Chapter 13
LAST MINUTE: An Empirical Experiment in User-*Companion* Interaction and Its Evaluation

Jörg Frommer, Dietmar Rösner, Rico Andrich, Rafael Friesen, Stephan Günther, Matthias Haase, and Julia Krüger

Abstract The LAST MINUTE Corpus (LMC) is a unique resource for research on issues of *Companion*-technology. LMC not only comprises 57.5 h of multimodal recordings (audio, video, psycho-biological data) from interactions between users—133 subjects in sum, balanced in age and gender—and a WoZ-simulated speech-based interactive dialogue system. LMC also includes full verbatim transcripts of all these dialogues, sociodemographic and psychometric data of all subjects as well as material from 73 in-depth user interviews focusing the user's individual experience of the interaction. In this chapter the experimental design and data collection of the LMC are shortly introduced. On this basis, exemplifying results from semantic analyses of the dialogue transcripts as well as from qualitative analyses of the interview material are presented. These illustrate LMC's potential for investigations from numerous research perspectives.

13.1 Introduction

In order to examine the system side as well as the user side of user-*Companion* interaction (UCI), naturalistic data of such interactions is indispensable. The LAST MINUTE corpus (LMC) includes approximately 57.5 h of such data. It is based on a widely standardized Wizard of Oz (WoZ) experiment in which the participants interacted with a simulated speech-based interactive dialog system [5, 29]. This system represented the main characteristics of future *Companion*-Systems, e.g. by pretending to be an individualized assistant in everyday life. The experimental modules simulated the following interaction situations: the collection

J. Frommer (✉) • M. Haase • J. Krüger
Universitätsklinik für Psychosomatische Medizin und Psychotherapie, Otto-von-Guericke-Universität, Magdeburg, Germany
e-mail: joerg.frommer@med.ovgu.de; matthias.haase@med.ovgu.de; julia.krueger@med.ovgu.de

D. Rösner • R. Andrich • R. Friesen • S. Günther
Institut für Wissens- und Sprachverarbeitung, Otto-von-Guericke Universität, Magdeburg, Germany
e-mail: roesner@ovgu.de; andrich@ovgu.de; friesen@ovgu.de; stguenth@ovgu.de

S. Biundo, A. Wendemuth (eds.), *Companion Technology*, Cognitive Technologies,
DOI 10.1007/978-3-319-43665-4_13

of personal as well as private user information for the purpose of personalization and individualization, and a mundane planning situation with the need for re-planning because of the accumulation of various planning constraints—a situation in which half of all participants got an emphatic system-initiated intervention. The following research questions led the development of the experiment and the design of the simulated system:

- What are the semantic markers in user utterances that enable us to identify negative dialog courses and the risk of a decrease in cooperativeness or even a communication break-up?
- How do the participants individually experience the interaction with the system, what do they ascribe to the system in order to develop their individual view on it and what emotions occur during the interaction?
- Which types of users can be differentiated by the semantic markers in participants' utterances, the sociodemographic as well as psychological characteristics of the participants and the subjective experiences of the system and the interaction with it?

The LMC provides multimodal recordings of all interactions including audio, video and psycho-biological data. These recordings are supplemented by transcripts of all interactions, data from well-established psychological instruments and material from in-depth user interviews focusing on the subjective experience of the experiment. The variety of recordings and supplementary material marks one of the ways the LMC goes beyond other naturalistic corpora [31]. This chapter is structured as follows: Initially, the LMC is introduced by a short overview on experimental design and data collection. On this basis, exemplifying insights from analyses of the collected data are presented—on the one hand with regard to semantic analyses of the interactions, and on the other hand with regard to qualitative analysis of the in-depth user interviews.

13.2 LAST MINUTE Experiment and Its Realization

The LAST MINUTE corpus contains multimodal recordings from a WoZ experiment that allows us to investigate how users interact with a *Companion*-System in a mundane situation with the need for planning, re-planning and strategy change. The design of this experiment is described in [27].

Before the interaction the subjects are instructed that they will interact with a speech-driven system and that they can begin the interaction by saying that they want to start. All system output is pronounced via a text-to-speech system (TTS).

The overall structure of an experiment is divided into (1) a personalization module, followed by (2) the LAST MINUTE module. These modules serve quite different purposes and are further substructured in a different manner (for more details, cf. [29]).

