gathering data from multiple sources can be beneficial. Lalmas, O'Brien, & Yom-Tov (2014a) describe three main approaches for measuring user engagement. The first approach involves some kind of self-report method, via questionnaires, surveys, or interviews. One can also collect physiological data to gain insight into subconscious processes, such as facial expression analysis, eye tracking, heart rate measures, or measuring the skin's electrodermal activity. Finally, analyzing web analytics can also shed light on the behavior behind user engagement; methods include clickthrough rate analysis, page view statistics, dwell time on sites, return frequency, etc.

Considering the scope of the thesis, self-report measures are described in detail. These are used for eliciting information about the respondent's behaviors, beliefs, attitudes, or intentions (P. Lavrakas, 2008). In contrast to physiological data and web analytics, self-report measures rely on the respondent's report – thus it is assumed that users are capable to understand the question and willing to report their own emotions and behavior. However, users always interpret questions subjectively which brings in an element of bias in the answers. Problems with recall can also influence the authenticity of the reported information – users can report about recent experiences more accurately than the ones that happened in the past. Nevertheless, the high correlation between self-report data and physiological data confirms the validity of self-reported information (Lopatovska & Arapakis, 2011).

Interviews are one of the most commonly used self-report methods in user research. Earlier, interviews were more exploratory and aimed to uncover how engagement is experienced by users. O'Brien used interview techniques to find out which attributes influence user engagement (O'Brien & Toms, 2008), the findings of which are discussed in the previous section in the user engagement model (Figure 2). Researchers conducting studies can now utilize this knowledge to focus on these specific aspects of user engagement in their interview protocols. For example, in the case of a mobile app for meditation, focused attention seems to be an important factor for user engagement – thus the interview protocol can contain questions like "Was there anything that distracted you during your previous meditation session?".

Another self-report approach is the *think-aloud* method, where the user is asked to verbalize their cognitive processes during a task. The participant can be asked to report their thoughts

while solving a task, or after completing the task – the former method sometimes called *concurrent think-aloud* and the latter *think-after* or *retrospective think-aloud*. Employing concurrent think-aloud can clarify user's actions, and retrospective think-aloud can be used to get to know the motive behind those actions, as users have the opportunity to reflect upon what they did. This can shed light on the motives behind their engagement.

Questionnaires are another method for collecting data from individuals or groups. They contain open- or closed-ended questions and can be administered either via paper, digital, or in a one-to-one interview format. Questionnaires administered via interview follow a similar strict protocol as a structured interview, whereas pen-and-paper or online questionnaires do not necessitate the researcher to be present – which gives way to bias, as e.g. the respondent cannot ask for clarification if they don't understand a question (C. Wilson, 2013). Apart from the wording, the order of questions can also influence the acquired data – for example, respondent fatigue at the end of the questionnaire can lead to nonrepresentative responses. Since attaining validity and reliability is critical, designing a questionnaire is a rigorous multi-step task and requires an iterative evaluation of the outcome (Peterson, 2013). There are several questionnaires for measuring user engagement (Lalmas et al., 2014b) – the one developed most recently by O'Brien & Toms (2010b) is denoted as the User Engagement Scale (UES). The UES uses a scale-type approach, where respondent data is collected in a quantitative form to measure phenomena that we cannot directly observe (DeVellis, 2016). The UES has been formerly validated via two large-scale studies, and it contains 31 items divided into six categories related to focused attention, perceived usability, aesthetics, endurability, novelty, and felt involvement. The questionnaire items are formulated as sentences related to the categories, and the user has to evaluate via a Likert-scale to which extent they agree or disagree with the statement. The initial six-factor version of the scale has been revised (O'Brien, Cairns, & Hall, 2018), and through factor analysis and experimental testing the items were recategorized into four subscales instead of six:

- focused attention (FA): feeling immersed in the activity and being unaware of the passing of time;
- perceived usability (PU): negative affect due to the interaction and effort of interacting;
- aesthetic appeal (AE): the attractiveness of the interface; and
- reward (RW): the combination of the novelty (being interested and curious), endurability (overall success and the likelihood that the user recommends the application to someone) and felt involvement (being immersed in the experience, having fun) categories devised in the earlier version of the UES.

This revision also yielded a shorter form of the questionnaire with only 12 questions, named as User Engagement Scale – Short Form (UES-SF). This contains only three items in each category, therefore the response scores of all items can be summed together without any weighting and divided by twelve to get the overall user engagement score. As the short form does not cause as much response fatigue as the standard UES, it can be used for repeated measuring of individual experiences across different devices, interfaces, etc.

2.3 Measuring usability

Usability itself is a complex concept to define and measure. Nielsen (1993b) defines the constituents of usability as: learnability (how easy is it to learn the system), efficiency (how

productively can the user use the system), memorability (does the user has to re-learn the system after not using it for a while), errors (the system has a low error rate), and satisfaction (users are satisfied using and like using the system). The International Organization for Standardization (ISO) defines usability in a more concise manner with only three constituent elements: effectiveness, efficiency, and satisfaction (International Organization for Standardization, 2018). Though one can study a particular aspect of the system (e.g. comparing whether one system has better usability than another), this thesis will not go in-depth, and instead the focus will be on the overall usability of the interfaces. Usability is an important constituent of engagement, as, according to the engagement model by O'Brien & Toms (2008) in Figure 2, poor usability can be a major factor of disengagement.

Usability is generally measured by assigning users several pre-described tasks to solve (Nielsen, 1993d), during which certain metrics are measured. These metrics are then compared across the interfaces to draw conclusions about their relative usability. For example, performance measures provide quantitative measurements that are easy to compare (Nielsen, 1993c), and also enable statistical analysis. Measures of effectiveness can be e.g. error frequency or task completion rate; efficiency can be measured via task times or the number of unnecessary actions to complete the task (Bevan et al., 2016). Most other methods elicit qualitative data, which require additional effort for analysis and comparison. The simplest method that can be coupled with the task-based evaluation protocol is the think-aloud approach, where a user is asked to verbalize what they are thinking (Nielsen, 1993c). This way, one can gain insight into the misconceptions users have about the system, and elicit remarks about the problems they face (Boren & Ramey, 2000; Van Den Haak, De Jong, & Schellens, 2003). However, it must be taken into consideration that think-aloud protocols can influence performance measures, as users might be slowed down due to the cognitive effort of vocalizing their thoughts (Nielsen, 1993c) – which might not be a problem for comparative studies where users are thinking aloud during teach task, as differences between performance measures will still be apparent.

Information about usability can also be elicited outside of the task setting – for example, via questionnaires. Questionnaires enable the researcher to collect self-report data not about the interface itself, but the opinions users have about the interface (Nielsen, 1993b). For eliciting valid data, users should first interact with the system to be able to form an opinion about it – otherwise their speculative answers can be misleading. As discussed in the previous section, questionnaires can be administered either in an interview, digital, or pen-and-paper form. Questionnaires can be especially useful for measuring one particular attitude: the user's satisfaction with the system (Nielsen, 1993d; Sauro & Lewis, 2009). O'Brien & Tom's (2008) model, Nielsen's (1993d) definition and the ISO standard all consider satisfaction or positive affect an important criterion during interaction – which makes questionnaires an ideal tool for eliciting valid data.

3 Progress in Mind as an information retrieval system

The Progress in Mind platform was launched in 2017. It stores data in an organized structure and enables users to retrieve those pieces of data via a search user interface (SUI). This classifies the system as an *information retrieval* system, where *information retrieval* (IR) denotes the discipline that is concerned with the storage, maintenance, indexing, and retrieval of data (Minker, 1977). Although this traditional definition solely concerns the capabilities of the IR system itself, nowadays the field of IR also takes into account the cognitive processes of users engaging with such systems.

3.1 Information retrieval theory

Information retrieval as a field emerged from librarianship, where, seeing the chaotic situation in public information access, practitioners of the field sought ways to make their databases efficiently browsable and establish guidelines for documentation to make retrieval of information more successful (Meadows, 2002). The arrival of the computer replaced the cumbersome human labor of document annotation and indexing due to digital accessibility and easy-to-manipulate database structuring. When the IR field was still maturing, IR systems were evaluated as if they were bibliographical collections: the main goal of the system was to only display items that are relevant to a field and eliminate everything that is not. This stipulated that the usability of the system shall be assessed exclusively on its functionality – assigning *relevance* as the main criterion of assessing usability. From here stems the traditional model of information retrieval (depicted in Figure 3). Users engaging with IR systems are driven by their *information need*, which is a gap between their actual information knowledge state and their desired information state (White & Roth, 2009).. This information need translates to a query, for which the system must retrieve a set of documents that satisfy the need of the user.

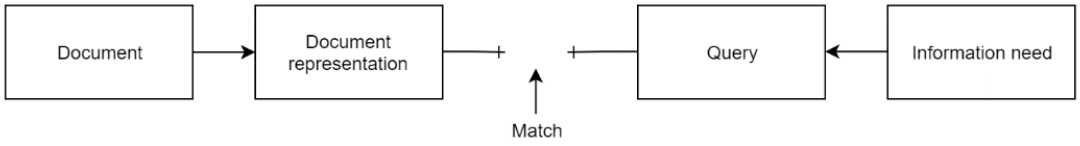

Figure 3: The traditional information retrieval model. Original diagram by Bates (1989).

As the field of human-computer interaction developed, the shortcomings of system-oriented approaches became clear (Shackel, 2009). Information scientists realized that the behavior of users needs to be addressed in order to ensure less demanding and more successful interactions. Instead of viewing IR as isolated query submissions, they approached entire "search episodes" holistically, and acknowledged that the information need of the user changes constantly throughout the interaction (Belkin, 2010). They also acknowledged that, based on the current knowledge of the user, information needs are not always well-defined, but they can be exploratory in nature (Ingwersen, 1996). Thus the discipline of *interactive information retrieval* (IIR) was developed, which considers the interplay between the user and the system in its entirety and evaluates the usability of an IR system from a user-centered point of view (Xie, 2008b). Here, the focus is not on whether the system can retrieve relevant documents, but whether the user is able to use the system to retrieve relevant documents (Kelly, 2009).

Users might engage in information retrieval due to various reasons, such as simple fact-checking to verify a piece of information, to learn about a field of interest, or to conduct a deep

analysis of data and arrive at a synthesized conclusion (Marchionini, 2006). Though an IR system should satisfy sophisticated IR tasks, most SUIs, such as the web search engines, cater mainly for simple information lookups. Users who engage, for example, in exploratory search are driven by information needs which cannot be satisfied via a simple lookup (Palagi et al., 2017), and therefore need advanced search functionalities to aid their browsing, e.g. suggestions for query reformulation, autocomplete, autocorrect or suggesting related keywords (Beckers & Fuhr, 2012; Russell-Rose & Tate, 2012c). The functionalities of the system, therefore, define the IR behavior of users by enabling/denying them the use of certain search tactics (Bates, 1979). In the following section, the functionalities of the Progress in Mind platform are assessed.

3.2 IR functionalities of Progress in Mind

3.2.1 Under the hood: the Solr search engine

In order to provide the system functionalities for IR, Progress in Mind is built upon a robust search engine called Apache Solr. It is an open-source search server written in Java and used widely in Online Public Access Catalogs (OPACs) (Serafini, 2013b). Though Progress in Mind does not classify fully as an OPAC (e.g. it is not owned by a library and it does not only contain bibliographical data), there are some aspects of its database structure that necessitate the usage of an advanced search engine. For example, similarly to OPACs, Progress in Mind also has dynamic content (new content can be added and content can be updated in real-time), enables keyword searching, and can be accessed from any location via a web interface (Xie, 2008a).

The Solr engine is built upon the Lucene library, which is a full-text search library in Java (Serafini, 2013c). Lucene's core elements are as follows:

- Document: an internal representation of data, which is essentially a collection of fields (e.g. an article about Treatment resistance among patients suffering from schizophrenia);
- Field: a piece of data that consists of a field name and a value, which describes the content of the document. These fields can be used for, for example, storing metadata and they can have multiple values (e.g. a field for storing the information about the keywords describing an article can be denoted as *field name="keyword",* and the values of the field can be *"Schizophrenia"*, *"Treatment resistance"* etc.);
- Term: the basic unit for indexing, usually consists of a single word (e.g. the *keyword* field's value *Treatment resistance* consists of two terms, *treatment* and *resistance*);
- Index: the in-memory structure where Solr performs the search. A document can be seen as a single record within the Index.

Figure 4 shows how the Lucene Index is structured: searching for the query Solr Book within the field title, the system is expected to return all documents with the field-value pair *title: "Solr Book"*. Solr, being a full-text search system, is not restricted to only search within the various fields of the documents, but it can also scrutinize the content of the document. Searching directly within the entire document texts, though, would be a slow process. Instead, the document content is *indexed*, which is a process where the most descriptive terms are extracted from a text. The process starts with collecting the entire vocabulary from a text, which is then "cleaned" by removing the most frequent (*a*, *an*, *the*, *and*, *of,* etc.) and least frequent terms, collective called *stopwords,* and transforming the remaining words into their lexical

roots via *stemming* (e.g. the words *stressed*, *stressing*, and *stressful* are combined into the single term *stress*). This way, ideally, only those terms remain in the vocabulary that are characteristic for the text. These terms are then collected from across the database and each document is assigned to one or more terms, depending on whether the document's vocabulary contains that term or not. This way we create an index table, where one term is assigned to one or more documents. This structure is a so-called inverted index (Dominich, 2008), which also contains information about how many times a term occurs in a specific document, and describes how the Lucene Index is structured.

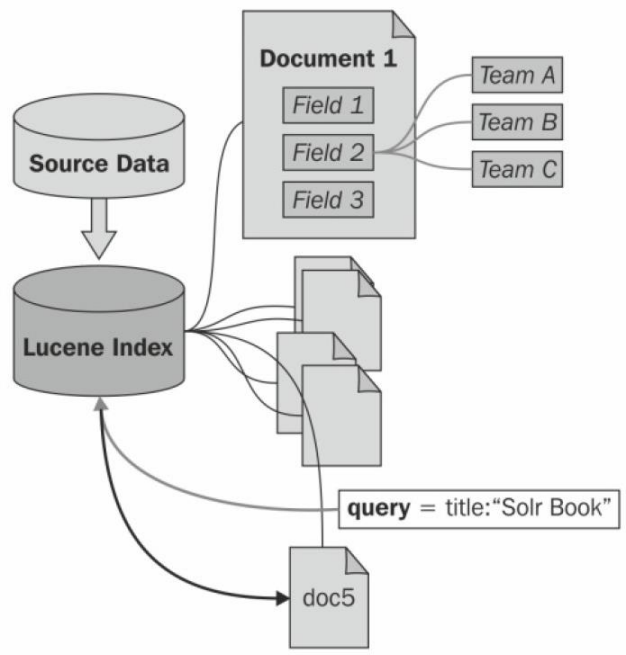

Figure 4: Schematics of the Lucene Index which Solr is built upon. Source: Serafini (2013a).

This inverted index ensures fast retrieval of documents since it is only the indices that are searched, not the document contents in their entirety. The retrieved documents which contain the keywords extracted from the query are then ordered by relevance according to the term frequency. The more times a searched-for term occurs within a specific document, the higher it gets in the ranking list. This way the documents which, ideally, contain the most relevant content in terms of the query are shown first in the result list.

Solr also supports advanced search functionalities, like query suggestions, faceted navigation, snippet creation, etc. For easier understanding, these functionalities are discussed in relation to how they are implemented in the user interface, both in the website SUI and in the chatbot.

