Towards task-sensitive assistance in public spaces*

Markus Kattenbeck
Information Science, University of Regensburg
Universitätsstraße 31, D-93053 Regensburg, Germany
markus.kattenbeck@ur.de

Melanie A. Kilian
Information Science, University of Regensburg
Universitätsstraße 31, D-93053 Regensburg, Germany
melanie.kilian@ur.de

Matthias Ferstl
Information Science, University of Regensburg
Universitätsstraße 31, D-93053 Regensburg, Germany
mathias2.ferstl@student.ur.de

Florian Alt
Research Institute CODE, Bundeswehr University Munich
Werner-Heisenberg-Weg 39, D-85579 Neubiberg, Germany
florian.alt@unibw.de

Bernd Ludwig
Information Science, University of Regensburg
Universitätsstraße 31, D-93053 Regensburg, Germany
bernd.ludwig@ur.de

Abstract

Purpose – Performing tasks in public spaces can be demanding due to task complexity. Systems that can keep track of the current task state may help their users to successfully fulfill a task. These systems, however, require major implementation effort. In this paper a least effort approach is taken to investigate *if* and *how* a mobile information assistant which has only basic task-tracking capabilities can support users. This means, we are interested in whether such a system is able to have an impact on the way a workflow in public space is perceived.

Design/methodology/approach – We implement and test AIRBOT, a mobile chatbot application that can assist air passengers in successfully boarding a plane. We apply a three-tier approach and, first, conduct expert and passenger interviews to understand the workflow and the information needs occurring therein; second, we implement a mobile chatbot application providing minimum task-tracking capabilities to support travelers by providing boarding-relevant information in a proactive manner. Finally, we evaluate this application by means of an in-situ study ($N = 101$ passengers) at a major European airport.

*This article is an extended version of a poster paper presented by the authors at the MobileHCI '18 conference (see Kattenbeck et al. 2018).

Findings – We provide evidence that basic task-tracking capabilities are sufficient to affect the users' task perception. AIRBOT is able to decrease the perceived workload airport services impose on users. It has a negative impact on satisfaction with non-personalized information offered by the airport, though.
Originality/value – The study shows that the number of features is not the most important means to successfully provide assistance in public space workflows. The study can, moreover, serve as a blueprint to design task-based assistants for other contexts.
Keywords Assistance system, Cooperative problem solving, Human-computer interaction, Mobile information needs
Paper type Research paper

1. Introduction

Chatbots have seen increasing interest in recent years (e.g. Chandel et al. 2018; Ma and Ho 2018). One particular focus has been on acquiring knowledge about the cohesiveness of discourse across multiple turns based on dialogue corpora (e.g. De Gasperis, Chiari, and Florio 2013; AbuShawar and Atwell 2013; Callejas-Rodrıguez et al. 2016). State-tracking capabilities of chatbots have seen less interest, though. This is a major research gap, in particular with respect to mobile chat-based information companions which may be used in scenarios where providing information in a timely manner is highly important. These activities, which are often performed in public spaces, are frequently standardized, resulting in two challenges for people: First, the correct order of sequences is less obvious for *inexperienced persons* (e.g. first time air passengers). Second, disruptions of the sequence of activities by *unforeseen events* may require users to find alternative solutions (e.g. switching modes of public transportation in case of breakdowns of a bus). In both cases, the perceived complexity of the activities performed in public spaces increases heavily, even for experienced persons. Personal information systems capable of tracking the current state of the activities can provide a means to reduce this perceived complexity. Building systems requires major implementation effort, though. With these costs in mind, it is useful to, first, understand the benefits a chatbot with *basic* capabilities to track a user's state within a particular workflow might even have.

This paper reports on an empirical investigation of this least effort question by focusing on support provided to departing passengers at an airport prior to boarding. To this end, we take a multi-perspective approach which is reflected in the structure of this paper: First, we report on a two-part interview study (see section 3.2) during which we interviewed both airport staff members (i.e. information experts) and departing air passengers. This study was a suitable means to learn about the ideal workflow from the staff members' perspective and the degree as to which passengers share their point of view. Taken together, both parts of the interview study reveal a rather simple workflow model and a set of requirements an information companion should fulfill. Subsequently, we take a least effort approach to understand whether a personal information companion with a minimal set of state-tracking abilities can positively influence the perception of passenger services[a]. We, therefore, implement a chatbot application called AIRBOT, which is able to inform its users proactively

[a]The term *passenger services* is used throughout this paper to denote all services offered by the airport and the operating airlines to handle departing passengers, for example check-in, baggage drop, security check and boarding services.

due to its basic state-tracking capabilities (see section 3.3). Using this prototype application we conduct an in-situ user study using live data streams (see section 4). We find that even a system with minimal capabilities in natural language understanding and generation has a positive effect on the perceived workload, resulting in an increased satisfaction with passenger services.

2. Related Work

Three different strands of prior evidence are important for the work presented in this paper. The first of these provides a justification for studying personalized assistance by reviewing literature on delivering information in public spaces. The second field of interest is concerned with state-of-the-art, goal-oriented chatbots which focuses on the importance of task awareness to provide meaningful personalized assistance. Finally, the third part of this section deals with existing mobile information systems providing assistance to airport passengers.

2.1. *Delivering Information in Public Spaces*

Digital displays to inform and guide persons in public spaces have seen increasing interest within the HCI community up to today (see e.g. Iwai et al. 2006; Sato et al. 2010; Sharifi, Payne, and David 2006; Väänänen-Vainio-Mattila et al. 2013; Tomitsch et al. 2014; Shi and Alt 2016). All of these share the conclusion that these displays are a useful means of presenting information, if and only if they are visible and noticed. This conclusion implies that it is a major challenge to draw the users' attention to these displays in public spaces (see e.g. Müller, Alt, et al. 2010; Müller, Walter, et al. 2012 for studies on capturing a user's attention with this display type).

Personalized information companions, such as smartphones, however, can push information to their users and, thus, can draw the users' attention more easily. Personalization is much easier to achieve using these systems compared to public displays, which first need to identify users (see e.g. Davies et al. 2014; Alt and Vehns 2016). Location-based tourism services for mobile devices have been among the first application areas with respect to personalization. Studies such as Schmidt-Belz et al. (2003) provide evidence that participants perceive this solution as more beneficial compared to other, non-personalized information sources like free web content, books, or maps. In addition to personalization, context awareness of information presented on smartphones has been revealed to be particularly important (see e.g. Sohn et al. 2008; Nugent et al. 2015). Overall, prior evidence suggests that a personal device can be a useful means of providing information to users in a personalized and context-aware manner. Proactively acting chatbots are one option, among others, to fulfill both goals.

2.2. *Chatbot Strategies*

Research on goal-oriented chatbots has largely been focusing on dialogues. Traditionally, these have been based on a set of rules represented either in lookup tables (see e.g. Wang and Y. Lin 2000) or Prolog-like rules (see e.g. Biermann et al. 1997). Very recently, however, neural networks trained on very large dialogue corpora have been applied to tackle a variety of challenges in goal-oriented chatbot dialogues. Each of these challenges relates to conversational capabilities, which are a major user requirement (see e.g. Zamora 2017; Jain et al. 2018). In these studies, neural networks are, for example, used to personalize ranking algorithms for answers given by chatbots (see B. Liu et al. 2018) or to create utterances based on a user model (see Serras, Torres, and Pozo 2017). Related to this, improving the coherence in dialogues has also been an important research focus, which is generally seen as a major success factor (see e.g. Følstad and Brandtzæg 2017). Mixing task-oriented dialogue techniques with chatbot techniques have been shown to increase dialogue persuasiveness (see Andrews, Manandhar, and De Boni 2008); Takanobu et al. (2018) used reinforcement learning to segment topics successfully, thereby addressing aspects of both local and global topic coherence. In a similar realm, Ilievski (2018) transferred behaviour learnt from one domain to another by means of deep reinforcement learning, a scenario in which the two domains share a common ground, e.g. one domain is an extension of the other. This approach stresses the importance of context. It is in line with the increasing focus on the peculiarities of a specific task in goal-oriented chatbot research. Asri et al. (2017) present a corpus of human-human dialogues on a trip planning setting which can be employed to learn dialogue flow and the way decisions are made within these dialogues. Lee, Kim, and Seo (2013) use neural networks to classify domain actions of users and simultaneously plan system actions. Similarly, Seon et al. (2014) focus on the classification of domain action and provide evidence that taking these into account can outperform corpus-only strategies. Being able to define individual tasks relevant for domain actions, however, requires a thorough understanding of the user needs involved. Ultimately, the users' motivation for using chatbots (see Brandtzaeg and Følstad 2018) need to be understood in order to make these applications successful. Qualitative interviews have yielded promising results in understanding these needs. M. Dirin, A. Dirin, and Laine (2015) identified goals by qualitatively analyzing interviews and deriving information needs from the data. Similarly, Vaccaro et al. (2018) conducted interviews with personal stylists to understand the type of advice those give to their costumers. Based on these insights they implemented and tested a stylist chatbot by means of a user study which revealed several similarities in terms of goals users try to achieve in both worlds.

This overview shows, first, that dialogue coherence is a major research focus, which requires very large corpora. Second, prior studies increasingly stress the importance of understanding user tasks to make chatbots successful. Qualitative interviews have been a useful means of understanding these tasks. We decided to choose passenger services at airports as a use case, which is a non-trivial task scenario that does not show too many degrees of freedom. It is, nevertheless, complex enough to pose a number of information needs which becomes evident by the large number of information kiosks that can be found in airport buildings. The information needs occurring also span an adequate range of different types, ranging from fact finding (e.g. "When does the check-in open?") and time management (e.g. "How can I spend my time until departure?") to wayfinding (e.g. "How do I get to my gate”). A passenger survey conducted worldwide by IATA in 2016 supports this view by revealing the need for personal information systems at airports (IATA 2016). Moreover, recent work suggests that specific user groups (e.g. small persons) may not benefit from signage in public spaces like airports (Lueg 2014; Cox et al. 2017; Olsson 2016). Finally, passenger services at aiports are a good example for a situation a larger number of people is not subjected to on a very regular basis. Hence, using this scenario will help us to understand if basic task-awareness of a chatbot can have an impact on a user’s perception of task-related services in public spaces.