In the personalization module the system welcomes the subject, gives a short self-description and prompts the subject to tell and spell his name and (further) introduce himself, prompts for missing information and then gives a summary of the collected information. The system then stipulates user narratives on recent positive and negative events, his hobbies and his experiences with technical devices.

In the bulk of LAST MINUTE—the problem solving phase—the subject is expected to pack a suitcase for a 2-week holiday trip by choosing items from an online catalogue with 12 different categories that are presented in a fixed order. The options of each category are given as a menu with icons on the subject's screen.

In a normal packing subdialog a user requests articles from the current selection menu (e.g. 'ten t-shirts') followed by a confirmation of the system (e.g. 'ten t-shirts have been added'). For the whole packing a total of 15 min are allocated.

There are two ways to finish the current category and to proceed to the next one: (1) the user explicitly asks for a change of the category (in the following *SICC* for subject-induced category change) or (2) the system changes the category because a time limit is reached (*WICC* for wizard-induced category change). In the latter case the system informs the user that the selection from the current category has to end and that the following category is now available.

The normal course of a sequence of repetitive subdialogs is modified for all subjects at specific time points. These modifications or barriers are:

- after the sixth category, the current contents of the suitcase are listed verbally (*listing* barrier),
- during the eighth category, the weight limit for the suitcase is reached (*weight limit* barrier, WLB).
- at the end of the tenth category, the subject is informed that he has to pack the suitcase for cold weather in Waiuku instead of summer weather (*weather info* barrier, WIB).

Additional difficulties for the subjects may occur depending on the course of the dialog. These are typically caused by user errors or limitations of the system or a combination of both.

Nearly half of the subjects—randomly chosen—get an empathic intervention designed according to the principles of Rogerian psychotherapy [26] after the weather info barrier.

After the optional intervention or, when no intervention was given, immediately after the weather info barrier, the subjects finish packing the suitcase. We distinguish between two types of endings of the selection phase: (1) the user ends the selection on his own (*early enders*) or (2) the system ends the selection due to the global time limit reached. From a total of $N = 133$ subjects, $n_1 = 41$ subjects (with $n_2 = 20$ from the elderly and $n_3 = 21$ from the young group) ended the selection on their own; in the other $n_4 = 92$ cases the system closed the session and blocked additional user input.

After the end of the selection phase the system prompts the user to rate the outcome of the packing and reveal his plans for the holiday trip.

The system closes the session, thanks the user for his cooperation and says goodbye. Many users answer with (variants of) goodbye as well.

13.3 LAST MINUTE Corpus

13.3.1 Sample

User characteristics have an important influence on human computer interaction. Currently, age and gender are most debated sociodemographic data in HCI (e.g. [14, 23, 35]). A lot of studies evaluate the effect of gender on performance with technical devices, usage behavior, and experience of self-competence as well as the amount of anxiety while dealing with those devices. Furthermore, there is a vast number of studies regarding age. Headwords are the effect of social media usage on child development or possibilities to improve and integrate the handling of technical devices for elder users. It is noticeable that age cohorts are analyzed mostly exclusively from each other, which considerably complicates an actual comparison regarding usage and interaction behavior or performance while using technical devices. In order to take this into account we differentiated between participants at the age of 18–28 years and participants over 60 years. We minded equal distribution of gender and the level of education. We distinguished between higher educational level (general matriculation standard, studies at a university or a university of applied sciences) and lower educational level (secondary school or secondary modern school certificate, apprenticeship as highest educational/occupational qualification). We additionally evaluated sociodemographic data like profession, experience and skills using technical devices (especially computer systems). After finishing the 'LAST MINUTE' module, we employed the AttrakDiff [8] questionnaire to evaluate the hedonic and pragmatic quality of the simulated computer system. Participants filled out further psychometric questionnaires on personality factors [20], interpersonal difficulties [10], coping strategies [18], technology affinity [12] and attributional style [24] during a separate appointment. On average, participants needed 90 min for this process. Participants were recruited via advertisement in local newspapers and bulletins in vocational schools and universities as well as through sporting or recreational associations. Altogether we recruited 135 participants. In two experiments, there were technical problems during recording. In addition, three subjects did not complete the psychometric questionnaires.