3.2.2 Functions of the search user interface

Session start and faceted navigation

Facilitating the start of a search session is of great importance for task success, especially for users who are at the beginning of a research project (Ellis, 1989). This is the stage where the user gets to know the capabilities of the system, thus the system functionalities must be exposed. The website immediately displays its *faceted navigational* capabilities for new

visitors by listing the various facets in the header area, which users can use to filter the content. The content is organized around the topics of depression, migraine, schizophrenia, Parkinson's disease, bipolar disorder, and Alzheimer's disease. Users can also filter content either by the topics of the diseases or by the type of publication: congress highlights (containing summaries and excerpts from medical conferences), podcasts, and expert views (video interviews with healthcare professionals). In IR terms, these categories constitute *search facets*, which serve as a way for partitioning the database along non-overlapping categories (Russell-Rose & Tate, 2012b). Every document stored within the database is indexed with both a *category* field that describes which disorder the content is about (denoted as Topics in the user interface) and a *type* field that describes the type of publication. Therefore, the same item can be found either by browsing through topics or by searching across the publication types – a model called polyrepresentation, where documents are represented with as many aspects of information as possible, leading to a so-called *intentional redundancy* of information (Ingwersen, 1994). This enables the IR system to flexibly comply with the user's cognitive model or information needs, which can vary from person to person (Beckers, 2009). For example, a healthcare specialist interested in the latest development of treating schizophrenia might browse the topic Schizophrenia, while a nurse who is currently doing routine paperwork might want to listen to some podcasts in the background – and by chance, they might listen to the same podcast about a novel schizophrenia treatment. If a facet is chosen, users are taken to a list of publications of the chosen topic, ordered chronologically by the date of publication.

On the landing page, users can sign up for the Progress in Mind newsletter which contains weekly digests of new publications on the website – a function which caters for the advanced IR task of "maintaining awareness of developments in a field through the monitoring of particular sources" (Ellis, 1989).

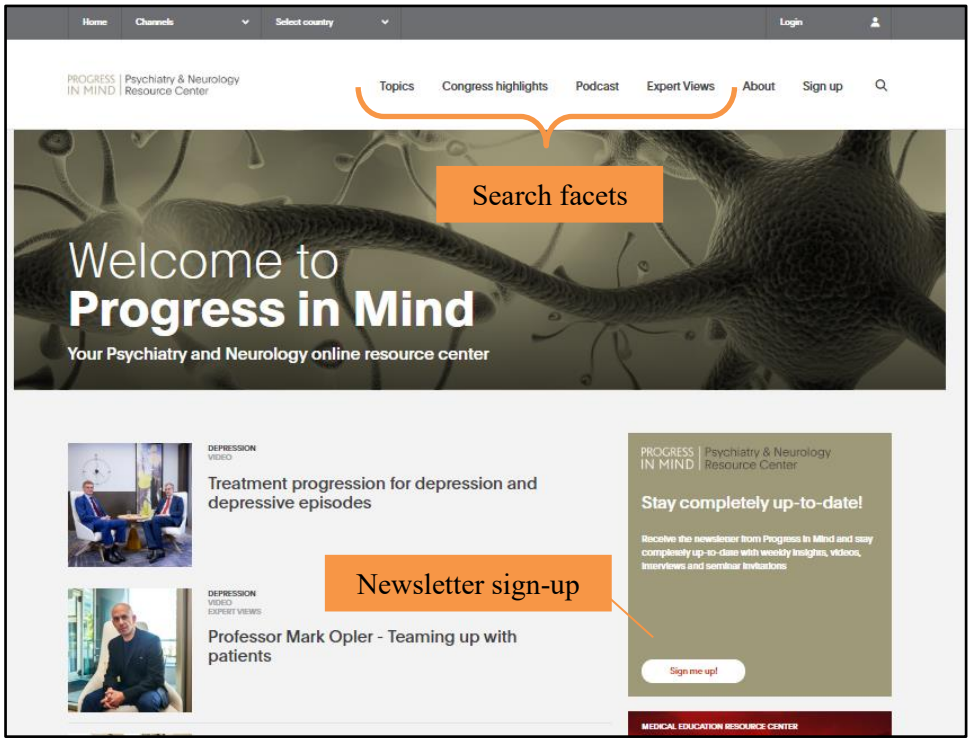

Figure 5: Landing page of Progress in Mind with exposed search facets.

When the user starts interacting with the Progress in Mind chatbot, it requests a confirmation from the user that they are a healthcare practitioner –to ensure that the user possesses the field knowledge to make use of the content[2]. Similar to the webpage, the chatbot also lists the search facets at the beginning of the interaction in a scrollable carousel (Figure 6), though the facet names are different. Users can choose from the category *Video*, *Articles*, *Congress Highlights*, *Interviews*, *Podcasts*, *Conferences* or choose *Mix it up!* which gives a blend of all publication types. The bot clarifies that the user can either submit a free-text query or choose one of the facets for browsing. It is noticeable that these facets do not fully correspond to those within the website, though the chatbot provides more elaborate and straightforward categories compared to the website. For example, the *Expert views* facet on the website UI contains mainly videos, therefore the name *Expert views* might lead to confusion – whereas the *Video* facet provided by the chatbot is more obvious.

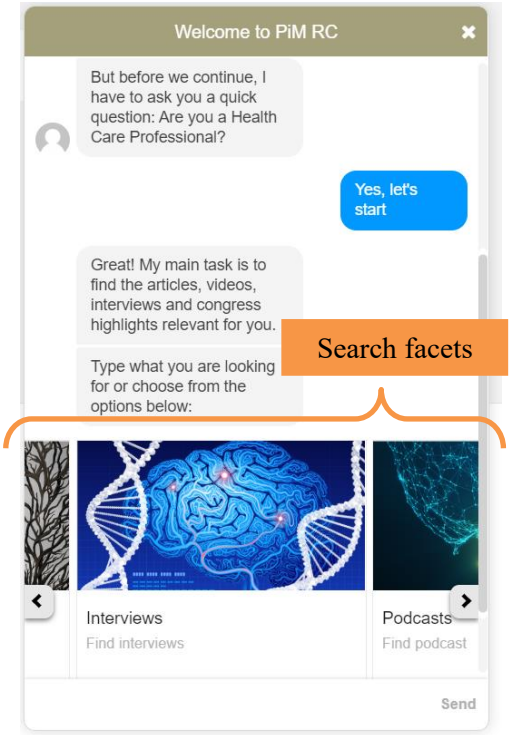

Figure 6: Progress in Mind chatbot revealing the search facets at the beginning of the interaction.

Search engine results page

On the website, if the user selects one of the facets, they are taken to a search engine results page (SERP) that contains all documents indexed with the corresponding facet field. Apart from topical browsing, the system allows free-text search queries and displays a search result page with content that contains matches to the submitted query (Figure 7). Users can use Boolean operators (AND, OR) to refine their queries, which is an efficient technique for enhancing literature searches in the medical field (Baumann, 2016; Lowe, 1994); although this functionality is not made apparent for the user. If no relevant results can be retrieved, the system displays some tips to avert such zero-result pages. This help text contains information about

[2] This might be limiting in a sense that it excludes people who, though not experts in neuropsychiatry, have a legitimate user need for using the site (e.g. students of the field, psychologists with an interest in medical practices etc.).

the possibility to use Boolean operators and exact-term searching by using "double quotes" around the query. The system uses Solr's default settings where a query with multiple terms are connected with OR logic (Serafini, 2013d) unless the user specifically adds AND in between the keywords. This means that submitting the query *depressive episode* will yield results that contain either the term *depressive* or *episode*. If modified as *depressive* AND *episode*, then only those documents are retrieved which contain both the terms *depressive* and *episode* – though not necessarily appearing next to one another in the text. When submitted with double quotes – *"depressive episode"* – the engine returns documents that contain the exact term *depressive episode*. This feature is called *phrase search*, and the search engine handles the phrase between the double quotations as one single term (Serafini, 2013d).

The results are comprised of metadata (topic, type of publication), which serve as previews for the result, the title, a picture, and a short query-oriented summary. The summary is an excerpt from the content with the query keywords highlighted and displayed in the context of sentences (Russell-Rose & Tate, 2012a). The highlighting of the matching keywords enhances relevance feedback, as the user can see how closely related the retrieved document's content is to the query (e.g. if the query terms appear after one another as one term or separately within the content) (Muramatsu & Pratt, 2001).

If the system detects a possibly misspelled keyword, it suggests reformulating the query via a "Did you mean" suggestion (Russell-Rose & Tate, 2012c). This happens according to a fuzzy matching algorithm (Serafini, 2013a), which, when encountering a possible misspelling, tries to match the query with an existing index with a certain confidence score. This feature does not only trigger when a potential misspelling is detected but also if there exists a differently formulated query that might yield better results. For example, in Figure 7 the system suggests "imaging" instead of "neuroimaging" (which demonstrates the defects of the suggestion algorithm, as this query edit might not be convenient in terms of results relevance). Thus, the feature acts both as a misspelling detection tool and a general query suggestion function.

The chatbot contains an alternative search pattern compared to the website. When the user the categories for publication type are displayed, the user can either select one of the facets or submit a free-text query. If they submit a query, the user can use the chatbot the same way as the website's free-text search box and search the entire database. If they select one of the facet categories, the chatbot will display 9 random results from that category (Figure 8, top left). The recommended items that should be displayed are selected by the database administrators. The results can be scrolled through in a carousel, and they contain a picture, the title, the beginning of the content, and a button that displays the publication type and leads to the publication URL on the website. Users can either select the New Search button or just type anything in the message field, and it will be treated as a query. After submitting a new query, the bot displays results that match the query, and which belong to the same publication type which the user selected before (Figure 8, top middle). The chatbot does not convey in any way that the submitted query will be searched only within the previously selected facet. If the user submits another query after this, then the chatbot will exit the facet and instead search for matches within the entire database (Figure 8, top right). After this point, the user can return to a facet by starting a New Search and typing in the name of the facet – these are defined as entities to the bot (see section 2.1.1) so that it understands that it's not a free-text query that needs to be searched across the database. This means that the user can only search within a facet only once, after which they have to return to the facet again to submit another within-facet query (Figure 8, bottom).

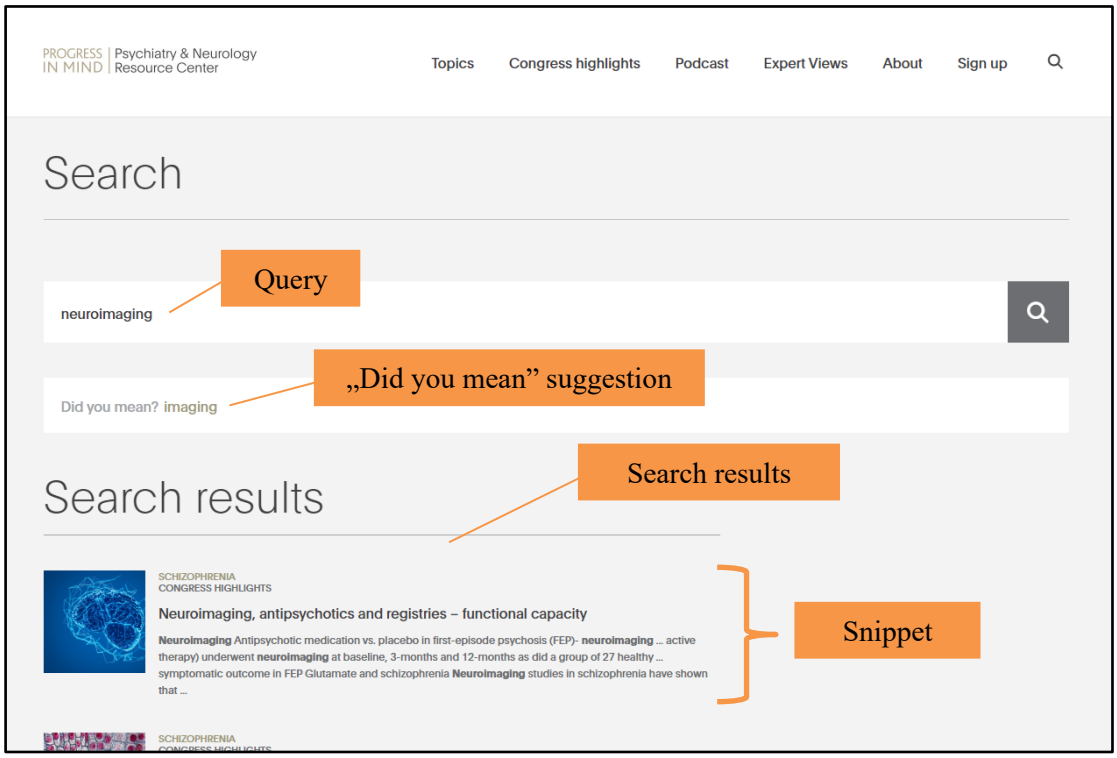

Figure 7: The search result page of the Progress in Mind platform and annotations that detail the certain search functionalities.

Considering the search interaction, the chatbot is a "Partial Item System - Free Text User" system (Radlinski & Craswell, 2017). These types of conversational search interfaces enable the user to submit free-text queries (as opposed to simply rating suggested result items) and utilize slot filling (see section 2.1.1). "Partial item" means that a result only contains a subset of the document information – similar to a preview or snippet. Using the taxonomy defined by Radlinski & Craswell (2017), the search interaction is as follows:

1. The system provides multiple categories to facilitate the beginning of the search (a_p^{2+});
2. The user sets their preference among the suggested clusters (r_p) or provides unstructured text to express their information need (r_t);
3. The system provides a set of partial items (a_p^{2+});
4. The user provides unstructured text (r_t) which either contains a facet name or search keywords, then the process continues from step 3.

The standard interaction form of the chatbot is $(a_p^{2+}) \rightarrow (r_t) \rightarrow (a_p^{2+}) \rightarrow (r_t) \ldots$ and the type of results the system displays in (a_p^{2+}) depends on the submitted user query and the context of the conversation. Figure 8 (bottom) details the chatbot-user interaction process along with context.

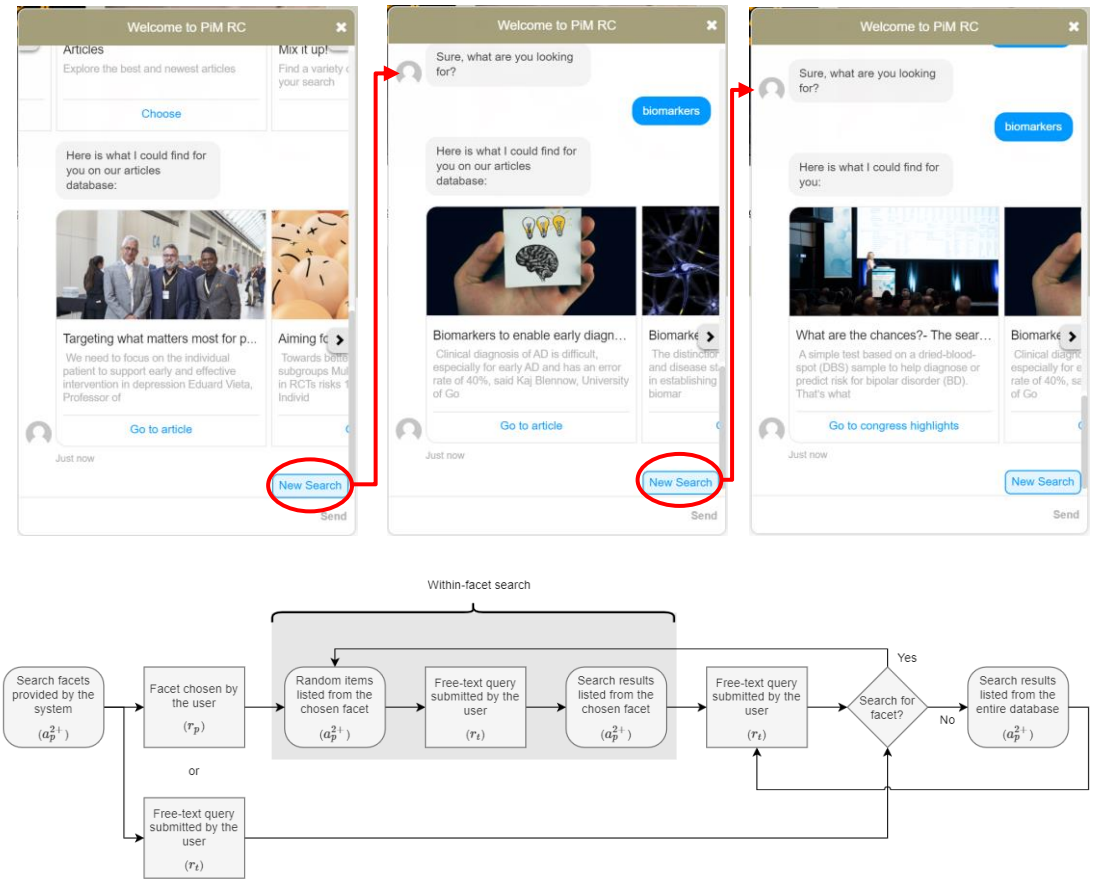

Figure 8: SERP display and the generic flow of search modes in the chatbot. The chatbot displays 9 randomly chosen publications from the category which can be scrolled in a carousel (**top left**). After clicking on New Search, the user can submit a free-text query that searches among documents of the previously selected category (**top middle**). After submitting a query again, the facet-search criterion is cleared and the newly submitted query will be searched throughout the entire database (**top right**). The generic flow of the search interaction is summarized in the flowchart (**bottom**).