2.3. *Existing Mobile Assistance Systems for Air Travelers*

A variety of mobile assistance systems provide air travelers with personalized and/or context-aware information. Some of these are research-related (e.g. *FlyTalk* Awori et al. 2012; *GatePal* Y. E. Liu et al. 2016) and many others[b] have a commercial focus (see Radaha and M. E. Johnson 2013 and Matthews 2015 for non-exhaustive overviews). All but *GatePal* (Y. E. Liu et al. 2016), *App in the Air* (Pronin 2018) and *FRAnky* (FraportAG 2018) provide context-sensitive features; none of the systems, however, tracks the user’s state within the workflow. The push notifications they offer are also not aligned with the current workflow-state of the user. While Y. E. Liu et al. 2016 did perform a task in a similar research context than we do, their goals was to specifically assist users with disabilities; they focused exclusively on the time consumption of passenger service activities. *App in the Air* and *FRAnky* are both commercial systems which are based on similar approaches than AIRBOT is. It is, therefore, important to note that we do not aim to present another fully developed system but report on a least effort study: AIRBOT is designed to have minimal task tracking capabilities needed to guide air travelers successfully through passenger services. Our goal is to understand whether these *basic* capabilities are sufficient to show a positive impact on passenger satisfaction.

3. Learning About Information Needs

The discussion above suggests that information, tailored to an individual user’s situation can be provided via smartphone, in particular as smartphones can be used to track the user’s current task. A model of the workflow to be tracked is, however, required to enable the development of suitable task-tracking algorithms. With this aim in mind we conducted interviews with both passengers and information personnel. We will, first, provide descriptive information about these interviews. We, then, move on to the most important aspects emerging from these interviews. While the interviews conducted with experts reveal which information needs are to be fulfilled by an information assistant, the passenger interviews reveal their mental model of the workflow. Two main conclusions can be drawn from these interviews. First, there is an ideal workflow leading to successful boarding according to the experts. The passengers’ mental model is in line with this workflow which seems to

[b]e.g. *GateGuru*, *Airports*, *Flight+*, Apple’s *Passbook*, *LoungeBuddy*, *MiFlight*, *Airport Transit Guide*, *FlySmart*, *Passngr*, *App in the Air*, *FLIO*, *FRAnky* and *Hamburg Airport Servicebot*.

be rather simplistic on first sight. Passenger interviews indicate, however, that it is sufficient to successfully complete boarding. The second conclusion relates to the importance of context because the situation passengers are in is unfamiliar to many of them.

3.1. *General Information*

3.1.1. *Interviews With Staff Members*

The first and the second author conducted semi-structured interviews with eight female[c] information personnel of a major airport in Europe. Although recruited via their employer, all interview partners volunteered to be interviewed. The interviews were conducted on-site (i.e. in a separate room at the airport during the interviewees' regular working hours), face-to-face between December 12th, 2016 and January 25th, 2017. Each interview lasted between 61 and 91 minutes, with a median duration of 77 minutes. Each interviewee had a minimum of three years experience in dealing with passenger information needs. The interviewees also were familiar with these needs from different work perspectives, i.e. face-to-face at information counters, at phone- or video telephony-based helplines and/or on social media channels. The interview guide for these semi-structured interviews was, consequently, designed to cover the interviewees' day-to-day experience and the various different information needs they encounter as broadly as possible. Interviewees were, for example, asked to elaborate on the different types of questions passengers have, their respective frequency and how they advise passengers to solve these.

3.1.2. *Passenger Interviews*

Six student experimenters conducted semi-structured interviews with 133 passengers ($\bar{x}_{age} = 38$ years, age range: $18 - 72$ years). 89 of these interviews were conducted in the public area of the airport, the remaining 44 within the security area[d]. The length of the interviews ranged between three and 30 minutes, with a median length of nine minutes. The short duration can be explained by the fact that the majority of interviewees was recruited in the public area next to their check-in; they were, consequently, not willing to spend much time before passing the security check. The interviews primarily focused on information needs occurring throughout the passengers' stay at the airport and about their wayfinding behavior. This focus yielded relevant information about the passengers' understanding of passenger services.

3.1.3. *Transcription and Presentation of Interviews*

The majority of the interviews were conducted in German and transcribed literally using the software f4transkript (Audiotranskription 2018). An ad hoc transcription guide was used to

[c]In general, the information center of this airport has few male staff members. All of these refused to be interviewed, although several appointments were offered within weeks and interviewees were allowed to participate during their work time.
[d]Interviewers were accompanied by an airport staff member when conducting interviews within the security area. The staff member did not intervene during the interviews, but waited standing aside in non-hearing distance.

label pauses, stuttering or abruption of a word, overlap of interviewee and interviewer statements and comments of transcriptors. To foster readability, we translate German interview excerpts idiomatically to English and present this translated version only.

3.2. *Interview Results*

Three major topics emerged from the interviews: First, workflow-related questions occur most frequently according to the interview partners; second, questions related to wayfinding are encountered often; third, passengers are interested in leisure/waiting time activities.

3.2.1. *Workflow*

A first insight from interviews with staff members is the high frequency with which workflow-related questions occur. This applies — as one would expect — primarily to first time air passengers.

Participant E06 [e] It is primarily first time passengers who have no idea. They arrive at the airport, have a look at the flight information display and then, for example, try to get to the gate without having their bags dropped. This is something I encounter frequently. They are not aware of the fact that they need to go to the check-in first, to register.

Workflow-related questions, however, also arise when more experienced air travelers are faced with the intricacies of air traffic procedures. One frequently encountered example is the confusion resulting from codesharing agreements between airlines:

Participant E07 Passengers are often confused by codesharing agreements between airlines. For example, Lufthansa has numerous cooperations with other airlines. Passengers do not know the concept of a codeshare marketing flight, e.g. they believe they have booked a United flight, but the operating carrier is Lufthansa. So they get confused because they think they are going to fly with United but in fact it is a Lufthansa flight.

Both examples are in line with several well-known models of human information behaviour (Kuhlthau 1993; J. D. Johnson and Meischke 1993; Byström and Järvelin 1995; Freund 2015; for an overview see Case and Given 2016, p. 152–171), all of which stress the impact a person's task has on shaping an individual's information needs and (seeking) behavior. Furthermore, from an information personnel perspective, the interviews echo work, which acknowledges the effect of available time or time pressure on human information behavior (e.g. Savolainen 1995; Marcella, Pirie, and Lockerbie 2013). Interviewees, in particular, stressed *time* as most important context variable for successfully boarding a plane:

Participant E06 We often explain why it is required to arrive at the airport two hours prior to departure. This amount of time is reasonable because one has to keep in mind

[e] The letter *E* denotes cases where experts were interviewed, the letter *T* denotes travellers.

that there are other passengers queuing at check-in and security checks.

In fact, an ideal workflow with respect to time constraints became evident from the interviews. The workflow derived from the interviews is fairly simple and comprises 5 steps:

(1) Arrive at the airport approximately 2 hours prior to departure.
(2) Go to Check-in and/or drop baggage.
(3) Keep in mind that boarding will close about 15 to 20 minutes prior to departure.
(4) Complete the security check ahead of time for several reasons (e.g. queues of other passenger; walking distance to gate etc.)
(5) Walk to the gate and wait for boarding to start.

While this workflow may seem overly simplistic on first sight, passengers first and foremost describe the passenger service procedures they underwent or were going to undergo when asked to describe their activities at the airport. Due to the large amount of interview data, we use an excerpt from a single interview to illustrate our findings. This quote indicates that the passenger's mental model of the steps required to board a plane successfully shows a large overlap with the experts' ideal workflow:

Participant T01 [f] so i arrive at the s-bahn, i went outside, go direct to the airports [sic], just check my bag really fast and went to the gate. really quick, was really easy

Conclusion These excerpts reveal two important aspects: Successful boarding relies, firstly, on proper time management and, secondly, on the correct order of steps prior to boarding. Thereby, time constraints rely on the flight schedule and can be derived from flight numbers. For this reason, AIRBOT shows the capability to use live flight data (see below, section 3.3) and users are required to enter their flight number on AIRBOT's start-up.

3.2.2. *Wayfinding*

Generally speaking, wayfinding accounts for a large proportion of questions asked. These are frequently intertwined with workflow-related information needs. This finding is in line with Church and Smyth (2009) who suggest that *geographical needs* are very common (>31%) in mobile contexts, such as commuting or traveling. Thereby, *geographical needs* are defined as non-informational needs, i.e. information needs that "are focused on the goal of finding an answer to a question, however, the information need and the answer expected is dependent on location in some form" (Church and Smyth 2009, p. 251–253). Hence, workflow-related explanations given by information personnel are often intertwined with wayfinding advice.

Participant E04 Most questions are related to the usual operations; when people arrive at the airport, they might not know their gate or their check-in. Similarly, they want to know where the post office is located at. Things like that…

[f] Participant T01 decided to answer our questions in English language, i.e. we quote her answers literally here.

Participant E03 Most frequently, in which direction? Where to?

The passenger interviews also show that a lack of information regarding location can discomfort passengers — albeit the fact that the necessary steps are known to them and, in general, perceived as "easy" to follow:

Participant T01 I did// the online check in, but I didn't know where to check my bag

Conclusion A personal information companion taking the current task and its contextual factors, such as time and space, into account will, therefore, likely enhance the passengers' airport experience. This evidence leads to the requirement that AIRBOT can provide route instructions at the airport to assist the wayfinding of passengers.