Thus, the total sample consists of $N_{total} = 130$ subjects of which one could not be assigned properly to one of the educational levels (see Table 13.1).

Table 13.1 Sample distribution with regard to age, gender and educational level

	Male	Female	Total
Age-group 18–29 years			
Higher educational level	22	23	45
Lower educational level	13	12	25
Age-group over 60 years			
Higher educational level	14	13	27
Lower educational level	14	18	32
Total	63	66	129

13.3.2 Recorded Data

The LAST MINUTE corpus comprises multimodal recordings (audio, high-resolution video from multiple perspectives, psychobiological indicators such as skin reductance, heartbeat, and respiration) and verbatim transcripts of all completed experiments (N = 133). A complete WoZ session takes approx. 30 min. The total lengths of sessions vary from 19 to 39 min. The technical setup and used hardware (cf. Figs. 13.1, 13.2, and 13.3) are described in detail in [28].

GAT 2 Minimal Standard The recordings of the experiments and of the interviews were transcribed according to the GAT 2 minimal standard [33]. Besides the textualization of utterances, the standard also considers pauses, breathing, and overlaps (simultaneously spoken utterances) and allows for descriptive annotations.

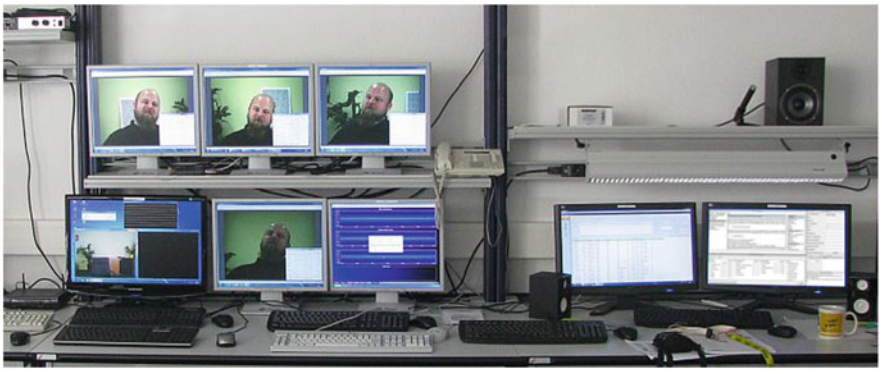

Fig. 13.1 Operator room. The functionality of the WoZ system was controlled by the *wizards* using the *right screens*. The recordings were coordinated by the operators using the *left screens*

Fig. 13.2 The hardware setting in the subject room. C=High resolution camera, H=Heart beat clip, M=Microphone, R=Respiration belt, S=Skin conductance clip, T=Stereo camera, W=Observation webcam. Not in the picture: Headwear microphone

Fig. 13.3 Subject room total view. The subjects sat in a comfortable chair in front of the subject screen in a room furnished like a living room

In the transcription process, we tried to come as close as possible to the actual wording and pronunciation of subjects. The transcripts, therefore, include typical phenomena of spontaneous spoken speech (e.g. discontinuities, repairs, restarts, incongruencies) as well as non-standard spellings (e.g. for dialect).

Quality Assurance The transcription was executed by trained personnel using the transcription software FOLKER [32]. Our own software was employed to support the transcribers in detecting and correcting possible misspellings and other defects. In addition, each transcript was examined multiple times by different transcribers.

13.3.3 Annotations

Dialog Act Representation (DAR) As reported in [30] the packing phase of the LAST MINUTE corpus was annotated with dialog acts. The utterances of the subjects show a large variance with respect to many linguistic features, so all dialog acts from the packing phase of each subject were annotated by a human rater. The annotation process was computer-assisted: raters verified automatic annotation suggestions and annotated remaining non-annotated segments. Most of the Wizard contributions were preformulated and could be annotated automatically by using regular expressions.

The DAR annotations are hierarchical. The first capital letter indicates the speaker (S for subject, W for wizard) while the second capital letter denotes the dialog act (x acts as a placeholder variable):

Sx: R = request, O = offtalk, Np = non-phonological and pauses, A = answer
Wx: A = accept, Rj = reject, I = information, Q = question/request

A third capital letter may refine a dialog act subtype:

SRx: P = packing, U = unpacking, E = exchange, F = finalization, L = listing, C = category change
SOx: T = general offtalk, Q = question
WAx: P = packing, U = unpacking, F = finalization
WRjx: P = packing, U = unpacking, Np = non-processable
WIx: C = category change, W = weather, F = finalization, L = listing
WQx: F = finalization, I = intervention, C = comment, E = elaboration

The lowercase letters (as in 'Rj') are inserted for readability.