Full publication view (website only)

If the user clicks on one of the results, either within the website or chatbot SERP, they are taken to the full publication (Figure 9). Apart from the main content, this page contains useful browsing tools, one of them being the tag system. Each document is identified by multiple tags (Figure 9), which are defined in the custom index *detail-tag* in the Solr database. If one of the tags is selected, the user is taken to a partitioned result page that lists all documents with the same tag. Unfortunately, these tags are only accessible once the user is already viewing a publication and they cannot be found in the SERP. Apart from the tags, users can find at the bottom of the page a list of references that are cited throughout the publication. This can be a useful feature for searchers who utilize citation chaining as an IR technique, as they can find further reading if they want to explore a certain topic (Bates, 1989).

Table 1 summarizes the search functionalities of the website-based SUI and the chatbot SUI.

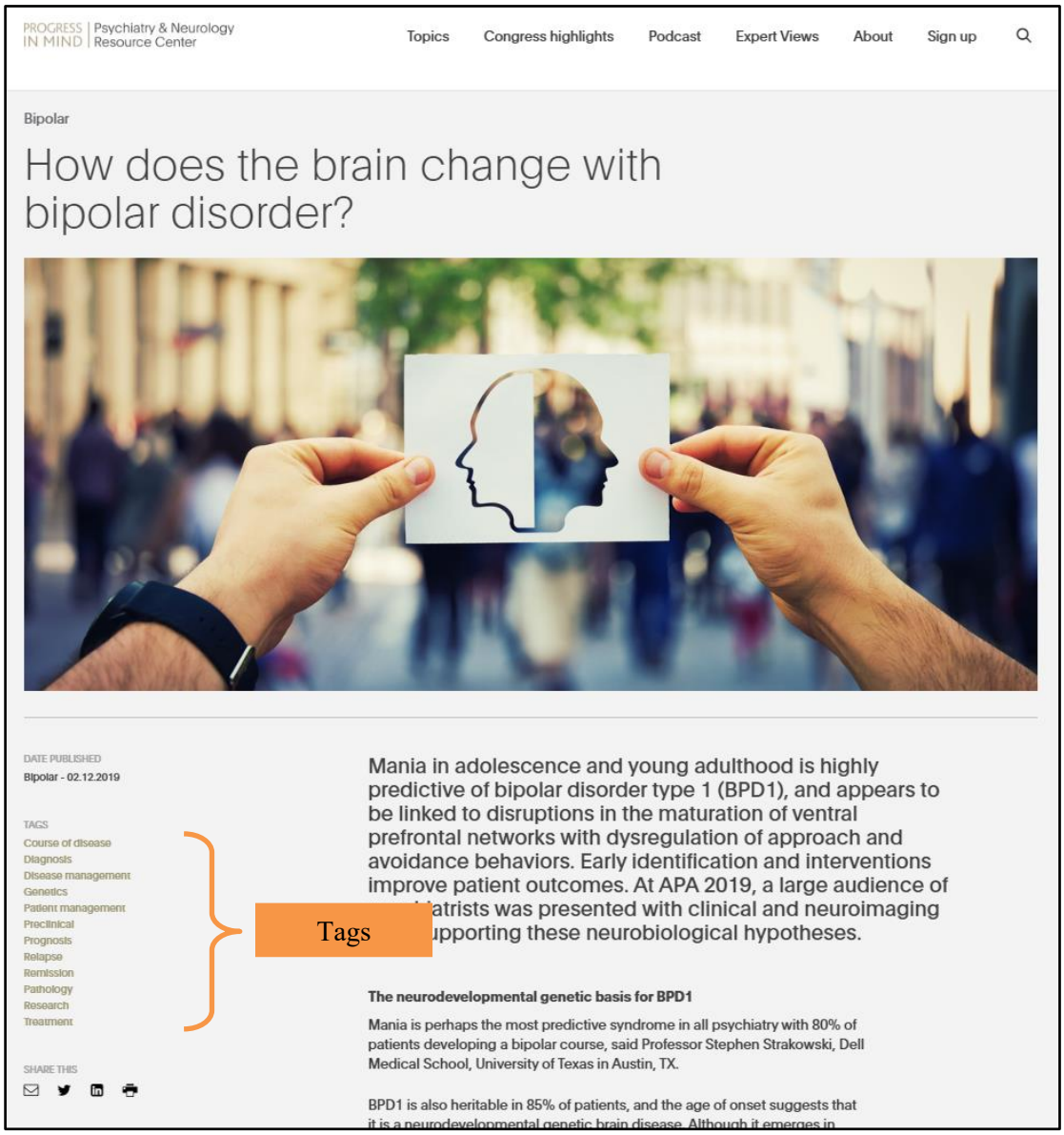

Figure 9: An example article. Topical tags are displayed on the left side.

Table 1: Comparison of search functionalities of the website and the chatbot SUI

Search feature	Website	Chatbot
Facilitating the start of the search	• Exposing facets • Newsletter for field monitoring purposes	• Exposing facets
Faceted navigation	• Topics (within that, 6 diseases), Congress	• Video, Articles, Congress Highlights, Interviews, Podcasts,

	highlights, Podcasts, and Expert views • Cannot be combined with query search	Conferences, and “Mix it up!” (i.e. every category combined) • Can be combined with query search, but one has to return to the facet again and again
SERP functions	• Search suggestions • Boolean operators • Results list is vertically scrollable • Displays every result	• Link to full publication • Results list is horizontally scrollable via a carousel • Displays a maximum of 9 results
Snippet components	• Topic • Publication type • Title • Content excerpt with query highlights • Picture	• Picture • Title • The first few sentences from the content

4 Related work

Conversational search and system usability

Conversational search is still a novel branch of IR, but it is becoming more popular thanks to the proliferation of voice assistants. As mentioned before, users may have difficulty adapting to conversational search, since the majority of search interfaces are based on a graphical user interface. Graphical SUIs set the standards for digital information search, and the majority of IR system design principles are based on graphical representation – e.g. faceted search (Tunkelang & Marchionini, 2009) or SERP control features like sorting, filtering or grouping (M. L. Wilson, 2011).

Due to this novelty, literature about his field is scarce, and most comparative studies do not focus specifically on search systems. For example, Ischen et al. (2020) compared a website, a human-like chatbot, and a machine-like chatbot and studied the effects of the interface on anthropomorphism and privacy concerns via questionnaires. One of their findings was that the website elicited more privacy concerns in users than the machine-like chatbot, which lead to less information disclosure (interestingly, no such difference was found between the human-like chatbot and the website). Celino & Re Calegari (2020) investigated whether administering surveys via a conversational interface is a reliable and user-friendly method for data gathering. They tested a website-based survey, a chatbot with informally formulated questions, and one with formally formulated questions via A/B testing and collected preference data via questionnaires. They found out that users have a preference towards the chatbot-administered survey, and that a chatbot-based method is at least as reliable in terms of inter-rater reliability as the website-based one. Sundar et al. (2016), whose work has been mentioned before, is the only study of this type that focused on an interface that is sued for search. They compared several types of interfaces for a movie search website with varying levels of message interactivity, which they manipulated by adding/removing search history functionalities and a chatbot for assisting users in their browsing. They found that providing interaction history and the possibility for chatting with a live agent significantly increased perceived contingency, and subsequently, interactivity, which affected user engagement positively. Apart from the latter, no literature was found that compares the performance of conversational and graphical search user interfaces – therefore, the focus will be on conversational search interfaces in general.

Vtyurina et al. (2017) explored users' preferences towards conversational search interfaces of various sentience. Participants completed exploratory search tasks with three types of chatbots: the first was a commercial chatbot, the second was a human expert (where participants knew they interacted with a human), and the third was a "wizard" where the chatbot was covertly operated by a human but participants thought they interact with a machine. They found that most users preferred the human or "wizard" chatbots as both were able to interpret half-sentences, whereas the machine struggled with reference resolution (see section 2.1.1), which also negatively affected participants' search task performance. Dubiel et al. (2018) found similar differences in task performance and user satisfaction where they used a Wizard-Of-Oz-style study to explore two hypothetical spoken dialog systems: a standard voice bot using a slot-filling algorithm and an intelligent "conversational search agent" with a memory component for handling contextuality. Participants were significantly more successful with their tasks when they used the agent with a memory component, and they found it less taxing and displayed a more positive sentiment towards it compared to the slot-filling agent. This points towards the users' need for more human-like conversations where chatbots have contextual awareness – preferably without asking too many questions for confirmation (Dubiel, 2018). However, user expectations about the capabilities of conversational interfaces are usually met with disappointment. Luger & Sellen (2016) conducted a qualitative study using

interviews and thematic network analysis to explore the mental models that users have about their voice assistants. They found, as they denoted, a "deep gulf of evaluation": users reported their confusion about the capabilities of the voice systems, as their expectations were not met. The in-built playful responses (e.g. the capability of telling jokes) also set unrealistic expectations about the sophistication of the system, and after continued disappointment users became reluctant to use their voice assistants for complex tasks.

Seeing that the discipline of conversational search still lacks profound research, Thomas et al. (2017) collected a rich dataset of search-oriented conversations called MISC (Microsoft Information-Seeking Conversation data). The participants of the conversations consisted of a searcher, who was given a search task, and an intermediary who had access to the internet and was tasked to follow the searcher's directions and provide feedback only via voice. These conversation recordings are created to help to establish desiderata for an optimal conversational search system and demonstrate users' desires for an *aligned* discourse with conversational interfaces. *Alignment*, in this case, means that the user and the system can match each other's style of communication in terms of involvement (chit-chattiness, verbosity, enthusiasm) or considerateness (more listening, hesitance, independence). If alignment succeeds, then task execution becomes more efficient (Thomas et al., 2018).

Supporting long-term engagement via conversational interfaces

Chatbots are used in many areas to engage users, but they can be particularly beneficial when it comes to longitudinal interventions. During longitudinal interventions (e.g. healthcare coaching or education), user engagement is crucial for ensuring continuous usage and reducing user dropout (Scherer et al., 2017), and conversational agents have a promising potential for sustaining engagement.

Regarding longitudinal interventions, behavioral health interventions are one of the main areas where conversational agents are proliferating, and a continuously growing body of literature exists to prove their effectiveness (Vaidyam et al., 2019). Perski et al. (2019) conducted an experimental study to find out whether the addition of a chatbot to a smoke cessation application lead to a higher rate of usage. The addition of the chatbot more than doubled the frequency of usage – though this did not clearly lead to successful behavior change. Chatbots were also shown to be useful for facilitating computer-assisted mental health therapies. Fitzpatrick, Darcy, & Vierhile (2017) conducted a randomized trial to study how a chatbot-based cognitive behavior therapy performs compared to an eBook-based therapy in helping students who suffer from depression and anxiety. Their conversational agent, Woebot, elicited a higher frequency of engagement coupled with a significant reduction in anxiety, compared to the control condition. Fulmer et al. (2018) conducted a study with a similar setting, where their conversational agent, Tess, elicited an even greater effect size compared to Woebot. They attributed the greater effect size to the fact that their chatbot, Tess, relied more on free-text-based conversations compared to Woebot, which mainly used buttons and quick-replies, and that Tess provided more personalized interventions. The positive effect of using a conversational agent for facilitating engagement during app-based mental health therapies was also demonstrated by Ly, Ly, & Andersson (2017), where the addition of a chatbot to a cognitive behavioral therapy mobile lead to higher interaction frequencies and decreased depressive symptoms – though the latter only applies to participants who fulfilled the completion criteria of the intervention. Personalization of content in conversational agents is shown to be an efficient method for motivating user engagement and improve user satisfaction in healthcare (Kocaballi et al., 2019). Recommendations and intervention plans which are tailored to the user create a feeling of connectivity and the humanlike dialog establishes a feeling of trust, which can be particularly useful in sensitive topics like mental health (D'Alfonso et al., 2017).

Apart from healthcare, education is another domain of interest where the application of conversational interfaces can facilitate engagement – though only a handful of reports employ experimental study designs to draw a statistically meaningful conclusion about their effectiveness. Milne et al. (2011) developed an embodied conversational agent for teaching conversational skills for children with autism spectrum disorder. Analyzing pre-test and post-test scores before and after the usage of the virtual tutor, they found that the agent contributed to the improved conversational skills of the children significantly, with an average improvement of 32%. Pereira (2016) developed a chatbot as a supplementary tool for helping students prepare for school exams through multiple-choice quizzes. Though they did not find a significant correlation between frequency of the chatbot use and final test scores (in fact, a slight negative correlation was found), qualitative self-reports indicated that students generally find chatbots useful for engaging with a subject.

Despite some of the promising results, the technology of chatbots still imposes considerable limitations, especially when used for facilitating engagement in language learning. Fryer et al. (2017); and Fryer, Nakao, & Thompson (2019) conducted longitudinal experiments where students undertook a language course either with human conversational partners or a chatbot. Analyzing students' self-report data, they found that a chatbot cannot meaningfully facilitate learning interest and task engagement if the student is lacking interest in human-to-human conversations. Thompson, Gallacher, & Howarth (2018) also found that conversational agents are not as capable of maintaining student interest in language learning over a longer period, compared to a human partner.

It shows the novelty of the field that most studies about conversational interfaces have been published relatively recently. When it comes to conversational research, human-operated conversational interfaces (live agents) seem to be notably preferred by users compared to NLP-driven systems, as the latter still cannot comprehend sophisticated information needs. This shows that conversational technology is still lacking when it comes to assisting in search, although its ability to incorporate humanlike communication into the interaction holds interesting potential for other purposes. The conversational modality can be an efficient method for increasing user engagement, and its humanlike communication style can elicit anthropomorphism, which can lead to increased trust. As only one study touched upon both the aspect of user engagement of conversational search (Sundar et al., 2016), the question still remains whether a chatbot is capable of, despite a generally poorer utility, facilitate engagement.

5 Methodology

The following section shortly details the research paradigm considered for the study, which is followed by the description of the research methodology. For designing the study, the recommendations by Kelly (2009) for evaluating interactive information systems have been followed.

5.1 Research paradigm

For establishing a basis upon which a research methodology can be built, the paradigmatic underpinnings of the research are discussed, according to Pickard (2013). To reiterate, the aim of the study is to find out the differences between the chatbot SUI and the website SUI in terms of engagement and usability. From an ontological stance, neither engagement nor usability are tangible concepts that can be measured directly with physical tools – which would be the basic requirement for a positivist approach. Moreover, a positivist viewpoint would also necessitate that one could describe engagement with absolute laws (i.e. users who are engaged would be "made subject to a single set of laws" of engagement (Lincoln & Guba, 1985)), which is not the case – as engagement manifests in many different forms. Therefore, a positivist approach must be ruled out – as that would stipulate that speculative thinking is to be reduced to a minimum.

On the other end of the scale, pure interpretivism would dictate that we accept multiple realities, and we cannot assign any generalizable truth to our observations. It would also stipulate that the researcher and the subject inadvertently influence each other. The uncertainty this paradigm imposes would make a comparative study futile, as it would mostly rule out quantitative methods, and deny the opportunity for hypothesis testing.

Therefore, a postpositivist paradigm will dictate the methodology, which allows quantitative approaches, but also puts great emphasis on acknowledging the context. It favors a mixed-methods approach, where quantitative hypothesis testing is coupled with qualitative data for interpretations of the results. The main difference postpositivism has compared to positivism is that a hypothesis can only be falsified, but not proven – that is, the purpose of the study is to disprove that a phenomenon exists. Therefore, the aim of the study is to disprove the null hypothesis, which is that the usage of the chatbot does not lead to a greater level of engagement. For this, an experimental methodology will be followed, which is described in the following section.