3.2.3. *Leisure Activities*

Beyond workflow- and wayfinding-related information needs, a third frequently occurring topic of interest relates to leisure activities, which the following three examples illustrate:

Participant E03 Is there a cinema at this airport? What kind of activities would you suggest to do during my stop-over? Other questions are related to one of our events and so on.

Participant E02 Most notably, passengers frequently ask about restaurants in the security area of Terminal 2. They want to know the different food options on offer.

Participant E01 Moreover, there are questions about how to spend 10 hours stop-over time.

The information needs concerned with these leisure activities depend heavily on contextual information, i.e. time and, e.g. the person's current location.

These two factors become also evident from the passenger interviews:

Participant T82 Usually, I'm traveling on business class and like to arrive ahead of time at the airport — I enjoy staying in the lounge for some time and have some decent food.

Participant T115 I've been comparing prices; getting a coffee can be really expensive here. One coffee shop sells a coffee for 3.70 Euro, the one left to it for 3.50 Euro, so I bought it there. To be frank, I have not seen other options, anyway.

This is, again, in line with the human information behaviour literature emphasizing the importance of contextual factors in human information seeking (see e.g. Wilson 1999). As a consequence, AIRBOT offers basic capabilities to spend waiting time (see section 3.3 below).

3.3. *Capabilities of* AIRBOT

Taken together, the information we get from the interviews indicate that information needs which are encountered frequently relate to the passenger services workflow, to wayfinding in relation to this workflow or leisure activities and to the way passengers can spend their waiting time at the airport. In all of these cases, context is particularly important. Therefore, a functional prototype chatbot system needs to take the following context-related aspects into account:

Flight data i.e. the system considers the departure terminal and gate, airline, departure and landing time as well as changes concerning these flight data.

Current activity i.e. the activity the user has to perform next, e.g. baggage drop, security check, boarding etc.

Time i.e. the time left to departure

We apply a finite state machine to build a passenger task model (PTM) to take these contextual factors into account. Once initialized, AIRBOT uses this PTM to keep track of the user's passenger service tasks based on information gained through interacting with the user regularly. It can predict a passenger's activities, react to the current context and approach the passenger proactively with reminders, notifications and route instructions to the correct location for the current activity. On successful initialization, users can choose between two options: The first option, "Pass my waiting time.", suggests a number of websites on general travel information to the user (e.g. information on weather and tourist attractions at their travel destination). The second option, "I'm fine.", will make AIRBOT wait for a trigger.

These system triggers represent the shortest path through the task model and can either be time- or event-based. Time-based triggers (see Figure 1) are used to assist the user in performing all activities that are mandatory to board a plane. They are intended to make users aware of the *ideal workflow* by pushing notifications: For example, if the passenger has not yet checked in when the check-in opens, the system will send a reminder and the number of the check-in counter; having successfully checked in, passengers are either guided to the baggage drop-off or straight to the security check. In contrast to time-based triggers, event-based triggers are issued on user input (e.g. requesting suggestions for leisure activities to spend waiting time) or changes in flight data in which case the user receives a notification about this change.

The system was implemented for Facebook Messenger (Facebook 2017), which serves 34k chatbot systems (O'Brien 2016). We used this popular framework to avoid passengers not using AIRBOT due to a lack of knowledge on how to use the app. By using dialogue trees, we modeled interaction consisting of pre-defined chatbot messages and user answers, which adhere to the Facebook Design Guideline for developers (Facebook 2017), i.e. the messages are short and clearly understandable (see Figure 2). This design decision helps us to overcome the aforementioned challenge of dialogue coherence in chatbots (see section 2.2), thereby providing quick and easy communication. The user, however, can always enter arbitrary free text. Depending on the current activity and the related dialogue tree, an appropriate set of keyphrase extraction patterns is activated to analyse the entered free text. The dialogue evolves step by step depending on whether a system trigger or a user input requires an update of the PTM, i.e. the finite state machine moves to another state.

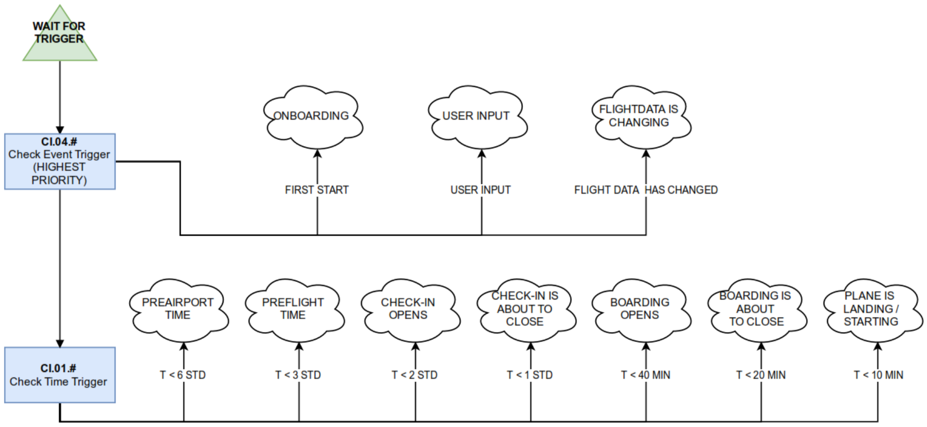

Fig. 1. Overview of the time- and event-based triggers eliciting messages of AIRBOT. This Figure was first published in Kattenbeck et al. 2018.

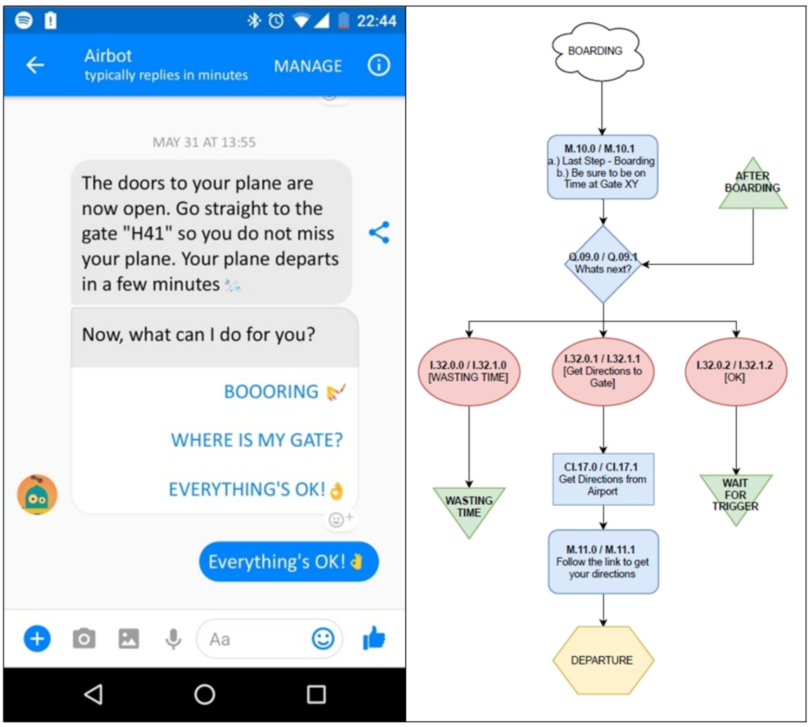

Fig. 2. The screenshot on the left (first published in Kattenbeck et al. 2018) visualizes generic messages and pre-defined user answers AIRBOT provides. Please note: Messages and prompts were translated from German to English. The Figure on the right shows the dialogue tree the screen on the left is based on.

4. Evaluation Method

This section details, first, the latent variables used in our study, the hypotheses about their relationships and their measurement, which, in combination, result in a structural equation model. We move on to justify the size of the sample in the study. Third, the acquisition of control (non-AIRBOT users) and treatment (AIRBOT users) group participants will be outlined. Finally, descriptive statistics will shed light on the participants in both groups.

4.1. *Latent Variables and Hypotheses*

We assessed the usefulness of AIRBOT's state-tracking capabilities by means of a between-subjects design in-situ experiment at a major European airport. We developed an ad-hoc survey to this end, involving the following latent variables (LVs).

SPAS (Satisfaction with PAssenger Services) The degree of perceived satisfaction a passenger shows with respect to all passenger services prior to boarding an airplane.

FPAS (Familiarity with PAssenger Services) The degree as to which a person is familiar with the process of passenger services prior to boarding.

DSIT (Day-related SITuational factors) The degree as to which passengers feel comfortable with walking distance, waiting time and time pressure prior to boarding.

SORI (Satisfaction with ORIentation opportunities) The degree as to which passengers feel oriented in literal as well as a figurative sense, i.e. they are aware about the correct next step.

In our study, *satisfaction with passenger services* (SPAS) is the key endogenous target construct. We are interested in providing assistance which comforts users and, hence, increases their satisfaction with passenger services. *Familiarity with these services* (FPAS), the *feeling of being oriented* (SORI) and the *day-dependent situation* (DSIT) are hypothesized to have a positive effect on SPAS (see Figure 3). In addition to that, FPAS and DSIT are hypothesized to have a positive influence on SORI. These hypotheses can be derived from our interview results (see section 3.2) which indicate that familiarity decreases the likeliness of information needs regarding workflow, that finding their way through passenger services is of key importance for passengers and that day-dependent aspects have an impact. These hypotheses describing the relationships between the LVs studied comprise the structural model part of our structural equation model.