Dialog Success Measures (DSMs) User requests may be either accepted and confirmed by the system or they may be rejected. The relation between subject requests and their acceptance or rejection allows us to define measures for the global dialog success (or failure) in the problem solving phase of LAST MINUTE [30]:

- first approach: ratio between the accepted subject requests and the total number of subject requests (termed **DSM1**)
- refinement: ratio between the accepted subject requests and the total number of turns (i.e. not only subject requests) in problem solving (termed **DSM2**)

  Thus for all subjects the following must hold: $0 \leq DSM2 \leq DSM1$.
  These measures are further addressed in Sect. 13.4.1 and Chap. 14.

LAST MINUTE Workbench The DAR is part of the LAST MINUTE workbench—a collection of tools for the exploration of the LAST MINUTE corpus. There are several visual representations of the DAR optimized for different tasks such as overview, search for patterns or calculation of statistics and other measures. The DAR and the LAST MINUTE workbench are described in more detail in [30].

An Example The DAR example in Table 13.2 is taken from a dialog segment where a subject (20110401adh) tries to pack a (winter) coat but the packing attempt is rejected several times (SRP WRjP pairs) and therefore the subject has to unpack several other items (SRU WAU) in order to create sufficient space. Please note the emotional expression of relief ('gott sei dank', Eng. 'thank god') when the subject finally succeeds.

Wizard Problems: Errors and Inconsistencies On the one hand the LMC experiment is a carefully designed and highly standardized experiment, based on a detailed manual [5], performed by intensively trained personnel (wizards) with elaborate computer support [27, 29].

Table 13.2 Transcript example

Tag	German text	English gloss
	...	...
SRP	ein mantel	A coat
WRjP	der artikel mantel kann nicht hinzugefügt werden (.) anderenfalls würde die maximale gewichtsgrenze des koffers überschritten werden	The item coat cannot be added (.) otherwise the weight limit of your suitcase will be exceeded
SNp	((raschelt)) ((schmatzt))	Rustles, smacks
SNp	(-)	(-)
SRU	ein buch raus	One book out
WAU	ein buch wurde entfernt	A book has been removed
SRP	ein mantel	A coat
WRjP	der artikel mantel kann nicht hinzugefügt werden (.) anderenfalls würde die maximale gewichtsgrenze des koffers überschritten werden	The item coat cannot be added (.) otherwise the weight limit of your suitcase will be exceeded
SRU	badelatschen raus	Beach slippers out
WAU	ein paar badelatschen wurden entfernt	A pair of beach slippers has been removed
SRP	ein mantel	A coat
WRjP	der artikel mantel kann nicht hinzugefügt werden (.) anderenfalls würde die maximale gewichtsgrenze des koffers überschritten werden	The item coat cannot be added (.) otherwise the weight limit of your suitcase will be exceeded
SOT	tja	Well
SNp	(1.77)	(1.77)
SOQ	was kann man denn noch rausnehmen	Well what else can be removed
SNp	(1.48)	(1.48)
SNp	pf pf pf pf pf pf pf	pf pf pf pf pf pf pf
SNp	(4.8)	(4.8)
SRU	zwei bh raus	Two bras out
WAU	zwei bhs wurden entfernt	Two bras have been removed
SRP	ein mantel	A coat
WAP	ein mantel wurde hinzugefügt	A coat has been added
SOT	gott sei dank	Thank god
	...	...

SNPs stand for nonphonological utterances. Following the GAT-2 minimal standard [33], short pauses are noted as (.) and (-), longer pauses with their duration in brackets, e.g. (1.77)

In spite of intensive training and a detailed manual [5] the wizards did not always operate consistently and accurately. We found, for example, inconsistent wizard behavior by analyzing the subject-initiated category changes. It turned out that some rejected wordings would have been accepted by different wizards or even by the same wizard in other situations.