5.2 Research design

An experimental approach (Bryman, 2016) will be followed to investigate whether the type of interface used for searching, the *independent variable*, influences the elicited user engagement and system usability, the *dependent variables* (Kelly, 2009, p. 44). In order to enable participants to compare the two systems (Kelly, 2009, p. 50), a within-group study approach is followed. A between-group study would not enable participants to make comparisons, and it would also necessitate a larger sample size to gather the same amount of user feedback about both systems, which is not feasible considering the current limitations.

Following a standard design of an IIR experiment design, the two interfaces will be compared through a series of tasks that the user has to complete with the interfaces (detailed in section 5.2.2). To gather an adequate amount of information from users, each user will interact with an interface twice, completing two tasks with each interface – therefore, a total of 4

different tasks are defined which are randomized across the two interfaces. This is to accommodate the need to learn the usage of the system, as users might initially focus on getting to know the system and concentrate less on the search task. As with all IIR experiments, the task and interface assignment to participants requires randomization in order to minimize the influence of ordering effects (Kelly, 2009, p. 50). A Greco-Latin square rotation is used to generate interface-task cases to which the users are assigned randomly.

5.2.1 Participants

Due to limited available resources, the recruiting method was restrained to a convenience sampling coupled with snowball sampling. Participants were recruited via social media, using the help of friends and acquaintances. Those that are contacted have been asked to recommend further potential participants who match the recruitment criteria.

As the publications on the Progress in Mind platform are written for an academic audience, it was a criterion that the participant understands the terminology. As recruiting healthcare professionals would be difficult considering the available resources, the criteria for selection were the ongoing or already finished studies in a graduate program related to psychology, and linguistic skills in English equivalent to the CEFR level of B2 because publications on the platform are written in English. Undergraduate students (especially those in their early years of education) were not considered knowledgeable enough in the field to fully comprehend the articles published on the platform. The advantage of recruiting younger participants is that they are more likely knowledgeable with the concept of chatbots, and in case they are not, they are eager to pick up novel technologies. This means that their user experience is less prone to be determined by fundamental technical difficulties since they will know how to interact with a chatbot. Recruiting students in IIR studies is also common practice (Kelly & Sugimoto, 2013).

5.2.2 Simulated work task

Two types of search tasks are considered in IIR: tasks that concern natural information needs (when users conduct tasks that they would conduct in their everyday lives (Kelly, 2009, p. 82)) or tasks that are artificially generated by defining a simulated search task (Kelly, 2009, p. 80). In the current case, only a simulated search task is feasible, as participants have not interacted with the system before. A simulated work task is a short "cover story" of why the user decided to use the system, provides the basic context, and clarifies the source of the user's information need (Pia Borlund, 2003). It comprises the simulated work task situation and the indicative request, together forming the simulated situation. Though these situations are artificially created, they can be adapted for the target user group for enhanced authenticity.

The work tasks had to be designed in a way so that they properly clarify the problem the user has but solving them does not pose too much of a cognitive burden. Therefore, a low-intermediate level of cognitive complexity was chosen, corresponding to the cognitive process defined by Kelly et al. (2015) as *Understand*ing. According to their definition, *Understand* tasks "require the searcher to provide an exhaustive list of items" by identifying "a list or factors in an information source and possibly compile the list from multiple sources if a single list cannot be found". *Understand* tasks seem ideal since their level of cognitive complexity will not overburden the user, and they do not take too much time to complete – in Kelly et al. (2015) the average task completion time was 5 minutes for this kind of tasks.

During each task, the basic goal of the user was to list three types of diseases that have a connection to the topic of the given task. The topics of the four tasks were: sleep disturbance (which diseases have a connection to it), cognitive impairment (which diseases have a connection to it), biomarkers (for which diseases could they be beneficial) and mobile health

(in the case of which diseases could it be applied). The simulated work task descriptions can be seen in Table 3. The randomization of the tasks can be seen in Table 2.

Table 2: Task randomization table.

Part 1		**Part 2**	
Interface	***Tasks***	***Interface***	***Tasks***
Chatbot	SLEEP, COGNITIVE	Website	BIOMARKER, MHEALTH
Chatbot	COGNITIVE, BIOMARKER	Website	MHEALTH, SLEEP
Chatbot	BIOMARKER, MHEALTH	Website	SLEEP, COGNITIVE
Chatbot	MHEALTH, SLEEP	Website	COGNITIVE, BIOMARKER
Website	SLEEP, COGNITIVE	Chatbot	BIOMARKER, MHEALTH
Website	COGNITIVE, BIOMARKER	Chatbot	MHEALTH, SLEEP
Website	BIOMARKER, MHEALTH	Chatbot	SLEEP, COGNITIVE
Website	MHEALTH, SLEEP	Chatbot	COGNITIVE, BIOMARKER
Chatbot	COGNITIVE, SLEEP	Website	MHEALTH, BIOMARKER
Chatbot	BIOMARKER, COGNITIVE	Website	SLEEP, MHEALTH
Chatbot	MHEALTH, BIOMARKER	Website	COGNITIVE, SLEEP
Chatbot	SLEEP, MHEALTH	Website	BIOMARKER, COGNITIVE
Website	COGNITIVE, SLEEP	Chatbot	MHEALTH, BIOMARKER
Website	BIOMARKER, COGNITIVE	Chatbot	SLEEP, MHEALTH
Website	MHEALTH, BIOMARKER	Chatbot	COGNITIVE, SLEEP
Website	SLEEP, MHEALTH	Chatbot	BIOMARKER, COGNITIVE

5.2.3 Measurement tools

User engagement

The dependent variable in focus, the user engagement, will be measured using the User Engagement Scale (section 2.2.2) in order to gather quantitative data about users' engagement. Using this standardized questionnaire instead of a self-created one will ensure greater reliability in terms of eliciting valid measures (Hornbæk & Law, 2007; Sauro & Lewis, 2009).

Taking the research design into account, the short form of the scale is used, because, compared to the standard version, the UES-SF is more convenient to use repeatedly during longer sessions (O'Brien et al., 2018) where user fatigue would have a huge influence on IIR experiment results (Kelly, 2009, p. 52). The UES-SF questionnaire is found in Appendix A.

As the UES-SF has been developed recently, only scarce literature is available about its practical applications. The first dilemma is the range of scale: not enough values might limit the option of users for expressing their opinions (Preston & Colman, 2000), but too many categories can induce additional cognitive load. There is an ongoing debate on whether to use a 5 or 7 points Likert scale, but since students are of higher cognitive capacities, it is less problematic to use 7-points (Weijters, Cabooter, & Schillewaert, 2010). Seven seems to be the optimal limit of items we can keep in our working memory (Miller, 1994), and additionally, O'Brien validated the UES for IR using a 7-points scale too (O'Brien & Toms, 2010a) – therefore a 7-point Likert scale will be used.

The mode of administering the questionnaire can also influence the elicited measures: articulating the user engagement questions in the form of an interview may cause participants to be more critical in their evaluations (Kelly, Harper, & Landau, 2008). Nevertheless, this is of no concern, because even if participants report lower engagement (due to the form of administration), they will do so for both interfaces. Therefore, the difference between the scores for the chatbot and the website should not be influenced by the questionnaire mode.

Usability

Apart from user engagement, the usability of both systems is also assessed. The definition of the ISO standard will be used to study the various elements of usability: effectiveness, efficiency and satisfaction (Bevan et al., 2016). The task-based evaluation setting is an optimal setup for measuring usability (see section 2.3). For eliciting self-report measures (Kelly, 2009, p. 101), the task sessions are supplemented with think-aloud protocols, where the user is asked to verbalize what they are doing. As the UES already contains a subscale that pertains to usability (PU), no additional post-task questionnaire is administered for measuring usability.

Following the work of Kelly et al. (2008), participants will also be administered an exit questionnaire at the end of the experiment, which contains questions about their most positive and negative experiences regarding each of the systems – which will also be used to gain insight into the overall user experience. Preference data will also be collected here, as it is a good indicator of usability – particularly, satisfaction (Hornbæk, 2006; Nielsen & Levy, 1994).

In terms of search interaction data, task completion time is shown to be a good indicator of system usability (Sauro & Lewis, 2009), as it pertains to efficiency (Bevan et al., 2016). For effectiveness, the number of completed tasks will be taken into account.

The number of queries submitted, and search results viewed by the participant will also be considered, although they can prognose either engagement or usability. More results viewed and queries submitted might correlate to higher engagement (Edwards & Kelly, 2016) – although they can also indicate higher effort and subsequently lower engagement (O'Brien, Arguello, & Capra, 2020). In light of this ambiguity, the results will be compared to the user engagement score reported by the participants to see whether there are any correlations.

5.2.4 Research model

Figure 10 summarizes the study protocol and the sources of data used for measuring the variables. The basic demographic data that are collected at the beginning of the experiment (age, sex, and education) will not be used as independent variables but simply for describing participants. Additionally, data about how frequently participants use conversational interfaces (chatbots or voice assistants) is also collected for descriptive purposes. Think-aloud data and self-report data from the exit questionnaire both comprise direct quotes from users, and they are used to draw conclusions about usability. Behavioral measures encompass task time, task completion (whether the task was completed), the number of submitted queries, and the number of results viewed; and the latter two can correlate with either user engagement or usability.

It must be noted that IIR evaluations put a great emphasis on relevance assessment, i.e. assessing how well the retrieved result set satisfies the user's information need (Borlund & Ingwersen, 1997). In an IIR evaluation, an important step would be assigning a (usually binary) value of relevance to each document by an assessor based on the search topic (Kelly, 2009, p. 70; Pia Borlund, 2003). This way, relevance judgments by users could be compared to the "benchmark" assessments and thus evaluate IR performance. Although the object of evaluation is indeed an IIR system, the focus is on user engagement and usability and not IIR performance. Therefore, no relevance assessments are made.

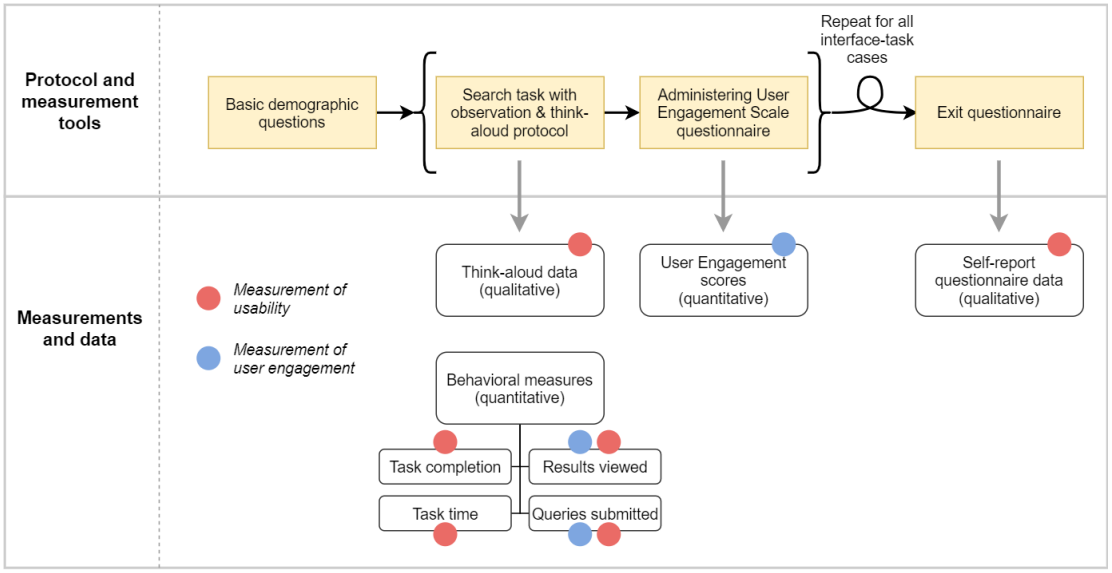

Figure 10: Research model: the measurement tools used in the study and the extracted measures.

5.3 Pilot study

To assess the approximate timeframe and the feasibility, a pilot study was conducted with two users who are not knowledgeable in the field of psychology, and one psychology practitioner who recently graduated as an MA in Psychology. The session length for the entire study was on average 50 minutes, which was deemed as an adequate timeframe. However, during the pilot study with the psychology student, the flaws in the simulated work task were discovered. The original simulated work task was constructed as follows:

You are conducting research for a school project where you explore the topic of biomarkers. You want to create a slide about biomarkers and for which diseases they can be useful. Your professor suggested using Lundbeck's database to search for information and references. Use the Chatbot of Progress in Mind to find at least 3 diseases where using biomarkers can be beneficial.

However, after exploring the site and reading some of the publications, the user articulated that they would not use the site at all for schoolwork as it is not a valid source for academic referencing (which is an accurate observation). Instead, they said that "I would check the references [at the end of the article] and I would refer to these" instead of using the article as a source. The participant also mentioned that they "already have some basic knowledge" about the topic and since "there are specific research sites for these topics" they would search within those academic repositories instead of Progress in Mind. The participant also did not feel urged to open the search results and read the article, as the SERP snippet contains the topical metadata and the query highlights, which give a clear indication about the disease the article covers. As they remarked "I can easily [see] the diseases which are connected to biomarkers" because "there are these keywords in the article highlighted here [in the] search result". Therefore, the work task had to be rewritten in a way so that it is not only authentic and does not conflict with the already established search behaviors of the participant but also prompts the participant to open the search results instead of relying solely on the information displayed in the snippet.

To alleviate the problem with the search task, the work task situation was modified in a way so that the participant has to find sources for one of their friends, instead of themselves. This

gives an explanation as to why the participant has to search within Progress in Mind rather than another, scientifically more established source. The indicative request was complemented with a prompt for the participant to point out the exact articles they would send to their friends. The aim was to motivate users to at least open the article and look through it to see if e.g. the text is not overly academic for their friend. An overview of all simulated situations can be seen in Table 3.

Table 3: Simulated work task descriptions

Task name	Simulated situation
SLEEP	You have a friend who needs help with a school project where they need to explore the causes and effects of sleep disturbance. He asks you to send him some easy-to-understand material about the topic, so you decide to use the Progress in Mind platform to search for resources. Use the Chatbot / Website to search for publications and find at least 3 diseases that may be linked to disturbed sleep, based on the publications. If you think you found a disease, point it out loud. When reading a publication, please also decide whether or not you would send it to your friend to help him with his project.
COGNITIVE	You have a friend who needs help with a school project where they need to explore the causes and effects of cognitive impairment. He asks you to send him some easy-to-understand material about the topic, so you decide to use the Progress in Mind platform to search for resources. Use the Chatbot / Website to search for publications and find at least 3 diseases that can have a connection to impaired cognitive functions, based on the publications. If you think you found a disease, point it out loud. When reading a publication, please also decide whether or not you would send it to your friend to help him with his project.
MHEALTH	You have a friend who needs help with a school project where they need to explore the possibilities of "mobile health". He asks you to send him some easy-to-understand material about the topic, so you decide to use the Progress in Mind platform to search for resources. Use the Chatbot / Website to search for publications and find at least 3 diseases, where applying the concept of mobile health can be beneficial, based on the publications. If you think you found a disease, point it out loud. When reading a publication, please also decide whether or not you would send it to your friend to help him with his project.
BIOMARKER	You have a friend who needs help with a school project where they need to explore the concept of biomarkers. He asks you to send him some easy-to-understand material about the topic, so you decide to use the Progress in Mind platform to search for resources. Use the Chatbot / Website to search for publications and find at least 3 diseases where using biomarkers can be helpful, based on the publications. If you think you found a disease, point it out loud. When reading a publication, please also decide whether or not you would send it to your friend to help him with his project.