The measurement model part, however, describes the relationships between a latent variable and those variables used to measure its value (MVs; usually depicted as boxes, see Figure 3). Table 1[g] presents which questions were used to measure each of the aforementioned latent variables. All LVs except the *day-related situational factors* (DSIT) are

[g] Two types of measurement are commonly distinguished. *Reflective* measurement is indicated through arrows pointing towards measured variables. This type of measurement, which is most widespread, assumes all indicators of a LV to provide equally good measurements of it and, consequently, a reasonable correlation between items is desired (see DeVellis 2012). This is in sharp contrast to *formative* measurement, where measured variables are supposed to be causes of the LV score and, therefore, must not be mutually interchangeable (see Jarvis, MacKenzie, and Podsakoff 2003, p. 203).

modeled in reflective measurement mode. Formatively measuring DSIT is reasonable in order to acknowledge that *day- and person-specific aspects*, which need not be correlated, may be important to this LV. Despite a possibly increased measurement error (see Diamantopoulos et al. 2012), single-indicator measurement was applied to avoid non-response bias induced by time effort needed to answer the survey. We use this type of measurement for *satisfaction* and *familiarity with passenger services* (SPAS and FPAS). A closed question format (5-point likert-like scale according to Rohrmann 1978) was used for each MV to minimize time effort in answering questions. Variable *o_ari* is the only exception to this closed question format. Here, users were asked to provide the number of times they needed to ask for (route) instructions (see 1). The comprehensibility of the wording of questions was cross-validated by two think-aloud protocols (see Raab-Steiner and Benesch 2015, p. 63 for this advice). The second of these think-aloud protocols took place at the airport itself.

LV	MV	Phrasing	ToM
Satisfaction with passenger services [SPAS]	s_pas	To what extent did you get along with passenger services? *(Wie gut kamen Sie mit dem Prozess der Flugabfertigung zu Recht?)*	R
Familiarity with passenger services [FPAS]	f_pas	To what extent are you familiar with passenger services? *(Wie vertraut sind Sie mit dem Ablauf der Flugabfertigung?)*	R
Day-related situational factors [DSIT]	d_wti	My idle time during passenger services was too long. *(Die Wartezeit während der Flugabfertigung empfand ich als zu lang.)*	F
	d_tip	To what extent did you feel pressed for time during passenger services? *(Wie stark standen Sie während der Flugabfertigung unter Zeitdruck?)*	F
	d_dis	The distance I had to travel during passenger services was too long. *(Die zurückgelegte Wegstrecke während der Flugabfertigung empfand ich als zu lang.)*	F
Satisfaction with orientation opportunities [SORI]	o_ari	How many times did you have to ask for directions and/or information during passenger services? *(Wie häufig mussten Sie während der Flugabfertigung nach dem Weg oder nach Hilfe fragen?)*	R
	o_sig	To what extent were signs giving directions or identifying locations helpful? *(Wie hilfreich empfanden Sie die allgemeine Beschilderung (d. h. die Wegbeschreibung oder Ortsbeschriftung) im Terminal?)*	R
	o_iwa	To what extent were you satisfied with those information systems located in the waiting area? *(Wie zufrieden waren Sie mit dem Informationssystem in den Wartebereichen?)*	R
	o_ori	To what extent was it easy for you to orient yourself at the airport? *(Wie leicht fiel es Ihnen sich am Flughafen zu orientieren?)*	R

Table 1. The variables (MVs) used to measure each of the latent variables (LVs) in this study. Column *ToM* indicates the type of measurement employed for the MV, where *R* denotes *reflective* and *F* means *formative* measurement. See footnote g for a very brief distinction of these different types of measurement. Please note: To ensure readability, all questions were translated from German to English; the German-language version is given in italics. The English language translations were first published in Kattenbeck et al. 2018.

4.2. *Sample Size Considerations*

While variance and covariance based approaches to estimate structural equation models exist, PLS-based (i.e. variance-based) techniques are known to be less demanding in terms of sample size than covariance-based estimation techniques. They have, therefore, a long-lasting tradition in being applied for smaller samples (see Ringle, Sarstedt, and Straub 2012 for a review of sample sizes). According to well-established rules of thumb (see Barclay, Higgins, and Thompson 1995 and Hair Jr. et al. 2014) $N = 30$ participants would be sufficient for the structural model proposed. This number equals 10 times the number formative causes for DSIT, which is the only formatively modeled construct. A sample of size $N = 100$ can, therefore, be regarded to be sufficient to detect large effects, at least. Furthermore, it is sufficiently large for logistic regression analyses (see section 5 below).

4.3. *Experimental Setup*

Data collection took place between May 24th, 2017 and May 28th, 2017. No technical errors in the airport infrastructure or other air traffic related faults (e.g. personnel on strike), which might have had an impact on the results, were reported throughout this week. The data collection for treatment group participants took place over several days of the week and at different times of day to reduce potential biases. They were recruited at the airport's metro station by posters making them aware of the study. By using a single pick-up location for AIRBOT users, we ensured equivalent starting conditions for all treatment group participants. Persons expressing their interest were required to be departing air passengers, have an active Facebook account, and Facebook's Messenger app installed on their own phone. The experimenter connected their phone to AIRBOT, explained the kind of assistance AIRBOT is able to give (including sending a first message to AIRBOT) and asked participants to answer the post-task survey preferably before take-off.

The participants in the control group (non-AIRBOT users), however, were acquired on May 26th 2017 in the security restricted area of all terminals[h]. This means, they were contacted after they successfully arrived at their gate, i.e. these participants already had completed passenger services. Using this point in time reduces potential bias in our data. Control-group participants were asked to answer the survey using a tablet device provided by the experimenter. Neither control nor treatment group participants were compensated for their participation.

4.4. *Describing Participants*

Table 2 presents summary statistics of both groups. The sample size for the treatment group is $N_t = 51$ (24 female) and for the control group $N_c = 50$ (26 female) [i]. Based on counts and a subsequent log-likelihood test (G-Test) the figures suggest that the groups are not

[h]Due to the security restrictions which applied at this particular airport in May 2017, a researcher was allowed to enter the secured area when being accompanied by an airport staff member. Due to this requirement, however, it was impossible to collect control group data on different days.

[i]The difference in sample size is a result of technical issues for a single control-group-participants.

different in terms of how often the participants visited the airport nor whether they are traveling for business or casual leisure reasons. With respect to *age*, however, both groups differ significantly with people being younger in the treatment group. The requirement of having an active Facebook account for treatment group participants may explain this difference. It reflects the known age group distribution of Facebook users in general with the majority of users being between 13 and 34 years of age (see Statista 2017). In order to exclude formal bias, *age* will be neglected during the analyses, consequently.

Group	**Age**					**Experience**				**Reason**		
	$[20,30[$	$[30,40[$	$[40,50[$	$[50,60[$	$[60,\infty[$	1	2	3	>3	b	c	s
Treatment	33	15	1	2	0	18	13	5	11	6	45	0
Control	24	11	7	6	2	11	12	9	18	7	41	2
	$G=10.768$					$G=4.52$				$G=3.026$		
	$df=2$					$df=3$				$df=2$		
	$p=0.005$**					$p=0.21$n.s.				$p=0.22$n.s.		

Table 2. An overview of the distribution of possibly confounding demographic data variables. *Age* was split into five groups, i.e. persons in their twenties, thirties, forties, fifties and older. The frequency of visits per year is used as a proxy for *experience* with airports. For variable *reason* possible values are **b**usiness, **c**asual, **s**tudent related. The G-Test results indicate no difference across the groups except for age. The significance level of $\alpha = 0.05$ was adjusted according to Bonferroni (Dunn 1959). The figures presented in this table were first published in Kattenbeck et al. 2018.

5. Data analysis

Based on these descriptive statistics, the data analysis conducted comprises three steps:

Step 1: PLS Path Modeling First, the proposed structural equation model is assessed in order to ensure the validity of ad-hoc survey results. Structural equation modeling has — in contrast to multiple regression approaches — the unique capability to take latent variable relationships and measurement model results simultaneously into account. We use PLS Path Modeling as an estimation technique because we are interested in maximizing the variance explained in satisfaction with passenger services.

Step 2: Binary Logistic Regression A binary logistic regression based on latent variable scores obtained through the structural model results is conducted. It will shed light on whether groups differ in terms of the constructs.

Step 3: Group Comparisons Based on the logistic regression results, we conduct pairwise tests to understand the influence of AIRBOT's basic task-tracking capabilities.

5.1. *Step 1: PLS Path Modeling*

5.1.1. *A Rationale for PLS Path Modeling*

While covariance- (see Jöreskog 1971) and variance-based methods (see Wold 1975) to estimate SEMs exist, the variance-based approach, i.e. PLS Path Modeling, is used here for three reasons. First, it is chosen due to the study's goal to assess the influence AIR-BOT shows on the key target construct (see Joe F. Hair, Ringle, and Sarstedt 2011, p. 144), the satisfaction with passenger services (SPAS). Second, PLS estimation can easily handle formative and reflexive measurement model parts simultaneously; this is a key ability due to the fact that DSIT is modeled formatively. PLS Path Modeling shows, finally, more statistical power, in particular with smaller sample sizes (see Joe F. Hair, Hult, et al. 2017, p. 118). In conducting the PLS-SEM analysis, we follow the suggestions by Henseler, Hubona, and Ray (2016), i.e. we will first assess the overall goodness-of-fit of the model, then move on to the measurement model results and will, finally, assess the direct, indirect and total effects of the structural model part. The estimations reported in this paper are based on this procedure and calculated with ADANCO software (Composite Modeling GmbH & Co. KG 2015) using the consistent PLS Path Modeling algorithm (see Dijkstra and Henseler 2015a; Dijkstra and Henseler 2015b).

5.1.2. *Overall Model Fit*

According to Dijkstra (2014) assessing overall model fit can reveal misspecification with respect to both structural and measurement model parts. The standardized root mean square residual (SRMR, Hu and Bentler 1998; Hu and Bentler 1999), suggests a reasonable approximate model fit ($SRMR = 0.03$) as well as the unweighted least squares discrepancy (d_{ULS}, see Dijkstra and Henseler 2015a) does ($d_{ULS} = 0.04$). Based on these results, it is reasonable to assess the measurement model parts next as the structural model results rely on these.