Table 13.3 Tau statistics for sociodemographic features and successful actions

Kendall's tau	Age	d_{Cohen}	Gender	d_{Cohen}
Successes (abs.)			0.209***	0.51
Successes (rel.)	−0.117*	0.39	0.166**	0.41

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$ (cf. [30])

We also found some wizard errors, meaning situations where a wizard did not operate according to the guidelines of the manual. One type of such a wizard error is the rejection of a subject request with 'your input could not be processed' (WRjNp) when indeed the intention of the subject was clearly recognizable and the intended action was performable.

In sum, a manual like [5] is *necessary*, but by *no means sufficient* for successful experiments. A manual gives overall structure of experiments, but for non-trivial interactions nearly necessarily many questions will remain open.

In other words, spontaneous improvisation by wizards seems unavoidable, but it has a price: unreflected and unsupervised improvisation may—very likely—result in inter-session inconsistencies in wizard behavior. Indeed, the LAST MINUTE corpus contains many occurrences of inter-session wizard inconsistencies.

An example of such W inconsistencies is accepting or rejecting synonyms. In some cases wizards accepted synonyms of packed items, in other cases they did not.

Recommendations Thorough analysis of wizard errors and inconsistencies across the full LAST MINUTE corpus gives strong support for the following recommendations for future Wizard of Oz experiments:

- Invest in wizards. They need not only training, they need continuous supervision and monitoring.
- Expect the unexpected; therefore define appropriate meta rules for wizards.
- Invest in transcription and annotation, i.e. at least three independent transcribers with majority voting.

DAR-Based Measures The DAR characterizes dialog courses with sequences of dialog act labels. This allows the definition of a variety of measures for dialog success or failure.

One approach is to calculate the relative frequency of successful subject commands (SRx). As already reported in [30], Table 13.3 shows the correlation of absolute and relative frequency of successful actions with age and gender calculated using Kendall's tau. Missing values were not significant on a $p < 0.05$ level.

Successful Packing and Unpacking Actions The variety of actions changes during the experiment. At the beginning the subjects (normally) only pack items and change the category—which is usually accepted by the system (SRP WAP).[1] After the weight limit barrier (WLB) other actions become necessary. The subjects have to successfully unpack items first in order to be able to pack others.

[1] Some subjects try to unpack items before it is necessary, but these are only single cases.

A more refined dialog success measure is given by the number of successful packing actions after the WLB. This quantitative measure uncovers a subgroup of eight subjects (of $N_{total} = 133$) that did not have a single successful packing action after the WLB at all. This is in sharp contrast to the top performer group in the fourth quartile with at least nine and up to 14 successful packings after the WLB. Why users fail in such a drastic way was then a matter of qualitative in-depth analyses of the transcripts. It turned out that virtually in all eight cases with zero successful packing actions after the WLB, wizard errors and inconsistencies caused or heavily contributed to the negative dialog courses that ended without any more progress after the WLB.

A closer look at those top performers with 10 or more packing successes after WLB reveals that from a total of 24 subjects in this subgroup we find 13 that are young and female (cf. Table 13.4), far more than would be expected in a random sample.

These observations motivate us to have a closer look at the differences between the age group- and gender-based subgroups of subjects with respect to the number of successful packing actions after the WLB (cf. Figs. 13.4 and 13.5). We exclude the eight outliers with zero successes and work with $N_{nozero} = 125$ subjects. A Kruskal-Wallis test with the four age group and gender pairings yields a significant result (Kruskal-Wallis chi-squared = 9.4945, df = 3, $p = 0.02339$*). The mean for young women is larger than the means of the other three groups each (the mean values in decreasing order are mean(yw) = 8.182, mean(ym) = 6.706, mean(ew) = 6.333, and mean(em) = 5.857). Wilcoxon tests give the following results when we test young women against the three other pairings: yw against ym yields W = 413, p = 0.06408; yw against ew W = 319, $p = 0.01499$*, $d_{Cohen} = 0.59$; yw against

Table 13.4 Top performers after weight limit (for details see text)