5.4 Protocol

The experiments were conducted online due to the ongoing pandemic (the limitations of which are going to be discussed in section 8.1). During the interview, screen and audio were recorded for later analysis – of which the participants were aware, as they were asked to fill out a consent form before the sessions. The participant was invited to a virtual conference room through Google Meet and given a short introduction to the experiment (Appendix B). After this, demographic data were collected and the link to the Progress in Mind website was sent to the participant, after which they were asked to share their screen. The search tasks were read out loud to the participant, after which the researcher switched off their camera and microphone

to avoid disturbing the participant. Users could spend as much time as they needed to complete each task, although after 10 minutes they were told that they can stop if they want. A task was deemed complete once the user named a third disease that is relevant to the topic. After each task completion, users were administered the UES-SF in an interview form, where each statement was read out loud. At the end of the experiment, the exit questionnaire items were also read out loud (see Appendix C).

6 Analysis

A total of 10 students from Eötvös Loránd University participated in the study, 8 females and 2 males. Their age ranged from 22 to 32 with a mean age of 24,5 years (SD=3). Almost all participants reported that they never used chatbots or only once or twice in their life; one participant used chatbots more than once a month for flight booking and online shopping assistance. Similarly, most participants never used voice assistants either; only two participants used them in the past, and both of them reported having bad experiences with them due to their lack of utility. All had Hungarian as their native language. On average, an entire experiment was 60 minutes long with a minimum length of 39 minutes and a maximum of 73 minutes.

Following the instructions by O'Brien et al. (2018), the questionnaire items for perceived usability (PU) were reverse coded. For each participant, the subscale scores for the chatbot were calculated by taking the two tasks which the participant completed using the chatbot, and taking the average of all scores within a subscale (RW, PU, FA, AE) – the same procedure was done for the website interface. For obtaining the final UES scores for the interface, the subscale-scores for the interface were averaged. Finally, a grand mean UES score for both interfaces was calculated by averaging the final UES scores of each participant. A Shapiro-Wilks test revealed that both the chatbot scores and the website scores follow a normal distribution (p = .865 and p = .428 respectively), and an F-test showed that the datasets for the two interfaces had a homogeneity of variance (p = .853), therefore a parametric test could be applied to test for any significant difference.

Behavioral measures were also analyzed for each participant. Task times are positively skewed because there is a minimum amount of time users need to complete the task. Therefore, instead of the arithmetic mean, the geometric mean was used, as it is more representative of the central tendency for the whole population, especially at smaller sample sizes (Sauro & Lewis, 2010). Task time denotes the time users spent on the task until they either successfully completed it or gave up (the latter occurred only in three cases). The number of submitted queries includes both full-sentence or keyword-based queries, and also queries which contain a misspelling, but not small-talk inputs like greeting the chatbot (of which there were only two cases). The number of results viewed incorporates only those results which the user clicked on (either in the SUI result list or via a link within an article) and viewed. If a participant clicked on a result and opened it in the background but did not look at it, it was not counted (because its content did not contribute to the participant's information evaluation process). The geometric average of task completion times and the average number of submitted queries and results viewed were calculated for each interface. A summary table of the behavioral measures and UES scores can be found in Appendix D.

A thematic analysis was conducted for analyzing the qualitative data provided by the think-aloud data and the exit questionnaire data. The recommendations by Braun & Clarke (2006) were followed: recordings were reviewed and utterances of interest were transcribed as user quotes – which were turned into codes. The main themes according to which quotes were collected were usability (any remarks about the usability of the system) and search behavior (any remark about how the participant searches and why). The codes were collated into sub-themes based on their topical similarity, the sub-themes were named and assigned to the main themes. Further partitioning of the codes was done where it was deemed necessary. In some cases, the quotes were supplemented with observational descriptions to clarify the context.

7 Results

7.1 Quantitative results

The analyses focused on understanding how the independent variable, the search interface, influenced user engagement, the dependent variable. According to Figure 11, most UES scores are from the upper half of the scale, which suggests that the majority of participants were engaged throughout the study. Comparing the average UES scores of the chatbot (M = 4.65, SD = 1.05) and the website's (M = 4.83, SD = 1.12), no significance difference was found between the two, $t(9) = -0.4$, $p = .69$. Moreover, the mean UES score of the chatbot was slightly lower than the website's. As we did not find substantial evidence to say that the usage of the chatbot results in higher user engagement, we do not reject the null hypothesis (H_0).

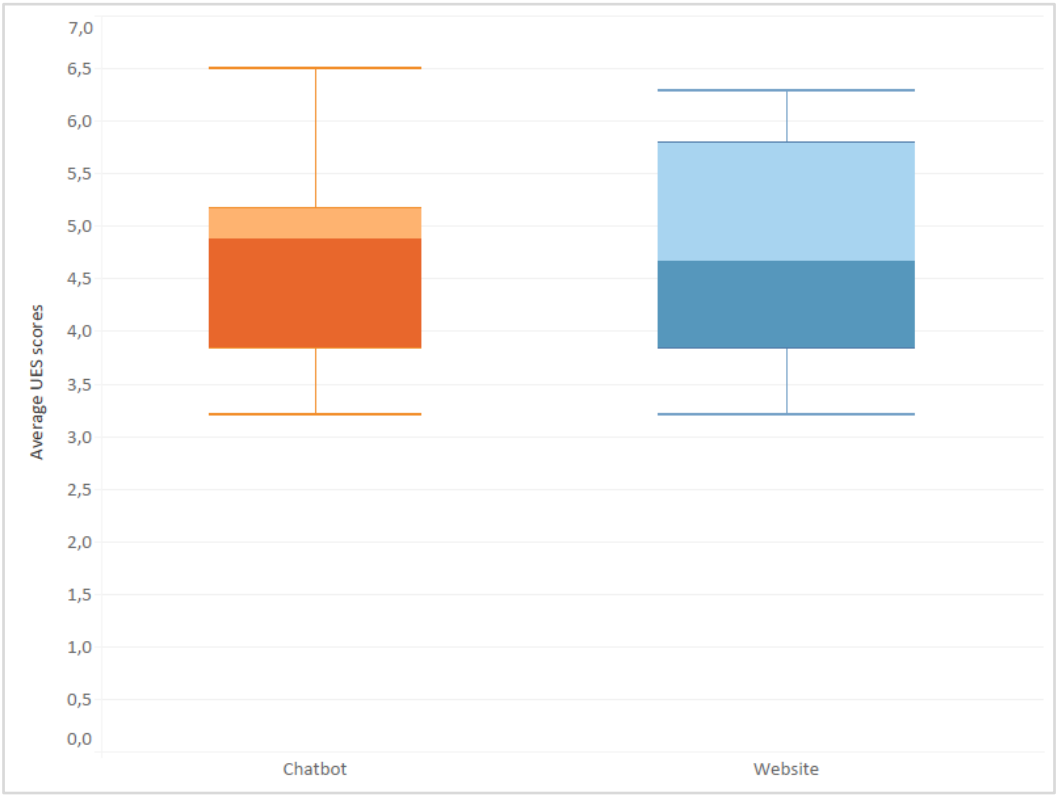

Figure 11: Distribution of the average UES scores for the chatbot and the website SUI. The darker shade represents data between the 25th and 50th percentile, the lighter shade, data between the 50th and 75th percentile.

In order to investigate whether the type of interface influenced any specific aspect of engagement, the UES scores have been broken down to subscales. Figure 12 shows the mean subscale scores per interface. No significant difference was found between the two interfaces in terms of subscale scores. The website SUI outperformed the chatbot interface in all but one aspect: Aesthetic Appeal (AE), where the chatbot got a 0.2 points higher score. The largest difference between the two interfaces can be observed in Perceived Usability (PU), where the website (M_{PU} = 5.6, SD = 1.19) outperformed the chatbot (M_{PU} = 5, SD = 1.34) by more than half a point. Both interfaces received a relatively low score for focused attention (FA). Reward

(RW) received the second-highest scores after PU, with only a slight difference between the two interfaces.

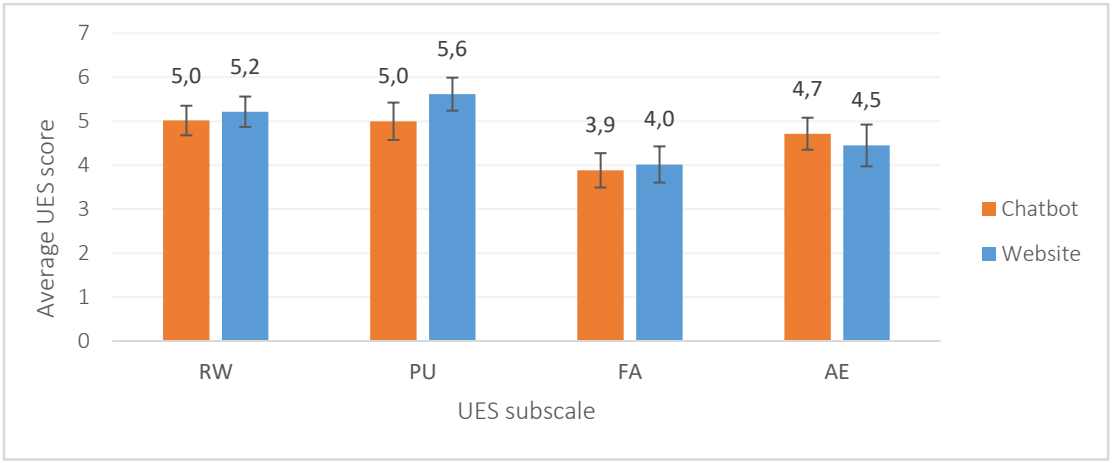

Figure 12: Average UES subscale scores for the chatbot and the website SUI.

Behavioral measures were analyzed (Figure 13) to see how the two interfaces compare in terms of task performance. On average, participants took more time (approximately 2 more minutes) to complete the tasks with the chatbot compared to the website, which indicates a lower efficiency within the chatbot. Users submitted on average almost 1.5 times more queries when using the chatbot. The number of viewed results was approximately equal across the two interfaces, with chatbot users viewing slightly more results than website users. No significant difference was found between the two interfaces in terms of any behavioral measures.

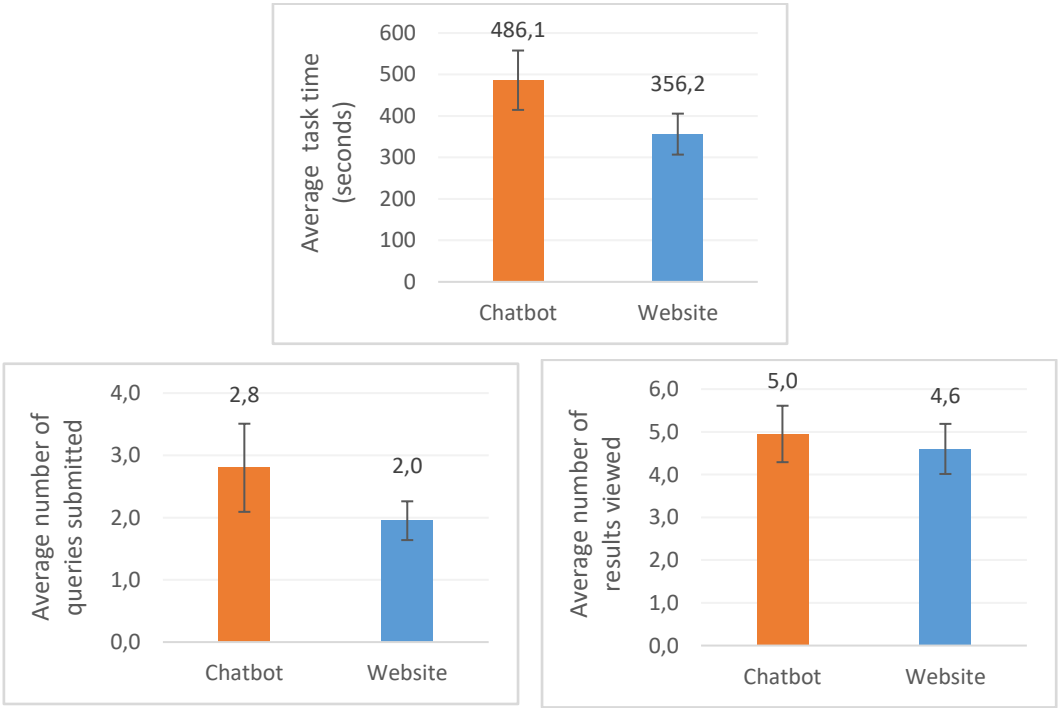

Figure 13: Summary of behavioral measures per interface: task time (**top**), number of queries submitted (**left**), and number of results viewed (**right**).

Preference data showed that users favor the website more: 9 participants reported that they would use the website for searching across the collection, and only one participant said that they would use the chatbot. This means that user satisfaction was higher while using the website than while using the chatbot.

In terms of task completion, there were only three instances where the user was not able to successfully complete the task, each time while using the chatbot. Therefore, while all participants managed to complete each task with the website SUI successfully, the usage of the chatbot led to unsuccessful task completions three times – which indicates a slightly lower effectiveness in the latter.

In order to see whether there is any correlation between behavioral measures and user engagement, the Pearson-product moment correlations have been calculated between the behavioral measures and the UES scores (Table 4). Correlations are almost exclusively negative, apart from 3 cases, which means that task time, number of submitted queries and number of results viewed elicit lower UES scores. Significant correlations have been found between website RW score and task time, website RW score and number of results viewed, and chatbot FA score and the number of results viewed. Behavioral metrics have overall stronger correlations with UES scores in the case of the website, and weaker correlations in the case of the chatbot.

Table 4: Pearson-product moment correlations between behavioral metrics and the grand mean UES subscale scores and the grand mean UES score per interface.

	Chatbot			**Website**		
	Geometric mean of task time	Avg. number of queries submitted	Avg. number of results viewed	Geometric mean of task time	Avg. number of queries submitted	Avg. number of results viewed
RW	-.077	.002	-.529	-.659*	-.333	-.717*
PU	-.469	-.323	-.571	-.464	-.248	-.346
FA	-.262	-.156	-.658*	-.485	-.035	-.546
AE	.144	.292	-.218	-.575	-.277	-.384
Final UES	-.206	-.069	-.568	-.616	-.249	-.552

Symbol "*" denotes significant r values ($p < .05$). Coloring represent the strength of the correlation, where white color represents very weak correlation ($|r| < .19$), and the darkening orange colors represents weak ($.2 < |r| < .39$), moderate ($.4 < |r| < .59$), and strong ($|r| > .6$) correlation.

7.2 Themes

This section details the themes that emerged from the qualitative data, organized around the two encompassing themes of usability and search behavior. Table 5 summarizes the themes and sub-themes that emerged from the think-aloud data and the exit interview. The detailed version of the thematic analysis diagram with user quotes can be seen in Appendix E.

Table 5: Thematic analysis table

Theme	Sub-theme	Code
Usability	User interface	Overlapping UI elements
		Interface aesthetics (visuals and sound)
		Ambiguity of UI elements
	Inconsistencies within the system	Newsletter among the results
		Website SUI's inconsistency with thumbnail images
	Chatbot utility	Uncertainty about when to use the chatbot
		Chatbot fails to handle complex input
		Doubtful disposition towards chatbots
	Importance of overviewable results	Viewing more results at once is important
		Chatbot window is too small
	-	Importance of response speed
Search behavior	Assessing relevance of documents	Relying on field knowledge
		Assessing results and content using metadata and keywords (within SERP and within content)
	Partitioning and query tactics	Filtering and faceting features are missing (in the SUIs and in the homepage)
		Phrase search is useful for experienced searchers
		Preferring keyword-search in chatbot
	-	Finding related items is difficult
	-	Lack of search transparency within the system
	-	Searching for explanations in case of unfamiliar topics

Usability

The majority of remarks about usability concerned the visual structure of the user interface, highlighting the differences between the SUI of the website and the chatbot. The main issue seemed to be the relatively **small size of the chatbot window**, which caused all navigational elements to be placed closely together. One user remarked how “the [chat] window was too

small, and you had to click this small right-arrow which was annoying", reflecting on the difficulty to navigate between results. This issue was particularly frustrating for three participants who used touchpad input, as each one remarked about their frustration with the small chatbot window. One touchpad user mentioned that "it was kind of tiring to click between the results [in the chatbot]. You have this small arrow and immediately next to it the scrollbar, plus the article. So, I can click on three things if I'm not accurate".