5.1.3. *Assessing the Measurement Model*

In case of the model suggested here, single indicator measurement is used for LVs SPAS and FPAS, i.e. no measurement model fit must be assessed for both latent variables. A composite model was devised for DSIT and, therefore, the multicollinearity of the measured variables should be low (see Joe F. Hair, Ringle, and Sarstedt 2011). It is assessed based on the variance inflation factor (VIF). A commonly used threshold is $VIF \leq 5$ (see ibid.) which is met for all MVs ($\text{VIF}(d\_wti) = 2.07, \text{VIF}(d\_dis) = 2.20, \text{VIF}(d\_tip) = 1.88$). Finally, SORI was modeled to be reflexively measured and, consequently, its reliability needs to be assessed. SORI shows a high reliability as both Dijkstra-Henseler's ρ_A (Dijkstra and Henseler 2015a) and Cronbach's α (Cronbach and Meehl 1955) are greater than 0.7 ($\rho_A = 0.92$ and $\alpha_{Cronbach} = 0.92$, respectively). Similarly, convergent validity based on the average variance extracted (Fornell and Larcker 1981, AVE) is given because $\text{AVE}(SORI) = 0.75$ is well above the threshold of 0.5 commonly used (see Joe F. Hair, Ringle, and Sarstedt 2011). In terms of discriminant validity, however, SORI turns out to share meaning with SPAS. The HTMT (see Henseler, Ringle, and Sarstedt 2015, p. 121) and the Fornell-Larcker criterion (see Fornell and Larcker 1981) are both not met: The HTMT-value of $\text{HTMT} = 0.938$ does not differ significantly from one and the squared correlation between SORI and SPAS equals 0.80, whereas the average variance extracted in

SORI equals 0.75. However, in the light of single indicator measurement for SPAS, which inevitably increases random error and the fact that cross loading requirements are met we proceed with the structural model analysis.

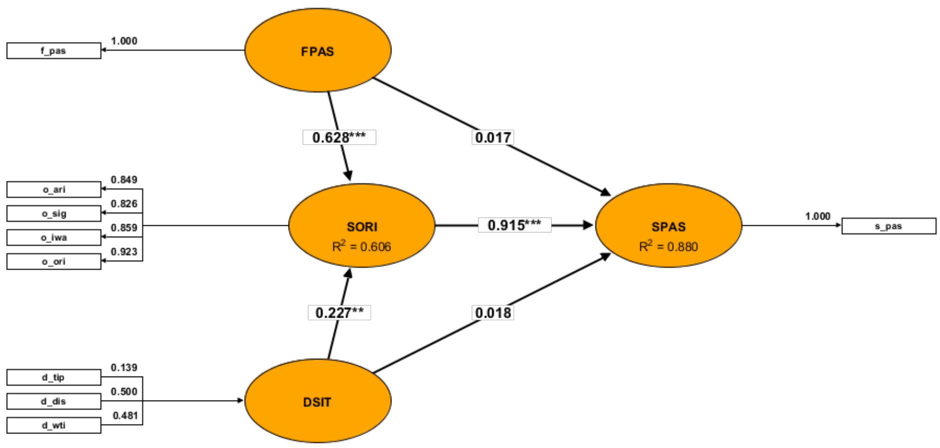

Fig. 3. The estimated model, indicating R^2-values in exogenous LVs and direct effects between all LVs. The figures attached to MVs of factor SORI are outer loadings, those of DSIT are outer weights. ** indicates $p < .01$ and *** means $p < .001$.

5.1.4. *Assessing Structural Relationships*

88 % of the variance present in the key target construct *satisfaction with passenger services* (SPAS) can be explained through the structural relationships proposed ($R^2 = 0.880$, see Figure 3). Feeling oriented (SORI) is most important for SPAS (see Table 3). The second largest effect size was found for the impact *familiarity with passenger services* (FPAS) has on SORI, which is an indicator for passengers getting used to the passenger services procedure.

The direct effect of *day-related situational aspects* (DSIT) on SPAS, however, is fully mediated as well as the direct effect of *familiarity* (FPAS) on *satisfaction* (SPAS) is. This means, both *familiarity* (FPAS) and *day-related situational aspects* (DSIT) contribute significantly to the *feeling of being oriented* (SORI) because both aspects can explain approximately 60 % of the variance present in SORI ($R^2 = 0.606$).

5.1.5. *Contrasting Theoretical Models — Tree-augmented Naïve Bayes Results*

Generally speaking, at least two different structural models exist for which a given dataset fits equally well (see Joseph F. Hair et al. 2010, p. 647). Instead of constructing a second theoretical model, we follow the arguments given in Wu 2010 and use, first, Tree-augmented Naïve Bayes to learn a structural model from the data gathered using WEKA

Effect	Direct	Indirect	Total	Cohen's f^2
SORI $\rightarrow$ SPAS	0.915 [0.744;1.087]	NA	0.915 [0.744;1.087]	2.760
DSIT $\rightarrow$ SPAS	0.0178 [-0.119;0.135]	0.208[0.0678;0.373]	0.226 [0.075;0.371]	0.002
DSIT $\rightarrow$ SORI	0.227 [0.079;0.380]	NA	0.227 [0.079;0.380]	0.091
FPAS $\rightarrow$ SPAS	0.017 [-0.151;0.172]	0.575[0.377;0.745]	0.592 [0.392;0.727]	0.001
FPAS $\rightarrow$ SORI	0.628 [0.452;0.758]	NA	0.628 [0.452;0.758]	0.690

Table 3. Direct, indirect and total effects of the theoretical model; the 95%-CIs, which are based on $B = 10000$ bootstrapping resamples are given in square brackets.

(Frank, Hall, and Witten 2016). Tree-based search is particularly useful with respect to learning a structural model from data (see Wu 2010, p. 136 for more technical details) as a directed a-cyclic graph with a single top-level node mirrors the goal of our analysis in judging the impact AIRBOT shows on SPAS as a key target construct. In order to reflect the nature of the measurement modes, we use a sum score to estimate the DSIT score and the mean to estimate the SORI score. The effects of the resulting structural model are, second, assessed on consistent PLS Path Modeling using, again, ADANCO. Figure 4 presents the results.

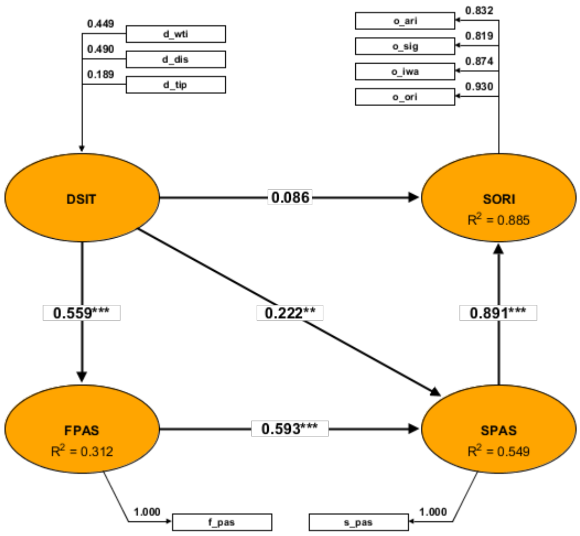

Fig. 4. The PLS-based estimation of the data-driven structural model. ** indicates $p < .01$ and *** means $p < .001$ for the significance of effects, which is based on $B = 10000$ bootstrapping resamples.

The overall model fit of this empirical structural model ($SRMR = 0.03$ and $d_{ULS} = 0.045$) is not worse than the values found for the theoretical model. Yet, comparing the empirical model to its theoretical counterpart, three differences occur: First, there is no direct effect of *familiarity with passenger services* (FPAS) on *satisfaction with orientation* (SORI); second, there is now a direct effect of *day-related situational effects* (DSIT) on

FPAS; third, *satisfaction with passenger services* (SPAS) has now an impact on SORI (and not vice versa). These results show two implausibilities: As orientation is needed to successfully finish the passenger services workflow, the direction of the effect suggested on empirical data is implausible due to time order. In addition to that, the empirical model explains less variance in the *satisfaction with passenger services* (SPAS) variable, which is the key target construct of this study. We, therefore, proceed with our analysis based on the theoretical model. Given the reasonable fit of the theoretical model revealed throughout step 1, the regression analysis and the group comparisons are based on the unstandardized latent variable scores. In doing so, we take the structural relationships and the measurement model results into account instead of using means based on equally weighted variables.

5.2. *Step 2: Binary Logistic Regression*

The assumption of independence of observations was given by the between-subject design and the dependent variable (DV) *group* is binary in nature by experimental design, too. The independent variables (IVs) show low Variance Inflation Factor values ($VIF(SPAS) = 1.4$, $VIF(SORI) = 1.30$, $VIF(DSIT) = 1.15$, $VIF(FPAS) = 1.18$). Moreover, the sample size is reasonably large according to the commonly applied rules of thumb[j].

Based on the fact that these assumptions were met, a reasonable amount of variance can be explained by the model ($McFadden - Pseudo - R^2$ = 0.58) using *group* as a dependent variable, thereby suggesting an acceptable model fit. Table 4 shows that both *day-related situational aspects* (DSIT) and *satisfaction with orientation opportunities* (SORI) have a significant effect: A unit increase in DSIT and SORI increases the log odds of group by 2.27 and 2.61, respectively.

	Unstandardized Coefficients		z	95% CI for B
	B	Std. Err.		
Intercept	-13.86	4.79	0.004**	$[-25.54;-6.11]$
SPAS	-1.24	1.25	0.322	$[-3.76;1.24]$
SORI	2.27	2.54	0.011*	$[0.71;4.25]$
DSIT	2.61	4.63	0.000***	$[1.68;3.95]$
FPAS	0.33	0.66	0.507	$[-0.61;1.36]$

Table 4. The unstandardized coefficients for the logit model, including their standard errors, 95% confidence intervals and significance based on Wald's *z*; * indicates $p < .05$, ** means $p < .01$ and *** equals $p < .001$.

Using χ^2-tests reveals a significant decrease of the residual deviance for factors DSIT and SORI. Hence, empirical evidence is provided to base group comparisons solely on these two factors.