Age/Sex	Male	Female
Young	5	13
Elderly	4	2

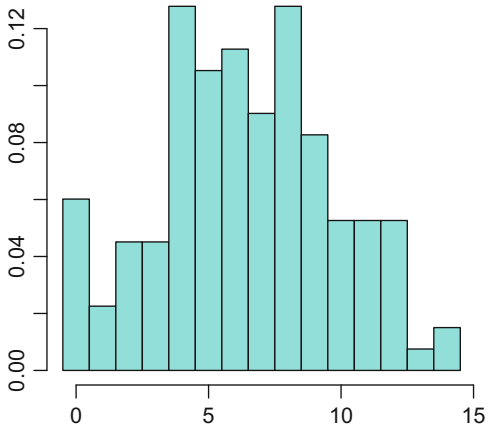

Fig. 13.4 Distribution of number of successful packing actions after WLB per transcript ($N = 133$). Please note the amount of zero values

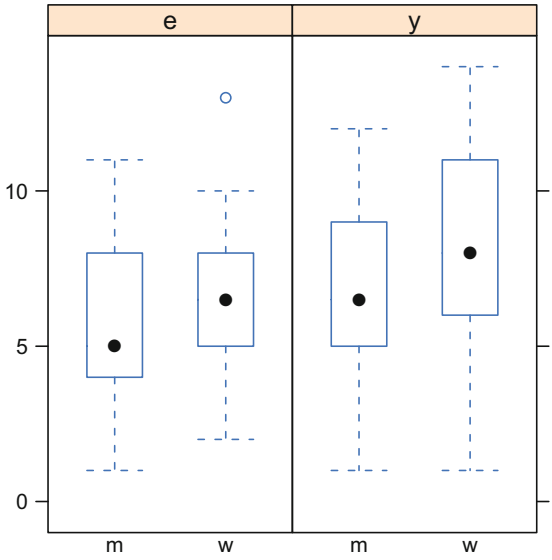

Fig. 13.5 Distributions of number of successful packing actions after WLB conditioned by age group and gender; eight outliers with zero actions removed (see text)

em W = 275, $p = 0.006328^{**}$, $d_{Cohen} = 0.73$). This finding replicates other results about age group- and gender-based differences in dialog success in the LMC. But as a dialog success measure, 'successful packing actions after WLB' has a major weakness, it is too task-dependent. For more general and task-independent success measures and the resp. results see Sect. 13.4.1 and Chap. 14.

Initiative in LAST MINUTE Dialogs An important aspect of dialogs is the initiative: Which participant has the initiative and is thus driving the dialog? While the system generally has the initiative during the personalization phase, the problem solving phase is primarily characterized by user initiative. The subject expresses a request (R) for a system action and, as a response, the system either confirms (A) or rejects (Rj) the request. An action may be rejected based on aspects of the user's utterance ('your request could not be processed') or—although the utterance was 'understood' and accepted—because the action cannot be performed for task-related reasons. In sum, the pairs of successive dialog acts are SRx WAx, SRx WRjx or SRx WRjNp, but there are exceptions to this general rule.

One example for system initiative is category change. If the allocated processing time of approximately 1 min for a selection category is over, the system takes initiative, informs the user about the timeout (WIF) and performs the change to the next category (WIC).

Self-initiated changes of the category for selection are an indicator for the degree of control of the dialog flow and for the taken initiative that a subject exhibits. As already reported in [30]: An in-depth analysis of self- vs. system-initiated category changes reveals a highly significant correlation between age and number of successful self-initiated category changes after the weight limit barrier; calculating Kendall's tau for these two quantities reveals that, with a tau statistic of −0.27, a

p-value smaller than 10^{-5} and $d_{Cohen} = 0.69$ as effect size, younger subjects have on average a higher number of such self-initiated category changes than the elderly.

13.3.4 Post-Hoc Interviews

Currently, there are research investigations regarding methods and techniques for the technical realization of *Companion*-Systems (e.g. [9]). Comparatively, only few theoretical considerations and hardly any empirical investigations about the users' inner processes, feelings and experiences during the interaction with such a system exist (e.g. [37]). In order to contribute to a more in-depth understanding of what goes on in users' minds during user-*Companion* interactions (UCI), it was decided to realize interviews with approximately half of all participants [17].

The interviews were conducted subsequent to the experiment. They were guided by the idea that the participants will make up their individual views on the simulated system by ascribing human-like characteristics, motives, wishes, beliefs, etc. to it (Intentional Stance, [4]). Besides intentional ascriptions to the system the interview focused on: emotions occurring during the interaction, the subjective experience of the speech-based interaction and the intervention (if given), the role of technical devices in autobiography and the general evaluation of the system.