Apart from navigational problems, the small chatbot window also hindered the relevance assessment of results. Almost all users noted the inadequacy of the chatbot to **display several results at once**, which, as one participant noted, "was weird because I couldn't have as much of an overview". In contrast, "the website was better because it showed the results below one another and I could overview them more easily". The constrained display window of the chatbot also limited the users in their search flow, as they "had to check the results in sequence, because I was afraid that I'll get lost". Another participant noted the same problem of losing track of the results, saying that they would prefer a full-screen view, so that "I would still be able to remember what the first result was even when I'm at the 6th". The website SUI enables the user to overview multiple results at once on the screen, which, as one participant commented, "made it easier for me to choose between them", whereas the chatbot was found lacking from this aspect. Only one participant noted that "I like that [the chatbot] displayed the results horizontally and I didn't have to scroll up and down", but the same participant commented later in retrospect that "you can overview a lot of results on the webpage search whereas in the chatbot you can only see one at once".

Due to a fault in the front-end code, some visual elements of the website (the navigation button and loading bar of the hero image, see Appendix F) are **overlapping the chatbot window**, which confused the participants on several occasions, as it would seem like there are two navigational arrows present for navigating between the search results. Users would sometimes click on the "wrong" arrow and some even remarked that "it is a bit annoying that you have these two arrows on top of one another". One participant who worked in product ergonomics pointed out further poorly designed and **ambiguous visual cues**: the "*send* button does not really look like a button", and that "the whitespace [at the bottom of the conversation window] and the text field look the same visually, and I was accidentally clicking on the whitespace".

Another negative aspect of the chatbot was the slow **response time**, as "it was slow compared to the webpage [...] that 3 seconds waiting was strange. It took some time to react". In contrast, the website was "faster [compared to the chatbot] and I didn't have to wait for an answer", as one participant remarked.

Users had an overall negative affect towards the **visual style** of the platform, as some pointed out that it "was quite dull and colorless. Those brown and grey colors weren't very attractive". Interestingly, one participant remarked positively about the poor visual appeal of the website, as "for some reason, it's important to me that if something [deals with] scientific [material], it should look a bit lame. This website looked trustworthy for me [...] because if this". Apart from visual aesthetics, the **sound** the chatbot made every time it sent a message was found to be annoying by some participants. Two participants said that the sound "irritates" them and that it was "weird and annoying", whereas one participant said they liked it.

Participants also made negative comments about the chatbot's utility. Users who tried to communicate with the chatbot with complete sentences realized that the **chatbot is not capable of handling complex inputs**, therefore all resorted to keyword inputs (discussed in the next theme). Interestingly, one participant who sent numerous full-sentence queries to the chatbot, reported that "I felt like [the chatbot] handles full sentences better", whereas this was not the

case at all[3] – the chatbot always showed irrelevant results or no results at all to these queries. Most participants were **uncertain about the actual use case for the chatbot**. One subject stated that "I'd probably use the chatbot to find random articles I might not find otherwise. But for targeted search, I wouldn't use it", highlighting the use case for exploratory search, whereas another participant remarked that "If I'd want to look up something [...] I would use the chatbot, but If I needed to e.g. write an essay, I would use the website because it would give me a broader overview", which highlights the use case of targeted lookup within the chatbot. These conflicting opinions show that users have a hard time pointing out the best way to utilize the chatbot. Some participants were even more **skeptical about the chatbot** and articulated that they "would not even think about using the chatbot for searching". One reason behind the doubtful disposition was the lack of faith in conversational technology: one participant told about their negative experience with voice assistants and "which made it clear that I don't want to use them again", and another reflected that "using [chatbots] only makes sense if you're talking to a real human". One participant even highlighted the overall futility of conversational search, saying "I don't think that searching needs to be innovated from an interface aspect".

Search behavior

Throughout their search, users demonstrated various tactics to choose which results to click and to assess whether the content they are reading is relevant for them or not. Most of the users could be observed **scanning the metadata in the snippets** for relevant keywords using either the title, topical tags, or query highlights – which was a fairly limited tactic in the chatbot. Users were missing the rich metadata from the chatbot snippets, because in the website SUI "you could see above the titles the diseases the article is about [...] whereas the chatbot does not display these keywords". In the chatbot, users could only use the title and the first few sentences of the content to assess the relevance of the result, but in some fortunate cases they "could see it [that the article is relevant] from the description". Half of the participants started immediately reading the result they clicked, the other half tended to open several results at once in the background then returned to read them one by one. Interestingly, two participants mentioned that "some of [the results] has an image, some of them hasn't" and that they "automatically undervalued those" which do not have one.

Users commented positively on the **easily scannable structure of the content**, as one participant said the "articles are structured neatly, and they are easy to overview". Abstracts were a main source of information for assessing relevance, so they "could already see from the first sentence that it's a relevant article". Instead of reading the entire article, most of them "only searched for keywords" – 5 participants were even observed using the *Find on page* tool of their browser to search for keywords of interest.

Apart from relying on keywords, another tactic of assessing relevance was **relying on field knowledge**. Users would collate what the content stated with their own knowledge to assess the relevance of the result, e.g. "this one shows major depressive disorder, which makes sense as that [and cognitive impairment] go hand in hand". Some participants expressed a certain level of confusion when they met with information which seemingly contradicted their former knowledge, with comments like "based on what the article states, I would not connect it [to the disease] but I know [by myself] that they are related" or "I was surprised that the system does not give me results related to anxiety, as it would be logical for me. Because I heard about a couple [mobile] applications that deal with this topic".

[3] The reason behind this behavior could be that the participant knew which system is being tested, and "reactivity suggests that subjects might react more favorably to the system they presume the researcher is really studying" (Kelly & Sugimoto, 2013)

More than half of the subjects expressed their frustration that the system does not provide adequate search functionality. The biggest issue was the **lack of filtering and faceting options** as there are “no options in the search bar to filter, like which source is it from, [or] when was it published”. One participant remarked positively that the chatbot provides at least some level of categorical browsing, saying that “I really liked at the beginning that it asked whether I want to read articles or listen to podcast”. However, users had no knowledge about how to conduct faceted searching within the chatbot (detailed in section 3.2.2) since it is not a straightforward process – therefore they were still missing the option to “choose exactly what [type of publication] I'm searching for”. Though the homepage displayed the facets explicitly, users found little use for them as “the website does not handle them very intuitively”, and some said they “don't know which direction I should go”. They also found it “unfortunate” that they could not search within these facets – only one participant searched in the Expert views facet but could not find anything relevant and abandoned the facets. Regarding topical search, only one of the participants clicked on the tags that are attached to the articles

In terms of query formulations, users almost exclusively resorted to **keyword-based search**, usually using the task topic as the query (e.g. *cognitive impairment*) – even within the chatbot, despite its conversational interaction. One reason behind this could be that, as one participant stated, “the chatbot phrased the question in a way that it didn't even occur to me to reply in full sentences”. The chatbot phrased its welcome message as “type what you are looking for”, which the users might have interpreted as a prompt for a keyword or search phrase. Nevertheless, users liked that “it was enough to write keywords and you got all types of publications”.

In the website, only four participants could be observed using **phrase search** (without knowing that the system is capable of that) to search for exact query matches and they all had positive opinions about it, although one participant commented that “they could have stated this somewhere that you can do this”.

Three participants raised the issue of **search transparency** – saying that “it wasn't really clear to me how [the chatbot] selects those articles”. One participant even remarked about this distrust, saying that “I had a bit of distrust in me about whether [the platform] actually shows me the relevant results”. This issue was even more relevant in the chatbot, where only a limited number of results were displayed. Users “didn't really know how to expand the number of results”, and one user mentioned that they were curious how those articles were selected, as they “couldn’t really see any pattern in it”.

8 Discussion

The results of the study will be discussed in accordance with the main research question of how the interface type influences user engagement (RQ) and the research sub-question of how the interface type influences usability (RQ-s).

How does the interface type influence user engagement?

The analysis revealed that using the chatbot for searching does not lead to greater engagement – the null hypothesis could not be rejected. In fact, the chatbot underperformed in all but one aspect of engagement, aesthetic appeal (AE). According to the thematic analysis, the aesthetics of the interface seems to be of a subjective matter, as a stylistic choice can elicit both negative and positive reactions from participants. Therefore, the reason behind the higher AE score may be attributed simply to the novelty of the interface – the chatbot might have grabbed the users' visual attention because of its unique way of searching, which might have resulted in an initial interest and a more favorable AE score. Still, the attractiveness of the interface was not enough to counterbalance the other aspects the interface was lacking – especially perceived usability (PU), which is going to be discussed separately.

Interestingly, both interfaces received a relatively low score for focused attention (FA), which suggests that neither interface managed to hold the attention of the participants to such an extent which could have led to deep involvement. The reason behind the low scores could be that the protocol of the experiment gave little room for substantial immersion: the *Understand* type tasks we used did not require high-level cognitive processing, only identifying and compiling information (Kelly et al., 2015). The online format of the experiment might have also played a role, as participants' focus could be easily disrupted by their external environment – which prevented them from being absorbed in the experience. The chatbot's slightly lower FA score could be due to its slow response time, which participants occasionally commented about. Participants might find it self-evident that search systems are generally quick to respond (like Google), thus the chatbot with such a response delay (approximately 2 seconds) may seem sluggish and it can interrupt the user's flow of thoughts (Nielsen, 1993a).

Reward (RW) received the second-highest average score among the subscales with only a small score difference between the two interfaces, which indicates that participants usually found their search experience interesting, worthwhile, and rewarding – regardless of the platform. This highlights the importance of the content, which – interface-independently – enhanced the reward factor of using the platform. Observations also reinforce this assumption, as many participants made sporadic comments about the platform's interesting content (e.g. "the content is [extremely] good…the articles were great and contained relevant information").

Regarding the behavioral measures, participants were less efficient in their tasks when using the chatbot, with higher task times, more queries sent, and more results clicked. The almost exclusively negative correlations between the UES scores and the behavioral measures also show that a higher "interaction cost" leads to lower engagement. This is in accordance with O'Brien et al. (2020) results, who found that a higher task effort correlates negatively with engagement. Edwards & Kelly (2017) also found that "increased search behaviors" signify frustration, rather than engagement. The number of results viewed seems to be a good indicator of low engagement. Participants who clicked on a large number of results might have experienced impatience and frustration, which led to lower engagement. In the case of the website, task time seems to be a good indicator of low engagement, with correlations ranging from moderate to high. However, in the case of the chatbot, only PU had a moderate negative correlation with task time – which resonates with the findings of Sauro & Lewis (2009) that

higher task times indicate poor usability. The number of query submissions was not shown to be good indicators of engagement as correlations were either weak or very weak.

Interestingly, the correlations were generally stronger in the case of the website and weaker in the case of the chatbot (Table 4), which implies that behavioral measures may not be as reliable in predicting users' engagement with chatbots.

How does the interface type influence usability?

Quantitative data did not show any significant differences between behavioral measures. Nevertheless, the chatbot elicited overall higher task times (pertaining to efficiency), the preference data (pertaining to satisfaction) shows a higher satisfaction in the case of the website, and task success (pertaining to effectiveness) also shows that users were slightly more successful in completing their tasks with the website. Considering all three aspects of usability, the website performed better compared to the chatbot.

The thematic analysis also shows that participants found the chatbot less usable than the website SUI, from multiple aspects. The greatest problem of the chatbot is the limited amount of information it displays due to its small size. Though a horizontally scrollable result list needs less effort to navigate through compared to vertical scrolling, in exchange of displaying the results in a compact area the overviewability is greatly impaired – and since this issue was mentioned by almost all users, it might be the greatest contributor to the reduced PU score of the system. The problem of overviewability ties closely to Shneiderman's Visual Information Seeking Mantra, which stipulates that a system must first provide the user a proper overview of the collection, before zooming in on items of interest and providing details on demand (Shneiderman, 2007). The chatbot violated this mantra as users could only see one result at a time.

Results in the chatbot also omitted certain metadata, which made assessing their relevance even more difficult. The lack of metadata and sorting functions also made users question how the system ranks the results. Jackson et al. (2016) also raised the issue of search transparency, stating that a search system should state according to which criteria the results are ranked, otherwise users become "instinctively distrustful of any mechanism they don't understand" – which is reflected in participants' comments.

Lack of filtering and faceting also impaired search efficiency for both interfaces. Although the chatbot does provide faceted browsing to some extent, accessing it is not straightforward, and none of the users managed to figure out how to search within facets. Topical tags are also accessible for each article, but they are not integrated enough in the search system and not salient enough so that users could find them easily. The possibility of issuing phrase-search or using search operators was also not communicated effectively. The system seems to provide more utility to experienced searchers who are already familiar with the platform and who can leverage the system's less visible functionalities (e.g. search operators or tags).

Further indicators of the chatbot's poor usability are the higher task times (Sauro & Lewis, 2009), and the lower PU score of the chatbot (Figure 12). The higher number of submitted queries and viewed results may also indicate lower search efficiency, as participants had a harder time finding relevant results with the chatbot. However, it must be noted that the chatbot displays only a limited number of items on the SERP, thus users had to submit further queries if they wanted to see more results. This could be another reason behind the large difference in the number of query submissions. It must also be noted that, since the chatbot omits certain metadata and thus makes relevance assessment difficult, users might have been more inclined to open the result and check the content itself to determine its relevance – hence the higher number of viewed results. Nevertheless, the behavioral measures also show that the chatbot required greater interactional effort from the participants, which translated into poor usability.

Besides, the user comments and the almost-exclusive preference for the website SUI show that users can hardly recognize any value the chatbot could add to their search process. For example, a participant commented that "a chatbot can create added value where a human person, a social interaction with a human needs to be substituted… and searching is not a social interaction". This signifies that the chatbot performs poorly not only in terms of usability but also utility – and highlights the fact that, despite the promising results in some areas, chatbots have still a long way to go until they become useful for information retrieval.

8.1 Limitations of the study

The global pandemic of COVID-19 crippled several aspects of the experiment, especially the amount of control one had over the setting of the study. The primary hindrance was the safety regulations that restrained personal interactions to a minimum – thus, personal one-to-one interviews had to be ruled out. Though conducting the experiments online was a convenient alternative, it did not provide the same amount of control as if the study was conducted in a laboratory setting. Apart from that, sampling of participants became even more troublesome, as people were more difficult to prompt to participate in the study due to the economic uncertainty and the general unrest in the society. Not to mention that the targeted user group were students, who were also preoccupied with their exams during the time of the study.

The importance of control becomes even more apparent if we look at how many factors can influence the outcome of an IIR study. White & Roth (2009) state that the success of information retrieval tasks is dependent on:

1. The user's mental model of the system's features (determined by the GUI);
2. The user's knowledge of the task domain;
3. The user's information seeking experience;
4. The user's physical setting.

Regarding the mental model (1) or the information-seeking experience (3) of the user, the thesis did not focus on these factors, therefore a certain level of fluctuation among the participants is inevitable. The user's knowledge about the domain (2) was taken into account during the sampling: participants were all knowledgeable in the field to at least some extent, although minor fluctuations may still be present e.g. due to different specialization courses or personal interests. However, the user's interest in the task topic, which is closely tied to domain knowledge, could have had an influence on engagement. O'Brien et al. (2020) showed that different task topics will result in different levels of engagement, and a greater interest in the topic results in a higher user engagement score. Although the UES-SF does contain a related item ("I felt interested in this experience"), it does not specifically pertain to interest in the topic, but rather as the whole experience. Therefore, there is no information about how topic interest influenced user engagement.

The physical setting of the experiment (4) was not possible to control due to the online presence. A range of factors could have influenced not only users' search behavior efficiency but also their user experience and subsequently their self-reported measures. Disturbances that were caused by some of the participant's environment (e.g. someone ringing their doorbell or talking to them while solving a task) may have reduced task efficiency or hindered engagement, and participants might have even evaluated their experiences more negatively because of that (hence the low FA scores). Even the type of device the participant is using to access the website might influence perceived usability, as participants who used a touchpad reported difficulties

with interaction – an issue which is detailed under the code "Chatbot window too small" in section 7.2. The need to think out loud might have also influenced the search process; one participant even mentioned that "I cannot really become engrossed in reading an article while knowing that you hear what I'm thinking about", implying that the task of verbalizing thoughts could have also hindered focused attention and led to lower FA scores. Lastly, since the user engagement questionnaire items were asked in an interview form, participants had the additional cognitive burden of envisioning the scale and remembering what the numbers on the UES represent. One participant even asked once during the questionnaire administration what the two ends of the scale represent. In contrast, during a physical interview participants could have been presented with a printed version of the scale with clear labeling and they could have pointed at the specific point on the scale – similar to how Kelly et al. (2008) conducted their study.