[j]Peduzzi et al. (1996) suggest 10 observations per IV divided by the smallest expected probability, i.e. $N_{min} = 10 \times 4/.5$.

	Df	Dev. Resid.	Df	Resid. Dev.	p-value
NULL			100	140.006	
SPAS	1	1.794	99	138.212	0.181
SORI	1	11.549	98	126.663	0.0007***
DSIT	1	67.835	97	58.828	$< 2.2 \times 10^{-16}$***
FPAS	1	0.454	96	58.374	0.500

Table 5. The difference between the null deviance and the residual deviances for the predictors. Significance values (*** equals $p < .001$) are based on χ^2-tests. The figures provide evidence that SORI and DSIT are very important factors. This table was first published in Kattenbeck et al. 2018.

5.3. *Step 3: Group Comparisons*

Due to non-normality of the data ($|excess| < 4$ and $|skewness| \approx 0$) a one-sided Mann-Whitney-Wilcoxon-Test was performed using a significance level of $\alpha = 0.05$, which was, again, corrected according to Bonferroni.

First, the tests reveal a large effect of AIRBOT on factor DSIT, i.e. AIRBOT users have a highly significantly decreased perceived waiting time, walking distance and feel less time pressure ($\tilde{x}_{bot} = 2.43$, $\tilde{x}_{control} = 4.43$, $Z = -7.38$, $p < 7.8e^{-14}$, $r = 0.74$). This result indicates that AIRBOT is effective in terms of how passengers experience their travel task at the airport. AIRBOT users experience less workload to complete required activities and in parallel feel entertained better during waiting periods due to the fact that the perceived waiting time is decreased.

Second, a medium-sized significant difference was found regarding the satisfaction with orientation opportunities provided ($\tilde{x}_{bot} = 4.0$, $\tilde{x}_{control} = 4.25$, $Z = -3.38$, $p < 0.01$, $r = 0.34$). This means, people using AIRBOT feel less oriented. This result is somewhat unexpected and we will return to it in the discussion section.

6. Discussion

In applying PLS Path Modeling the user study presented yields three major results:

(1) The ad-hoc survey used can explain a large proportion of the variance present in *satisfaction with passenger services*.
(2) The *feeling of being oriented* mediates the impact of *familiarity with passenger services* and *day-related situational factors* on the *satisfaction with passenger services*.
(3) The day-related situational factors are not equally important.

The further analysis of AIRBOT's impact on *satisfaction with passenger services* based on the latent variable scores provides evidence that

(1) AIRBOT has a positive impact on satisfaction with passenger services and that
(2) the use of a personalized information companion can have adverse effects on non-personalized orientation facilities.

Based on these results, the discussion section has two parts. First, we discuss these major findings of the in-situ experiment and comment on its implications. We then move on to limitations which apply to the work presented.

6.1. *Discussing the Findings and Its Implications*

Comparing the structural model results to a data driven model (see Figure 4) suggests that the theoretical model proposed is reasonable and, hence, support the validity of the results. While SORI mediates the effects DSIT and FPAS have on the *satisfaction with passenger services*, DSIT and FPAS account for approximately 60% of the variance present in SORI; this indicates the importance of these two variables for SORI. The feeling of orientation is influenced by day-related situational factors and the familiarity with services, but not fully explained through these factors. This leaves room for further investigations on factors such as the Big Five personality traits and/or sense of direction may have. Among the causal-formative indicators of DSIT, *felt time pressure* (d_tip) is less important than *perceived waiting time* (d_wti) and *walking distance* (d_dis) are. This means, reducing walking distance and waiting time can both have an equally sized positive impact and are, therefore, both a worthwhile goal for adjustment by airports.

From a methodological point of view these results provide evidence that the use of structural equation modeling can benefit user studies. Taking multiple relationships between latent variables simultaneously into account is a unique capability of the method and provides valuable insights. In general, structural equation modeling is a useful means to achieve sound statistical results when using ad-hoc surveys, which are widely used in HCI studies. The possibility to base subsequent analyses on factor scores is a key advantage in terms of measurement theory, as compared to using simple means or sums: Despite the fact that error variances cannot be modeled in PLS Path Modeling (see Vinzi et al. 2010, p. 48), taking structural and measurement model relationships into account increases the reliability of the measurement by loosening major assumptions of classical test theory in surveys (see Feldt and Brennan 1989 for a general overview of these).

Overall, the results of this field study suggest that basic task-tracking capabilities are sufficient to induce a positive effect regarding the users' perceived experience of service provisions offered in public spaces such as airports. Compared to control group participants, the use of the implemented chatbot system makes a difference regarding the users' perception of their passenger service experience. AIRBOT users experience less workload to complete all required activities. In addition, the design idea of AIRBOT is also effective in terms of how passengers experience their travel task at the airport because AIRBOT report on decreased waiting time between subsequent activities. These findings are in line with work exposing the importance and benefit of personalized information provision in other public settings (Schmidt-Belz et al. 2003; Nugent et al. 2015). In terms of research practice this provides evidence that a thorough understanding of the users' task is a key element in enabling the design of simple, yet highly effective algorithmic solutions. Understanding and modeling tasks users are required to perform in a certain context and implementing only basic task-related functionalities based on these modelled tasks may, thus, serve as design principle that guides practitioners to provide mobile assistance in public spaces successfully and with minimal effort.

Despite its usefulness, AIRBOT users were at the same time less satisfied with signs, displays or other stationary information systems offered by the airport and their feeling of orientation. This result suggests that personalized information systems can affect the perception of non-personalized information negatively. This explanation is in line with general work that indicates that smartphone usage has an effect on the way users walk and where they look at (Timmis et al. 2017) and has an effect on the user's awareness of surroundings (M.-I. B. Lin and Huang 2017). This is the most important implication our work provides. It is a significant insight that task-tracking is a suitable means to assist people in public spaces. Providing users with assistance on their smartphones, however, may distract them from their surroundings. Ultimately, this can lead to a decreased feeling of orientation within public spaces. Thus, challenges in terms of pushing information at users arise, in particular because recent research also suggests the degree of distraction is positively correlated to the immersiveness of the screen content (Haga et al. 2015). This has major design implications for future UI designs in public space route guidance scenarios. One remedy to these negative effects may be to remind users by visual cues on the smartphone screens to orient themselves in their current environment.

6.2. *Limitations*

Three limitations apply to the work presented, but do not harm the reported results. The first of these applies from the perspective of measurement theory. We used single item measurement to reduce the time effort for participants; in doing so, the random error present in our measurement may have been increased (see Diamantopoulos et al. 2012). Having said this, random error is expected to decrease the likeliness of significant results, but has been achieved nonetheless. The second limitation relates to the difference in age distribution between control and treatment group participants. Prior evidence suggests that satisfaction with airport service quality is positively correlated with age (see Clemes et al. 2008, p. 59). Thus, while the significant difference reflects the age distribution patterns of Facebook and non-Facebook users, it may have resulted in underpredicting the positive effect AIRBOT has because participants of our study were younger in the treatment group (AIRBOT users). A third limitation may result from the fact that control group participants were required to have an active Facebook account and the Facebook Messenger app installed on their smartphone. This prerequisite may have induced bias with respect to the influence found for *day-related situational factors* (DSIT): Social media users may value the AIRBOT capabilities more because they might know similar capabilities from other digital contexts. This may have lowered the size and significance of the found effect.

7. Conclusion and Future Work

This paper reports on a least-effort study about the impact of task-tracking capabilities in chat-based mobile information companions. To this end, we implemented AIRBOT, a chat-based smartphone companion which provides assistance to air passengers during passenger services. Given the goal of a least-effort study, AIRBOT only provides minimal functionality to track passengers between arriving at an airport and boarding their plane; its main feature is the ability to monitor changes in workflow states over time and, thereby, to provide information proactively at the right time. Understanding this workflow is, consequently, a key element of the system design. Conducting interviews with both information personnel and passengers revealed that it is most important that the application can assist users in handling tasks at the right point in time or within the required time frame. Based on an in-situ user study contrasting AIRBOT users and non-AIRBOT users, we provide evidence that task-tracking capabilities, albeit very basic by design, have an impact on the way passengers perceive waiting time, covered walking distances and time pressure at airports. While AIRBOT's workflow is specific for the airport domain, the general approach employed can be transferred to other domains which require users to perform a task by adhering to a predefined workflow. In scenarios like these the integrating basic task-tracking capabilities in proactive and personalized mobile chatbots is a suitable means to increase user satisfaction.

Given the results and bearing the limitations in mind, we plan four different strands of future work. We will, first, conduct a second, qualitative field study. Collecting qualitative data from AIRBOT users will, on one hand, foster our understanding how cooperative workflow models have to be designed and implemented. On the other hand, this data will allow to investigate the reasons for the dissatisfaction with non-personalized information. Second, in order to fully personalize the assistance AIRBOT can give, we work on methods to determine a user's cognitive state in comparison to the workflow implemented. This is an important step towards personalization. We are, third, interested in detecting user struggles in fulfilling a specific workflow task by exploiting interaction data in a continuous manner. Related to this, we are, finally, working on algorithms which are capable to take a user's task-state into account: There might be, for example, occasions when a personal information companion cannot resolve a problem appropriately and, therefore, needs to redirect its users to information personnel.

Acknowledgements

We would like to thank our user study participants and interview partners, in particular the eight information experts. Many thanks to the InfoGate Information Systems GmbH for their support in conducting this research. We are grateful to the anonymous reviewers and our colleague, David Elsweiler, all of whom helped to improve the paper through their valuable comments.

References

AbuShawar, Bayan and Eric Atwell (2013). "Automatic Extraction of Chatbot Training Data from Natural Dialogue Corpora". In: *RE-WOCHAT: Workshop on Collecting and Generating Resources for Chatbots and Conversational Agents-Development and Evaluation Workshop Programme (May 28 th, 2016)*, p. 29.