The interviews were semi-structured to enable the participants to express their individual views and experiences freely and non-restrictedly (e.g. [7]). A preluding non-specific narration stimulus evoked a so-called initial narration regarding the experience of the experiment ("You have just done an experiment. Please try to put yourself back in this experiment. Please tell me how you did during the experiment. Please tell me in detail what you thought and experienced!"). If necessary, questions for clarifying and specifying vague narration sequences were given (immanent questions). Afterwards, pre-determined open questions (exmanent questions) from an interview guide were used, which dealt with the topics mentioned above and were handled flexibly with regard to formulation and chronological order. These questions were extended if unexpected relevant aspects were reported.

The interviewees were recruited according to a qualitative sampling scheme. Therein, age, educational level, sex and assignment to experimental or control group were considered to get a heterogeneous sample in order to maximize variance of subjective experience. Finally the sample included 73 participants of the total sample of the LMC, nearly balanced in regard to the mentioned variables. In total, 96 h and 22 min of interview material were recorded and transcribed in large parts according to the GAT-2 minimal standard [33].

13.4 Insights from Analyses

13.4.1 Discourse Analysis of Transcripts

Turn-Based Measures For the analysis of (two party) dialogs, so-called *'adjacency pairs'*—i.e. pairs of consecutive dialog acts or turns of the two participants—are a fundamental unit [11].

In LAST MINUTE typical adjacency pairs have two different structures [30]:

- In the personalization phase, the typical adjacency pair is a wizard question or prompt followed by the subject's answer or narrative.
- In the problem solving phase, a typical adjacency pair is made up of a user request for some action (packing some items, unpacking, changing category, etc.) followed by either a confirmation or a rejection by the wizard.

Given this dialogic structure, the number of turns of the participants is a measure of the total extent of the complete interaction or of its subparts. Taking the total number of turns in the problem solving phase (without intervention) we find a high variance within the whole cohort of $N = 133$ (cf. Figs. 13.6 and 13.7).[2,3]

The primary result about differences in total number of turns in problem solving (without intervention) is that gender matters (cf. Fig. 13.7 and Table 13.5). Men (fewer turns) significantly differ from women; young men significantly differ from

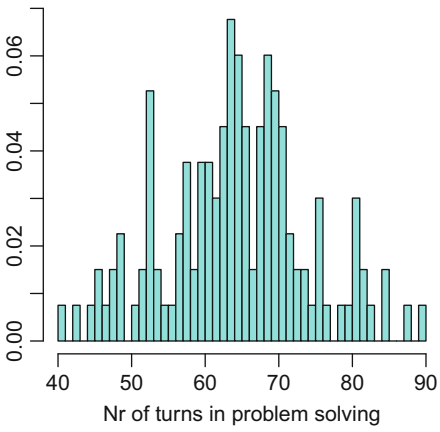

Fig. 13.6 Distribution of total number of user turns in problem solving (without intervention) per transcript (mean = 64.75, median = 65.00, sd = 9.79, N = 133)

[2]The distribution is visualized here—and in other figures—as a trellis boxplot: the rectangles represent the interquartile range (i.e. the range of 25% of the values above and below the median resp.); the filled dot gives the median; the whiskers extending the rectangle extend to the range of values, but maximally to 1.5 of the interquartile range; outlier values beyond the maximal whisker range are given as unfilled dots (cf. [1]).

[3]Unless noted otherwise, all statistical tests and calculations have been performed with the R, language [1, 25].

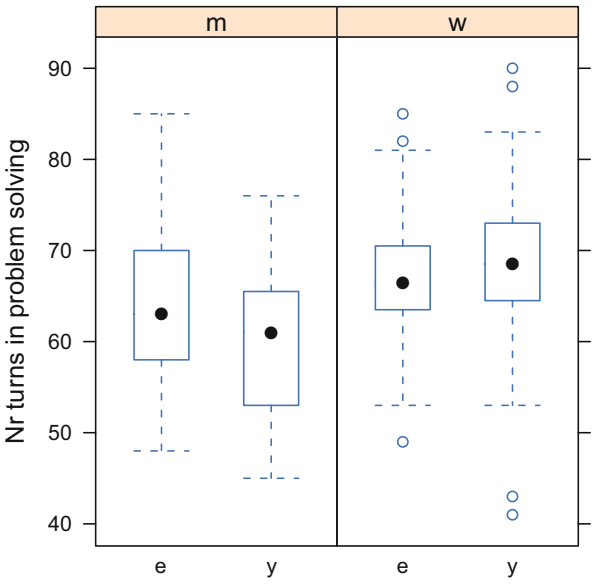

Fig. 13.7 Total nr of turns in problem solving (without intervention) per transcript conditioned by age group (e = elder, y = young subjects) and gender (m = men, w = women; please note outliers in the subgroup of young and elder women)