The generalizability of the findings must also be discussed. The convenience sampling used for gathering participants undoubtedly limits generalizability, since the entire sample consists of participants from Hungary. From the aspect of language skills, the Hungarian population is underperforming compared to other European countries[4]. The overall lack of language skills might have impaired the task efficiency of the participants, and although each participant possessed – according to their report – language skills equal to level B2, the general lack of language practice could still have influenced their performance. This shows in the observation that, on three occasions, participants asked whether they are writing the keywords right, and some of them expressed not understanding what certain acronyms within the articles mean (e.g. MDD for major depressive disorder, PD for Parkinson's disease). Apart from the language, the sample mainly consisted of female participants, which was possibly due to the snowball sampling method, as the first few participants recruited were female. The lack of gender diversity in the sample must be taken into account because gender can have significant effects on search behavior (Roy & Chi, 2003) and search efficacy (Zhou, 2014).

Choosing Progress in Mind as the object of analysis might also seem counterintuitive from the perspective of generalizability, as it is evident that the interface studied in this thesis is highly specific for a medical domain. It is arguable that for studying the user engagement of conversational search interfaces in a broader context, one should opt to use a more general-purpose chatbot with more "everyday" content (e.g. a news site or a library). The reason behind the choice for Progress in Mind was the immediate availability: both the chatbot and its GUI counterpart (the website) is easily accessible without having to create them from scratch, both of them is used for searching for content, and both of them taps into the same database – which means that they are ideal for comparison. Procuring a database, setting up a search engine, and creating both a graphical and a conversational SUI and on top of the system would have provided more options for tailoring the system to the study, but it would not have been feasible considering the limited time and resources.

[4] According to the statistics by Eurostat, 40% of young adults in Hungary do not speak any foreign language, which is twice as much as the European average.

9 Conclusion

A website-based graphical SUI and a chatbot-based conversational SUI of a resource center have been compared in terms of user engagement and usability. No evidence was found that the usage of the chatbot would lead to a higher level of engagement. On the contrary, the website interface outperformed the chatbot in all aspects of engagement except aesthetic appeal – the latter supposedly due to the chatbot's novel way of displaying information. Higher interactional efforts during chatbot use led to poor usability, which is believed to be a primary contributor to reduced engagement. Subjects also made more negative comments about the usability of the chatbot than the website SUI – which is in line with the preference data that is almost exclusively in favor of the website. The chatbot's limited overviewability, small display size, and the lack of information displayed in the SERP hindered efficient searching, and participants saw little value in using the chatbot instead of the website.

As the document collection was quite field-specific, further research should investigate the differences between graphical and conversational SUIs for more generic collections, like library books or music collections. Specific attention should be given that the chatbot SUI displays the same SERP information (i.e. metadata) and provides the same search features like the graphical SUI, so that differences between functionality do not influence user engagement – only the way the information is communicated.

10 Bibliography

Allen, J. F. (1987). Introduction to Natural Language Understanding. In *Natural Language Understanding* (pp. 1–17). Rochester, New York: The Benjamin/Cummings Publishing Company.

Allison, D. (2012). Chatbots in the library: Is it time? *Library Hi Tech*, *30*(1), 95–107. https://doi.org/10.1108/07378831211213238

Bassett, C. (2018). The computational therapeutic: exploring Weizenbaum's ELIZA as a history of the present. *AI and Society*, *34*(4), 1–10. https://doi.org/10.1007/s00146-018-0825-9

Bates, M. J. (1979). Information Search Tactics. *Journal of the American Society for Information Science*, *30*(4), 1–10. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/asi.4630300406/full%0Afile:///Files/3D/3D3EA76B-6809-4438-AB84-B095E5E73F73.pdf

Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review*, *13*(5), 407–424. https://doi.org/10.1108/eb024320

Baumann, N. (2016). How to use the medical subject headings (MeSH). *International Journal of Clinical Practice*, *70*(2), 171–174. https://doi.org/10.1111/ijcp.12767

Beckers, T. (2009). Supporting Polyrepresentation and Information Seeking Strategies. In *FDIA 2009: Symposium on Future Directions in Information Access* (pp. 56–61).

Beckers, T., & Fuhr, N. (2012). Towards the Systematic Design of IR Systems Supporting Complex Search Tasks. In *Task Based and Aggregated Search (TBAS 2012)* (pp. 15–19).

Belkin, N. J. (2010). On the Evaluation of Interactive Information Retrieval Systems. In B. Larsen, J. W. Schneider, F. Åström, & B. Schlemmer (Eds.), *The Janus Faced Scholar: A Festschrift in Honour of Peter Ingwersen* (pp. 13–21). Copenhagen: Det Informationsvidenskabelige Akademi.

Bevan, N., Carter, J., Earthy, J., Geis, T., & Harker, S. (2016). New ISO Standards for Usability, Usability Reports and Usability Measures. In *In International Conference on Human-Computer Interaction* (pp. 268–278). Cham: SAGE Publications. https://doi.org/10.1007/978-3-319-39510-4

Bickmore, T., & Cassell, J. (2001). Relational agents: A model and implementation of building user trust. *Conference on Human Factors in Computing Systems - Proceedings*, (3), 396–403.

Bohus, D., & Rudnicky, A. I. (2005). A principled approach for rejection threshold optimization in spoken dialog systems. *9th European Conference on Speech Communication and Technology*, 2781–2784.

Boren, M. T., & Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication*, *43*(3), 261–278. https://doi.org/10.1109/47.867942

Borlund, P., & Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, *53*(3), 225–250. https://doi.org/10.1108/EUM0000000007198

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Bryman, A. (2016). Research designs. In *Social research methods* (4th ed., pp. 45–77). Oxford University Press.

Cassell, J. (2000). Embodied conversational interface agents. *Communications of the ACM*, *43*(4), 70–78. https://doi.org/10.1145/332051.332075

Celino, I., & Re Calegari, G. (2020). Submitting surveys via a conversational interface: An evaluation of user acceptance and approach effectiveness. *International Journal of Human Computer Studies*, *139*, 1–16. https://doi.org/10.1016/j.ijhcs.2020.102410

Chung, M., Ko, E., Joung, H., & Kim, S. J. (2018). Chatbot e-service and customer satisfaction regarding luxury brands. *Journal of Business Research*, 1–9. https://doi.org/10.1016/j.jbusres.2018.10.004

Csikszentmihalyi, M. (1985). A Theoretical Model for Enjoyment. In *Beyond Boredom and Anxiety: The Experience of Play in Work and Games* (pp. 35–54). London: Jossey-Bass.

D'Alfonso, S., Santesteban-Echarri, O., Rice, S., Wadley, G., Lederman, R., Miles, C., … Alvarez-Jimenez, M. (2017). Artificial intelligence-assisted online social therapy for youth mental health. *Frontiers in Psychology*, *8*, 1–13. https://doi.org/10.3389/fpsyg.2017.00796

Dale, R. (2016). The return of the chatbots. *Natural Language Engineering*, *22*(5), 811–817. https://doi.org/10.1017/S1351324916000243

Denecke, K., Hochreutener, S. L., Pöpel, A., & May, R. (2018). Self-Anamnesis with a Conversational User Interface: Concept and Usability Study. *Methods of Information in Medicine*, *57*(5–6), 243–252. https://doi.org/10.1055/s-0038-1675822

DeVellis, R. F. (2016). Understanding the Latent Variable. In L. Bickman & D. J. Rog (Eds.), *Scale Development Theory and Applications* (4th ed., pp. 36–48). Los Angeles: Sage Publications.

Dominich, S. (2008). Basics of Information Retrieval Technology. In *The Modern Algebra of Information Retrieval* (pp. 65–104). Berlin: Springer Verlag.

Dubiel, M. (2018). Towards Human-Like Conversational Search Systems. In *CHIIR'18: 2018 Conference on Human Information Interaction & Retrieval* (Vol. 2018-March, pp. 348–350). New York, New York, USA: ACM Press. https://doi.org/10.1145/3176349.3176360

Dubiel, M., Halvey, M., Azzopardi, L., & Daronnat, S. (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* (pp. 1–8). ACM Press.

Edwards, A., & Kelly, D. (2016). Engagement in Information Search. In *Why Engagement Matters* (pp. 157–176). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-27446-1_7

Edwards, A., & Kelly, D. (2017). Engaged or frustrated? disambiguating emotional state in search. *SIGIR 2017 - Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 125–134. https://doi.org/10.1145/3077136.3080818

Ellis, D. (1989). A behavioural approach to information retrieval system design. *Journal of Documentation*, *45*(3), 171–212. Retrieved from http://www.emeraldinsight.com/doi/10.1108/eb026404

Exalto, M., De Jong, M., De Koning, T., Groothuis, A., & Ravesteijn, P. (2018). Conversational commerce, the conversation of tomorrow. In *Proceedings of the 14th European Conference on Management, Leadership and Governance, ECMLG 2018* (pp. 76–83).

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health*, *4*(2), 1–11. https://doi.org/10.2196/mental.7785

Flach, J. (2011). A Framework for Ecological Interface Design (EID). In *Display and Interface Design* (pp. 109–140). CRC Press. https://doi.org/10.1201/b10774-7

Fryer, L., Ainley, M., Thompson, A., Gibson, A., & Sherlock, Z. (2017). Stimulating and sustaining interest in a language course: An experimental comparison of Chatbot and Human task partners. *Computers in Human Behavior*, *75*, 461–468. https://doi.org/10.1016/j.chb.2017.05.045

Fryer, L., Nakao, K., & Thompson, A. (2019). Chatbot learning partners: Connecting learning experiences, interest and competence. *Computers in Human Behavior*, *93*(December 2018), 279–289. https://doi.org/10.1016/j.chb.2018.12.023

Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., & Rauws, M. (2018). Using Psychological Artificial Intelligence (Tess) to Relieve Symptoms of Depression and Anxiety: Randomized Controlled Trial. *JMIR Mental Health*, *5*(4), 1–15. https://doi.org/10.2196/mental.9782

Graesser, A. C., Li, H., & Forsyth, C. (2014). Learning by Communicating in Natural Language With Conversational Agents. *Current Directions in Psychological Science*, *23*(5), 374–380. https://doi.org/10.1177/0963721414540680

Gratzer, D., & Goldbloom, D. (2019). Open for Business: Chatbots, E-therapies, and the Future of Psychiatry. *Canadian Journal of Psychiatry*, *64*(7), 453–455. https://doi.org/10.1177/0706743719850057

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human Computer Studies*, *64*(2), 79–102. https://doi.org/10.1016/j.ijhcs.2005.06.002

Hornbæk, K., & Law, E. L. C. (2007). Meta-analysis of correlations among usability measures. In *Conference on Human Factors in Computing Systems - Proceedings* (pp. 617–626). San Jose: ACM Press. https://doi.org/10.1145/1240624.1240722

Ingwersen, P. (1994). Polyrepresentation of Information Needs and Semantic Entities: Elements of a Cognitive Theory for Information Retrieval Interaction. In W. B. Croft & C. J. van Rijsbergen (Eds.), *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 101–110). New York: Springer-Verlag. https://doi.org/10.1007/978-1-4471-2099-5_11

Ingwersen, P. (1996). Interaction : Elements of a Cognitive IR Theory. *Journal of Documentation*, *52*(1), 3–50.

International Organization for Standardization. (2018). Ergonomics of human-system

interaction — Part 11: Usability: Definitions and concepts.

Ischen, C., Araujo, T., Voorveld, H., van Noort, G., & Smit, E. (2020). Privacy Concerns in Chatbot Interactions. In A. Følstad, T. Araujo, S. Papadopoulos, E. L.-C. Law, O.-C. Granmo, E. Luger, & P. B. Brandtzaeg (Eds.), *Chatbot Research and Design. CONVERSATIONS 2019.* (pp. 34–48). Amsterdam, Netherlands: Springer International Publishing. https://doi.org/10.1007/978-3-030-39540-7_3

Jackson, A., Lin, J., Milligan, I., & Ruest, N. (2016). Desiderata for Exploratory Search Interfaces to Web Archives in Support of Scholarly Activities. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries - JCDL '16* (pp. 103–106). New York, New York, USA: ACM Press. https://doi.org/10.1145/2910896.2910912

Jurafsky, D., & Martin, J. H. (2000a). Discourse. In S. Russel & P. Norvig (Eds.), *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition* (pp. 663–713). Upper Saddle River: Prentice Hall.

Jurafsky, D., & Martin, J. H. (2000b). Introduction. In S. Russel & P. Norvig (Eds.), *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition* (pp. 1–18). Upper Saddle River, NJ, USA: Prentice Hall.

Jurafsky, D., & Martin, J. H. (2000c). Morphology and Finite-State Transducers. In S. Russel & P. Norvig (Eds.), *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition* (pp. 57–90). Upper Saddle River, NJ, USA: Prentice Hall.

Jurafsky, D., & Martin, J. H. (2000d). Regular Expressions and Automata. In S. Russel & P. Norvig (Eds.), *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition* (pp. 21–89). Upper Saddle River, NJ, USA: Prentice Hall.

Jurafsky, D., & Martin, J. H. (2000e). Representing Meaning. In S. Russel & P. Norvig (Eds.), *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition* (pp. 497–571). Upper Saddle River, NJ, USA: Prentice Hall.

Kelly, D. (2009). *Methods for evaluating interactive information retrieval systems with users*. *Foundations and Trends in Information Retrieval* (Vol. 3). https://doi.org/10.1561/1500000012

Kelly, D., Arguello, J., Edwards, A., & Wu, W. C. (2015). Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework. *ICTIR 2015 - Proceedings of the 2015 ACM SIGIR International Conference on the Theory of Information Retrieval*, 101–110. https://doi.org/10.1145/2808194.2809465

Kelly, D., Harper, D. J., & Landau, B. (2008). Questionnaire mode effects in interactive information retrieval experiments. *Information Processing and Management*, *44*(1), 122–141. https://doi.org/10.1016/j.ipm.2007.02.007

Kelly, D., & Sugimoto, C. R. (2013). A systematic review of interactive information retrieval evaluation studies, 1967-2006. *Journal of the American Society for Information Science and Technology*, *64*(4), 745–770. https://doi.org/10.1002/asi.22799

Kocaballi, A. B., Berkovsky, S., Quiroz, J. C., Laranjo, L., Tong, H. L., Rezazadegan, D., … Coiera, E. (2019). The personalization of conversational agents in health care: Systematic

review. *Journal of Medical Internet Research*, *21*(11), 1–15. https://doi.org/10.2196/15360

Lalmas, M., O'Brien, H., & Yom-Tov, E. (2014a). Approaches Based on Self-Report Methods. In G. Marchionini (Ed.), *Measuring User Engagement* (pp. 11–29). Morgan & Claypool.

Lalmas, M., O'Brien, H., & Yom-Tov, E. (2014b). Introduction and Scope. In G. Marchionini (Ed.), *Measuring User Engagement* (pp. 1–10). Morgan & Claypool.

Lavrakas, P. (2008). Self-Reported Measure. In P. J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods* (pp. 805–806). Thousand Oaks: Sage Publications. https://doi.org/10.4135/9781412963947

Lee, K., Jo, J., Kim, J., & Kang, Y. (2019). Can Chatbots Help Reduce the Workload of Administrative Officers? - Implementing and Deploying FAQ Chatbot Service in a University. In C. Stephanidis (Ed.), *HCII: International Conference on Human-Computer Interaction 2019* (pp. 348–354). Orlando: Springer Nature. https://doi.org/10.1007/978-3-030-23522-2_45

Liang, P. (2014). Talking to computers in natural language. *XRDS: Crossroads, The ACM Magazine for Students*, *21*(1), 18–21. https://doi.org/10.1145/2659831

Lincoln, Y. S., & Guba, E. G. (1985). Postpositivism and the Naturalist Paradigm. In *Naturalistic Inquiry* (pp. 14–46). London: SAGE Publications.