Alt, Florian and Julia Vehns (2016). "Opportunistic deployments: challenges and opportunities of conducting public display research at an airport". In: *Proceedings of the 5th ACM International Symposium on Pervasive Displays*. ACM, pp. 106–117.

Andrews, Pierre, Suresh Manandhar, and Marco De Boni (2008). "Argumentative Human Computer Dialogue for Automated Persuasion". In: *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*. SIGdial '08. Columbus, Ohio: Association for Computational Linguistics, pp. 138–147.

Asri, Layla El et al. (2017). "Frames: a corpus for adding memory to goal-oriented dialogue systems". In: *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Saarbrücken, Germany, August 15-17, 2017*, pp. 207–219. URL: https://aclanthology.info/papers/W17-5526/w17-5526.

Audiotranskription (2018). *f4transkript. Zeitsparend transkribieren*. URL: https://www.audiotranskription.de/f4.

Awori, Kagonya et al. (2012). "Flytalk: social media to meet the needs of air travelers". In: *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. ACM, pp. 1769–1774.

Barclay, D. W., C. A. Higgins, and R. Thompson (1995). "The partial least squares approach to causal modeling: personal computer adoption and use as illustration". In: *Technology Studies* 2.2, pp. 285–309.

Biermann, Alan W. et al. (1997). "Goal-Oriented Multimedia Dialogue with Variable Initiative". In: *Foundations of Intelligent Systems, 10th International Symposium, ISMIS '97, Charlotte, North Carolina, USA, October 15-18, 1997, Proceedings*. Ed. by Zbigniew W. Ras and Andrzej Skowron. Vol. 1325. Lecture Notes in Computer Science. Springer, pp. 1–16. URL: https://doi.org/10.1007/3-540-63614-5%5C_1.

Brandtzaeg, Petter Bae and Asbjørn Følstad (2018). "Chatbots: Changing User Needs and Motivations". In: *Interactions* 25.5, pp. 38–43. ISSN: 1072-5520. URL: http://doi.acm.org/10.1145/3236669.

Byström, Katriina and Kalervo Järvelin (1995). "Task complexity affects information seeking and use". In: *Information processing & management* 31.2, pp. 191–213.

Callejas-Rodrıguez, Ángel et al. (2016). "From dialogue corpora to dialogue systems: Generating a chatbot with teenager personality for preventing cyber-pedophilia". In: *International Conference on Text, Speech, and Dialogue*. Springer, pp. 531–539.

Case, Donald O. and Lisa M. Given (2016). *Looking for Information*. 4th ed. Bingley, UK: Emerald.

Chandel, Sonali et al. (2018). "Chatbot: Efficient and Utility-Based Platform". In: *Science and Information Conference*. Springer, pp. 109–122.

Church, Karen and Barry Smyth (2009). "Understanding the intent behind mobile information needs". In: *Proceedings of the 14th international conference on Intelligent user interfaces*. ACM, pp. 247–256.

Clemes, Michael D et al. (2008). "An empirical analysis of customer satisfaction in international air travel". In: *Innovative Marketing* 4.2, pp. 50–62.

Composite Modeling GmbH & Co. KG (2015). *ADANCO 2.0*. last access on May 25th, 2017. URL: http://www.composite-modeling.com/ (visited on 03/05/2017).

Cox, Andrew M et al. (2017). "What everybody knows: embodied information in serious leisure". In: *Journal of Documentation* 73.3, pp. 386–406.

Cronbach, L. J. and P. E. Meehl (1955). "Construct Validity in Psychological tests". In: *Psychological Bulletin* 52.4, pp. 281–302.

Davies, Nigel et al. (2014). "Personalisation and privacy in future pervasive display networks". In: *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, pp. 2357–2366.

De Gasperis, Giovanni, Isabella Chiari, and Niva Florio (2013). "AIML knowledge base construction from text corpora". In: *Artificial intelligence, evolutionary computing and metaheuristics*. Springer, pp. 287–318.

DeVellis, Robert F. (2012). *Scale Development. Theory and Applications*. 3rd. Thousand Oaks, CA et al.: SAGE Publications.

Diamantopoulos, Adamantios et al. (2012). "Guidelines for choosing between multi-item and single-item scales for construct measurement: a predictive validity perspective". In: *Journal of the Academy of Marketing Science* 40, pp. 434–449.

Dijkstra, Theo K. (2014). “PLS’ Janus Face – Response to Professor Rigdon’s ’Rethinking Partial Least Squares Modeling: In Praise of Simple Methods’”. In: *Long Range Planning* 47.3, pp. 146–153.

Dijkstra, Theo K. and Jörg Henseler (2015a). “Consistent and asymptotically normal PLS estimators for linear structural equations”. In: *Computational Statistics & Data Analysis* 81, pp. 10–23. ISSN: 0167-9473.

— (2015b). “Consistent Partial Least Squares Path Modeling”. In: *Management Information Systems Quarterly* 39.2, pp. 297–316.

Dirin, Mitra, Amir Dirin, and Teemu H. Laine (2015). “User-centered Design of a Context-aware Nurse Assistant (CANA) at Finnish Elderly Houses”. In: *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication*. IMCOM ’15. Bali, Indonesia: ACM, 39:1–39:8.

Dunn, Olive Jean (1959). “Estimation of the Medians for Dependent Variables”. In: *The Annals of Mathematical Statistics* 30.1, pp. 192–197.

Facebook (2017). *Messenger-Platform*. online. last access Sept 2nd, 2017. URL: https://developers.facebook.com/docs/messenger-platform.

Feldt, L. S. and R. L. Brennan (1989). “Reliability”. In: *Educational measurement*. Ed. by R. L. Linn. 3rd. Phoenix, AZ: Ornyx, pp. 105–146.

Følstad, Asbjørn and Petter Bae Brandtzæg (2017). “Chatbots and the New World of HCI”. In: *Interactions* 24.4, pp. 38–42. ISSN: 1072-5520. URL: http://doi.acm.org/10.1145/3085558.

Fornell, Claes G. and David F. Larcker (1981). “Evaluating Structural Equation Models with Unobservable Variables and Measurement Error”. English. In: *Journal of Marketing Research* 18.1, pp. 3–50. ISSN: 00222437.

Frank, Eibe, Mark A. Hall, and Ian H. Witten (2016). *The WEKA Workbench. Online Appendix for “Data Mining: Practical Machine Learning Tools and Techniques”*. 4th. Morgan Kaufmann.

FraportAG (2018). *FRAnky. Erkunden Sie mit mir den Flughafen!* URL: frankfurt-airport.com/de/reisen/am-flughafen.detail.suffix.html/article/reisen/services-a-z/franky.html.

Freund, Luanne (2015). “Contextualizing the information-seeking behavior of software engineers”. In: *Journal of the Association for Information Science and Technology* 66.8, pp. 1594–1605.

Haga, Shigeru et al. (2015). “Effects of using a smart phone on pedestrians’ attention and walking”. In: *Procedia Manufacturing* 3, pp. 2574–2580.

Hair Jr., Joseph F. et al. (2014). *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*. Los Angeles et al.: SAGE Publications.

Hair, Joe F., Tomas Hult, et al. (2017). “Mirror, mirror on the wall: a comparative evaluation of composite-based structural equation modeling methods”. In: *Journal of the Academy of Marketing Science* 45, pp. 616–632. DOI: 10.1007/s11747-017-0517-x.

Hair, Joe F., Christian M. Ringle, and Marko Sarstedt (2011). “PLS-SEM: Indeed a Silver Bullet”. In: *Journal of Marketing Theory and Practice* 19.2, pp. 139–151.

Hair, Joseph F. et al. (2010). *Multivariate Data Analysis. A Global Perspective*. 7th. Global Edition. Upper Saddle River, NJ: Person Education.

Henseler, Jörg, Geoffrey Hubona, and Pauline Ash Ray (2016). “Using PLS path modeling in new technology research: updated guidelines”. In: *Industrial Managment & Data Systems* 116.1, pp. 2–20.

Henseler, Jörg, Christian M. Ringle, and Marko Sarstedt (2015). “A new criterion for assessing discriminant validity in variance-based structural equation modeling”. In: *Journal of the Academy of Marketing Science* 43.1, pp. 115–135.

Hu, Li and Peter M. Bentler (1998). “Fit indices in covariance structura modeling: Sensitivity to underparameterized model misspecification”. In: *Psychological Methods* 3.4, pp. 424–453.

Hu, Li and Peter M. Bentler (1999). “Cutoff Criteria for fit indexes in covariance structure analysis: Conventional Criteria versus new alternatives”. In: *Structural Equation Modeling: A Multidisciplinary Journal* 6.1, pp. 1–55.

IATA (2016). *2016 Global Passenger Survey Results Highlights*. URL: http://www.iata.org/publications/store/Documents/GPS-2016-Highlights-Final.pdf.

Ilievski, Vladimir (2018). “Building Advanced Dialogue Managers for Goal-Oriented Dialogue Systems”. In: *CoRR* abs/1806.00780. arXiv: 1806.00780. URL: http://arxiv.org/abs/1806.00780.

Iwai, Takafumi et al. (2006). “Paravision: The Entertaining Visualizer in Public Space”. In: *Proceedings of the 2006 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology*. ACE ’06. Hollywood, California, USA: ACM. URL: http://doi.acm.org/10.1145/1178823.1178928.

Jain, Mohit et al. (2018). “Evaluating and Informing the Design of Chatbots”. In: *Proceedings of the 2018 Designing Interactive Systems Conference*. DIS ’18. Hong Kong, China: ACM, pp. 895–906. URL: http://doi.acm.org/10.1145/3196709.3196735.

Jarvis, Cheryl B., Scott B. MacKenzie, and Phil M. Podsakoff (2003). “A Critical Review of Construct Indicators and Measurement Model Misspecification in Marketing and Consumer Research”. In: *Journal of Consumer Research* 30, pp. 199–218.

Johnson, J David and Hendrika Meischke (1993). “A comprehensive model of cancer-related information seeking applied to magazines”. In: *Human Communication Research* 19.3, pp. 343–367.

Jöreskog, Karl (1971). “Simultaneous Factor Analysis in Several Populations”. In: *Psychometrika* 36, pp. 409–426.

Kattenbeck, Markus et al. (2018). “Airbot: Using a Work Flow Model for Proactive Assistance in Public Spaces”. In: *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. MobileHCI ’18. Barcelona, Spain: ACM, pp. 213–220.

Kuhlthau, Carol C (1993). “A principle of uncertainty for information seeking”. In: *Journal of documentation* 49.4, pp. 339–355.

Lee, Hyunjung, Harksoo Kim, and Jungyun Seo (2013). “An Integrated Neural Network Model for Domain Action Determination in Goal-Oriented Dialogues”. In: *JIPS* 9.2, pp. 259–270. URL: https://doi.org/10.3745/JIPS.2013.9.2.259.

Lin, Ming-I Brandon and Yu-Ping Huang (2017). “The impact of walking while using a smartphone on pedestrians’ awareness of roadside events”. In: *Accident Analysis & Prevention* 101, pp. 87–96.

Liu, Bingquan et al. (2018). “Content-Oriented User Modeling for Personalized Response Ranking in Chatbots”. In: *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 26.1, pp. 122–133. ISSN: 2329-9290. URL: https://doi.org/10.1109/TASLP.2017.2763243.

Liu, Yilin Elaine et al. (2016). “GatePal–Universal Design for Airport Navigation to Allow Departing Travellers to Stay Informed”. In: *International Conference on Universal Access in Human-Computer Interaction*. Springer, pp. 586–594.

Lueg, Christopher P (2014). “Characteristics of human perception and their relevance when studying information behavior”. In: *Journal of Documentation* 70.4, pp. 562–574.

Ma, Shang-Pin and Ching-Ting Ho (2018). “Modularized and Flow-Based Approach to Chatbot Design and Deployment.” In: *Journal of Information Science & Engineering* 34.5.

Marcella, Rita, Tracy Pirie, and Hayley Lockerbie (2013). “The information seeking behaviour of oil and gas industry workers in the context of health, safety and emergency response: a discussion of the value of models of information behaviour.” In: *Information Research Volume 18 Part 3*.

Matthews, K. (2015). *7 Amazing Airport Apps That Make Your Flight Suck Less*. URL: https://www.afar.com/magazine/7-amazing-airport-apps-that-make-your-flight-suck-less.

Müller, Jörg, Florian Alt, et al. (2010). “Requirements and design space for interactive public displays”. In: *Proceedings of the 18th ACM international conference on Multimedia*. ACM, pp. 1285–1294.

Müller, Jörg, Robert Walter, et al. (2012). “Looking glass: a field study on noticing interactivity of a shop window”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, pp. 297–306.

Nugent, Tim et al. (2015). “”It’s Useless for That”: Finding, Frustration, and Fun with Mobile Technology in Outdoor Markets”. In: *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*. ASIST ’15. St. Louis, Missouri: American Society for Information Science, 56:1–56:9.

O’Brien, Chris (2016). *Facebook Messenger chief says platform’s 34,000 chatbots are finally improving user experience*. online. last access on Sept 2nd, 2017. URL: `https://venturebeat.com/2016/11/11/facebook-messenger-chief-says-platforms/34000-chatbots-are-finally-improving-user-experience/`.

Olsson, Michael (2016). “Making sense of the past: The embodied information practices of field archaeologists”. In: *Journal of Information Science* 42.3, pp. 410–419.

Peduzzi, P. et al. (1996). “A simulation study of the number of events per variable in logistic regression analysis”. In: *Journal of Clinical Epidemiology* 49.12, pp. 1373–1379.

Pronin, S. (2018). *App in the Air. Personal flight assistant*. URL: `https://www.%5C-appintheair.mobi/`.

Raab-Steiner, Elisabeth and Michael Benesch (2015). *Der Fragebogen: Von der Forschungsidee zur SPSS/PASW-Auswertung*. 4., aktualisierte und überarbeitete. Wien: Facultas-Verlag.

Radaha, Timothy R and Mary E Johnson (2013). “Mobile Indoor Navigation Application for Airport Transits”. In:

Ringle, Christian M., Marko Sarstedt, and Detmar Straub (2012). “A Critical Look at the Use of PLS-SEM in MIS Quarterly”. In: *MIS Quarterly (MISQ)* 36.1, pp. iii–xiv.

Rohrmann, Bernd (1978). “Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung”. In: 9, pp. 222–245.

Sato, Munehiko et al. (2010). “Highly Integratable Large-scale Displays for Public Spaces”. In: *Proceedings of the 12th ACM International Conference Adjunct Papers on Ubiquitous Computing - Adjunct*. UbiComp ’10 Adjunct. Copenhagen, Denmark: ACM, pp. 427–428. URL: `http://doi.acm.org/10.1145/1864431.1864470`.

Savolainen, Reijo (1995). “Everyday life information seeking: Approaching information seeking in the context of “way of life””. In: *Library & information science research* 17.3, pp. 259–294.

Schmidt-Belz, Barbara et al. (2003). “Location-based mobile tourist services: first user experiences”. In: ENTER 2003: 10th International Conference on Information Technology in Travel & Tourism.

Seon, Choong-Nyoung et al. (2014). “Improving domain action classification in goal-oriented dialogues using a mutual retraining method”. In: *Pattern Recognition Letters* 45, pp. 154–160. URL: `https://doi.org/10.1016/j.patrec.2014.03.021`.

Serras, Manex, Marıa Inés Torres, and Arantza del Pozo (2017). “Regularized Neural User Model for Goal-Oriented Spoken Dialogue Systems”. In: *Advanced Social Interaction with Agents - 8th International Workshop on Spoken Dialog Systems, IWSDS 2017, Farmington, PA, USA, 6-9 June 2017, revised selected papers*. Ed. by Maxine Eskénazi, Laurence Devillers, and Joseph Mariani. Vol. 510. Lecture Notes in Electrical Engineering. Springer, pp. 235–245. URL: `https://doi.org/10.1007/978-3-319-92108-2%5C_24`.

Sharifi, Matthew, Terry Payne, and Esther David (2006). “Public display advertising based on bluetooth device presence”. In: *Mobile Interaction with the Real World (MIRW 2006)* 52.

Shi, Jiamin and Florian Alt (2016). “The Anonymous Audience Analyzer: Visualizing Audience Behavior in Public Space”. In: *Proceedings of the 2016 CHI Conference Extended Abstracts* *on Human Factors in Computing Systems*. CHI EA '16. Santa Clara, California, USA: ACM, pp. 3766–3769. URL: http://doi.acm.org/10.1145/2851581.2890256.

Sohn, Timothy et al. (2008). "A diary study of mobile information needs". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, pp. 433–442.

Statista (2017). *Facebook: global user age and gender distribution 2017*. last access Sept 2nd, 2017. URL: https://www.statista.com/statistics/376128/facebook-global-user-age-distribution/.

Takanobu, Ryuichi et al. (2018). "A Weakly Supervised Method for Topic Segmentation and Labeling in Goal-oriented Dialogues via Reinforcement Learning". In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.* Ed. by Jérôme Lang. ijcai.org, pp. 4403–4410. URL: https://doi.org/10.24963/ijcai.2018/612.

Timmis, Matthew A et al. (2017). "The impact of mobile phone use on where we look and how we walk when negotiating floor based obstacles". In: *PLoS one* 12.6, e0179802.

Tomitsch, Martin et al. (2014). "Who Cares about the Content? An Analysis of Playful Behaviour at a Public Display". In: *Proceedings of The International Symposium on Pervasive Displays*. PerDis '14. Copenhagen, Denmark: ACM, 160:160–160:165. URL: http://doi.acm.org/10.1145/2611009.2611016.

Väänänen-Vainio-Mattila, Kaisa et al. (2013). "Experiencing Interactivity in Public Spaces (Eips)". In: *CHI '13 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '13. Paris, France: ACM, pp. 3275–3278. URL: http://doi.acm.org/10.1145/2468356.2479665.

Vaccaro, Kristen et al. (2018). "Designing the Future of Personal Fashion". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal QC, Canada: ACM, 627:1–627:11. URL: http://doi.acm.org/10.1145/3173574.3174201.

"PLS Path Modeling: From Foundations to Recent Developments and Open Issues for Model Assessment and Improvement" (2010). In: *Handbook of Partial Least Squares*. Ed. by Vincenzo Esposito Vinzi et al. Springer Handbooks of Computational Statistics. Berlin Heidelberg: Springer-Verlag, pp. 47–82.

Wang, Huei-Ming and Yi-Chung Lin (2000). "Goal-oriented table-driven design for dialogue manager". In: *Sixth International Conference on Spoken Language Processing, ICSLP 2000 / INTERSPEECH 2000, Beijing, China, October 16-20, 2000*. ISCA, pp. 599–602. URL: http://www.isca-speech.org/archive/icslp%5C_2000/i00%5C_2599.html.

Wilson, Tom D (1999). "Models in information behaviour research". In: *Journal of documentation* 55.3, pp. 249–270.

Wold, Herman Ole Andreas (1975). "Path models with latent variables: The NIPALS approach". In: *Quantitative sociology: International perspectives on mathematical and statistical modeling*. Ed. by H. M. Blalock et al. New York: Academic Press, pp. 307–357.

Wu, Wei Wen (2010). "Linking Bayesian networks and PLS path modeling for causal analysis". In: *Expert Systems with Applications: An International Journal* 37.1, pp. 134–139. ISSN: 0957-4174.

Zamora, Jennifer (2017). "I'm Sorry, Dave, I'm Afraid I Can'T Do That: Chatbot Perception and Expectations". In: *Proceedings of the 5th International Conference on Human Agent Interaction*. HAI '17. Bielefeld, Germany: ACM, pp. 253–260. URL: http://doi.acm.org/10.1145/3125739.3125766.