Lopatovska, I., & Arapakis, I. (2011). Theories, methods and current research on emotions in library and information science, information retrieval and human-computer interaction. *Information Processing and Management*, *47*(4), 575–592. https://doi.org/10.1016/j.ipm.2010.09.001

Lowe, H. J. (1994). Understanding and Using the Medical Subject Headings (MeSH) Vocabulary to Perform Literature Searches. *JAMA: The Journal of the American Medical Association*, *271*(14), 1103–1108. https://doi.org/10.1001/jama.1994.03510380059038

Luger, E., & Sellen, A. (2016). "Like having a really bad PA": The gulf between user expectation and experience of conversational agents. In *Conference on Human Factors in Computing Systems - Proceedings* (pp. 5286–5297). ACM Press. https://doi.org/10.1145/2858036.2858288

Ly, K. H., Ly, A. M., & Andersson, G. (2017). A fully automated conversational agent for promoting mental well-being: A pilot RCT using mixed methods. *Internet Interventions*, *10*(August), 39–46. https://doi.org/10.1016/j.invent.2017.10.002

Marchionini, G. (2006). Exploratory search: from finding to understanding. *Communications of the ACM*, *49*(4), 41–46. https://doi.org/10.1145/1121949.1121979

McTear, M., Callejas, Z., & Griol, D. (2016a). Conversational Interfaces: Past and Present. In *The Conversational Interface* (pp. 51–68). Cham: Springer International Publishing.

McTear, M., Callejas, Z., & Griol, D. (2016b). Dialog Management. In *Spoken Dialogue Technology* (pp. 209–233). Cham: Springer International Publishing.

McTear, M., Callejas, Z., & Griol, D. (2016c). Spoken Language Understanding. In *The Conversational Interface* (pp. 161–185). Cham: Springer International Publishing.

Meadows, J. (2002). S.C. Bradford and documentation: A review article. *Journal of Librarianship and Information Science*, *34*(3), 171–174.

https://doi.org/10.1177/096100002401010763

Miller, G. A. (1994). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *101*(2), 343–352. https://doi.org/10.1037/0033-295X.101.2.343

Milne, M., Luerssen, M., Lewis, T., Leibbrandt, R., & Powers, D. (2011). Designing and Evaluating Interactive Agents as Social Skills Tutors for Children with Autism Spectrum Disorder. In *Conversational Agents and Natural Language Interaction* (pp. 23–48). IGI Global. https://doi.org/10.4018/978-1-60960-617-6.ch002

Minker, J. (1977). Information storage and retrieval: a survey and functional description. *ACM SIGIR Forum*, *12*(2), 12–108. https://doi.org/10.1145/1095515.1095516

Muramatsu, J., & Pratt, W. (2001). Transparent queries: Investigating users' mental models of search engines. In *SIGIR Forum (ACM Special Interest Group on Information Retrieval)* (pp. 217–224). New Orleans: ACM Press.

Nielsen, J. (1993a). Response Times: The 3 Important Limits. Retrieved July 23, 2020, from https://www.nngroup.com/articles/response-times-3-important-limits/

Nielsen, J. (1993b). Usability Assessment Methods beyond Testing. In *Usability Engineering* (pp. 207–226). Mountain View, California: Morgan Kaufmann Publishers.

Nielsen, J. (1993c). Usability Testing. In *Usability Engineering* (pp. 165–206). Mountain View, California: Morgan Kaufmann Publishers.

Nielsen, J. (1993d). What Is Usability? In *Usability Engineering* (pp. 23–48). Mountain View, California: Morgan Kaufmann Publishers.

Nielsen, J., & Levy, J. (1994). Measuring usability: Preference vs. Performance. *Communications of the ACM*, *37*(4), 66–75. https://doi.org/10.1145/175276.175282

Norman, M., & Thomas, P. (1990). Informing HCI Design from Conversation Analysis. In P. Luff (Ed.), *Computers and Conversation* (1st ed., pp. 51–66). London: Academic Press.

O'Brien, H. (2016). Theoretical Perspectives on User Engagement. In *Why Engagement Matters* (pp. 1–26). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-27446-1_1

O'Brien, H., Arguello, J., & Capra, R. (2020). An empirical study of interest, task complexity, and search behaviour on user engagement. *Information Processing and Management*, *57*(3), 1–19. https://doi.org/10.1016/j.ipm.2020.102226

O'Brien, H., & Cairns, P. (2015). An empirical evaluation of the User Engagement Scale (UES) in online news environments. *Information Processing and Management*, *51*(4), 413–427. https://doi.org/10.1016/j.ipm.2015.03.003

O'Brien, H., Cairns, P., & Hall, M. (2018). A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human Computer Studies*, *112*(December 2017), 28–39. https://doi.org/10.1016/j.ijhcs.2018.01.004

O'Brien, H., & Toms, E. G. (2008). What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, *59*(6), 938–955. https://doi.org/10.1002/asi.20801

O'Brien, H., & Toms, E. G. (2010a). Is there a universal instrument for measuring interactive

information retrieval? In *Proceeding of the third symposium on Information interaction in context - IIiX '10* (pp. 1–8). New Brunswick, New Jersey: ACM Press. https://doi.org/10.1145/1840784.1840835

O'Brien, H., & Toms, E. G. (2010b). The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, *61*(1), 50–69. https://doi.org/10.1002/asi.21229.1

Palagi, E., Gandon, F., Giboin, A., & Troncy, R. (2017). A Survey of Definitions and Models of Exploratory Search. In *Proceedings of the 2017 ACM Workshop on Exploratory Search and Interactive Data Analytics - ESIDA '17* (pp. 3–8). New York, New York, USA: ACM Press. https://doi.org/10.1145/3038462.3038465

Pereira, J. (2016). Leveraging chatbots to improve self-guided learning through conversational quizzes. In *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality - TEEM '16* (pp. 911–918). New York, New York, USA: ACM Press. https://doi.org/10.1145/3012430.3012625

Perski, O., Crane, D., Beard, E., & Brown, J. (2019). Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? An experimental study. *Digital Health*, *5*, 1–13. https://doi.org/10.1177/2055207619880676

Peterson, R. (2013). The Process of Questionnaire Construction. In *Constructing Effective Questionnaires* (pp. 13–28). Thousand Oaks: Sage Publications. https://doi.org/10.4135/9781483349022.n2

Pia Borlund. (2003). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, *8*(3), 1–31. Retrieved from http://informationr.net/ir/8-3/paper152.html

Pickard, A. J. (2013). Major research paradigms. In *Research Methods in Information* (2nd ed., pp. 5–24). London: Facet Publishing.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*(1), 1–15. https://doi.org/10.1016/S0001-6918(99)00050-5

Radlinski, F., & Craswell, N. (2017). A theoretical framework for conversational search. *CHIIR 2017 - Proceedings of the 2017 Conference Human Information Interaction and Retrieval*, 117–126. https://doi.org/10.1145/3020165.3020183

Reed, K., & Meiselwitz, G. (2011). Teacher agents: The current state, future trends, and many roles of intelligent agents in education. In *Lecture Notes in Computer Science* (Vol. 6778 LNCS, pp. 69–78). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-21796-8_8

Rich, E. (1984). Natural-Language Interfaces. *Computer*, *17*(9), 39–47.

Roy, M., & Chi, M. T. H. (2003). Gender differences in patterns of searching the Web. *Journal of Educational Computing Research*, *29*(3), 335–348. https://doi.org/10.2190/7BR8-VXA0-07A7-8AVN

Russell-Rose, T., & Tate, T. (2012a). Displaying and Manipulating Results. In *Designing the Search Experience* (pp. 129–166). Elsevier Inc. https://doi.org/10.1016/b978-0-12-396981-1.00006-9

Russell-Rose, T., & Tate, T. (2012b). Faceted Search. In *Designing the Search Experience* (pp.

167–218). Elsevier Inc. https://doi.org/10.1016/b978-0-12-396981-1.00007-0

Russell-Rose, T., & Tate, T. (2012c). Formulating the Query. In *Designing the Search Experience* (pp. 99–128). Elsevier Inc. https://doi.org/10.1016/b978-0-12-396981-1.00005-7

Sauro, J., & Lewis, J. R. (2009). Correlations among prototypical usability metrics: Evidence for the construct of usability. In S. Greenberg, S. E. Hudson, K. Hinckley, M. R. Morris, & D. R. Olsen (Eds.), *CHI '09: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1609–1618). Boston: ACM Press. https://doi.org/10.1145/1518701.1518947

Sauro, J., & Lewis, J. R. (2010). Average task times in usability tests: What to report? In *CHI 2010 - Conference on Human Factors in Computing Systems* (pp. 2347–2350). Atlanta: ACM Press. https://doi.org/10.1145/1753326.1753679

Scherer, E. A., Ben-Zeev, D., Li, Z., & Kane, J. M. (2017). Analyzing mHealth Engagement: Joint Models for Intensively Collected User Engagement Data. *JMIR MHealth and UHealth*, *5*(1), 1–9. https://doi.org/10.2196/mhealth.6474

Serafini, A. (2013a). Extending Search. In *Apache Solr Beginner's Guide* (pp. 179–220). Birmingham: Packt Publishing.

Serafini, A. (2013b). Getting Ready with the Essentials. In *Apache Solr Beginner's Guide* (pp. 33–56). Birmingham: Packt Publishing.

Serafini, A. (2013c). Indexing with Local PDF Files. In *Apache Solr Beginner's Guide* (pp. 56–100). Birmingham: Packt Publishing.

Serafini, A. (2013d). Searching the Example Data. In *Apache Solr Beginner's Guide* (pp. 137–178). Birmingham: Packt Publishing.

Shackel, B. (2009). Human-computer interaction - Whence and whither? *Interacting with Computers*, *21*(5–6), 353–366. https://doi.org/10.1016/j.intcom.2009.04.004

Shneiderman, B. (2007). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. *The Craft of Information Visualization*, 364–371. https://doi.org/10.1016/b978-155860915-0/50046-9

Sundar, S. S. (2007). Social psychology of interactivity in human-website interaction. In A. Joinson, K. McKenna, T. Postmes, & U.-D. Reips (Eds.), *The Oxford Handbook of Internet Psychology* (pp. 89–104). Oxford: Oxford University Press.

Sundar, S. S., Bellur, S., Oh, J., Jia, H., & Kim, H. S. (2016). Theoretical Importance of Contingency in Human-Computer Interaction: Effects of Message Interactivity on User Engagement. *Communication Research*, *43*(5), 595–625. https://doi.org/10.1177/0093650214534962

Sutcliffe, A. (2016). Designing for User Experience and Engagement. In H. O'Brien & P. Cairns (Eds.), *Why Engagement Matters* (pp. 105–126). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-27446-1_5

Sutcliffe, A., Thew, S., De Bruijn, O., Buchan, I., Jarvis, P., McNaught, J., & Procter, R. (2010). User engagement by user-centred design in e-Health. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1926), 4209–4224. https://doi.org/10.1098/rsta.2010.0141

Thomas, P., Czerwinski, M., McDuf, D., Craswell, N., & Mark, G. (2018). Style and alignment in information-seeking conversation. In *CHIIR 2018 - Proceedings of the 2018 Conference on Human Information Interaction and Retrieval* (pp. 42–51). New York: ACM Press. https://doi.org/10.1145/3176349.3176388

Thomas, P., McDuff, D., Czerwinski, M., & Craswell, N. (2017). MISC: A data set of information-seeking conversations. In *Proceedings ofIn- ternational Workshop on Conversational Approaches to Information Retrieval (CAIR'17)* (pp. 1–6). Tokyo. Retrieved from http://search.proquest.com.ezaccess.library.uitm.edu.my/docview/1095546335?accountid=42518

Thompson, A., Gallacher, A., & Howarth, M. (2018). Stimulating task interest: human partners or chatbots? In *Future-proof CALL: language learning as exploration and encounters – short papers from EUROCALL 2018* (Vol. 2018, pp. 302–306). Research-publishing.net. https://doi.org/10.14705/rpnet.2018.26.854

Traum, D., & Larsson, S. (2003). The information state approach to dialogue management. In J. van Kuppevelt (Ed.), *Current and New Directions in Discourse & Dialogue* (pp. 325–353). Dordrecht: Springer Science + Business Media.

Tunkelang, D., & Marchionini, G. (2009). Front-End Concerns. In G. Marchionini (Ed.), *Faceted Search* (pp. 57–68). Morgan & Claypool.

Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. *Canadian Journal of Psychiatry*, *64*(7), 456–464. https://doi.org/10.1177/0706743719828977

Vakulenko, S., Markov, I., & de Rijke, M. (2017). Conversational Exploratory Search via Interactive Storytelling. In *ICTIR'17 Workshop on Search-Oriented Conversational AI (SCAI 2017)* (pp. 1–5). Amsterdam, Netherlands. Retrieved from http://arxiv.org/abs/1709.05298

Van Den Haak, M. J., De Jong, M. D. T., & Schellens, P. J. (2003). Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour and Information Technology*, *22*(5), 339–351. https://doi.org/10.1080/0044929031000

Vtyurina, A., Savenkov, D., Agichtein, E., & Clarke, C. L. A. (2017). Exploring conversational search with humans, assistants, and wizards. *Conference on Human Factors in Computing Systems - Proceedings*, *Part F1276*, 2187–2193. https://doi.org/10.1145/3027063.3053175

Ward, D. (2005). Why Users Choose Chat. *Internet Reference Services Quarterly*, *10*(1), 29–46. https://doi.org/10.1300/J136v10n01_03

Watt, W. C. (1968). Habitability. *American Documentation*, *19*(3), 338–351. https://doi.org/10.1002/asi.5090190324

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, *27*(3), 236–247. https://doi.org/10.1016/j.ijresmar.2010.02.004

Weizenbaum, J. (1966). ELIZA---a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.

https://doi.org/10.1145/365153.365168

White, R. W. (2018). Viewpoint opportunities and challenges in search interaction. *Communications of the ACM*, *61*(12), 36–38. https://doi.org/10.1145/3195180

White, R. W., & Roth, R. A. (2009). Defining Exploratory Search. In G. Marchionini (Ed.), *Exploratory Search: Beyond the Query-Response Paradigm* (pp. 9–23). Morgan & Claypool.

Wiebe, E. N., Lamb, A., Hardy, M., & Sharek, D. (2014). Measuring engagement in video game-based environments: Investigation of the User Engagement Scale. *Computers in Human Behavior*, *32*, 123–132. https://doi.org/10.1016/j.chb.2013.12.001

Wilson, C. (2013). Questionnaires and Surveys. In *Credible Checklists and Quality Questionnaires: A User-Centered Design Method* (1st ed., pp. 30–77). Morgan Kaufmann.

Wilson, M. L. (2011). Modern Search User Interfaces. In G. Marchionini (Ed.), *Search User Interface Design* (pp. 29–80). Morgan & Claypool.

Winograd, T. (1972). Understanding natural language. *Cognitive Psychology*, *3*(1), 1–191. https://doi.org/10.1016/0010-0285(72)90002-3

Woods, W. A. (1973). Progress in natural language understanding — An application to lunar geology. In *AFIPS Conference Proceedings - 1973 National Computer Conference and Exposition, AFIPS 1973* (pp. 441–450). New York, NY, United States: Association for Computing Machinery.

Xie, I. (2008a). Interactive IR in OPAC Environments. In *Interactive Information Retrieval in Digital Environments* (pp. 29–52). Hershey, New York: IGI Global.

Xie, I. (2008b). User-Oriented IR Research Approaches. In *Interactive Information Retrieval in Digital Environments* (pp. 1–28). Hershey, New York: IGI Global.

Zhou, M. (2014). Gender difference in web search perceptions and behavior: Does it vary by task performance? *Computers and Education*, *78*, 174–184. https://doi.org/10.1016/j.compedu.2014.06.005

Zhu, P., Zhang, Z., Li, J., Huang, Y., & Zhao, H. (2018). Lingke: A Fine-grained Multi-turn Chatbot for Customer Service. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations* (pp. 108–112). Santa Fe, New Mexico: Association for Computational Linguistics. Retrieved from http://arxiv.org/abs/1808.03430