Table 13.5 Results from t tests for differences in *number of turns in problem solving* for subgroups of all subjects

g1	rel	g2	p value	d_{Cohen}
m	<	w	0.001219**	0.57
e	<	y	n.s	n.a.
ym	<	em	0.04275*	0.53
ym	<	ew	0.0007628***	0.86
ym	<	yw	0.000746***	0.83
em	<	yw	n.s.	n.a.
em	<	ew	n.s.	n.a.
ew	<	yw	n.s.	n.a.

Please note: Shapiro normality tests allow normality assumptions for all investigated subgroups
Significance levels: *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

the other three groups. Both findings have a domain-dependent reason: men, and especially young men, do *on average* pack fewer items and therefore employ fewer packing requests.

Dialog Success Measures Measures based on simply counting turns have a major fault: They do not differentiate between successful and unsuccessful adjacency pairs. This is remedied by measures taking (local) success and failure into account. In the following, we therefore work with the dialog success measures as defined in Sect. 13.3.3.

For differences between sociodemographic groups with respect to both dialog success measures we find that age group matters (cf. Table 13.6), whereas gender

Table 13.6 Results from Wilcoxon tests for differences between sociodemographic groups in dialog success measures *DSM1* and *DSM2*

g1	rel	g2	p DSM1	d_{Cohen}	p DSM2	d_{Cohen}
y	>	e	0.007459**	0.32	0.001061**	0.51
w	>	m	n.s	n.a	n.s.	n.a
yw	>	ym	0.01281*	0.54	0.02627*	0.43
yw	>	em	0.01589*	0.635	0.002896**	0.69
yw	>	ew	0.001619**	0.77	0.0005762***	0.79
ym	>	em	n.s.	n.a	n.s.	n.a
ym	>	ew	n.s.	n.a	n.s.	n.a
em	>	ew	n.s.	n.a	n.s.	n.a

Significance levels: *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

alone does not give a significant difference. In detail: The difference between young and elderly subjects is significant. The subgroup of *young women* shows significantly **higher** values (of DSM1 and DSM2) than the other three groups.

Phasewise Dialog Success Values The three major subphases of problem solving in the LAST MINUTE experiment are demarcated by the weight limit barrier (WLB) and the weather info barrier (WIB). Up to the weight limit barrier there is no need for unpacking (although some subjects do some unpacking, for example as part of exchange actions). After the weight limit barrier, unpacking becomes crucial. Without at least one successful unpacking there is no further progress possible. Finally, the weather info demands a strategy change. Now items for cold and rainy weather are needed in exchange for bathing suits and other summer items.

The three phases reach

- P2s: from the start of problem solving to the weight limit barrier (WLB),
- P2b: from the weight limit barrier to the weather info (WIB), and finally
- P2w: from the weather info to the end of the experiment.

Global success values differ remarkably within the different phases. The distributions for the successful requests to all request ratios (DSM1) are presented in Fig. 13.8. Note that in the P2s phase (before the weight limit barrier) more than 75% of the subjects have DSM1 values over 0.95, whereas in the subsequent phases, P2b and P2w, the means drop to 0.6877 and 0.7096 resp. The differences between the mean values for DSM1 before the WLB and those of P2b and P2w are highly significant (Wilcoxon tests: P2s vs. P2b, W = 2932.5, $p < 2.2 \times 10^{-16}$, $d_{Cohen} = 1.35$; P2s vs. P2w, W = 490.5, $p < 2.2 \times 10^{-16}$, $d_{Cohen} = 2.06$), whereas the slight differences between P2b and P2w are insignificant ($W = 9451$, $p = 0.3337$, $d_{Cohen} = 0.10$).

We get a similar picture for the phasewise values of the successful requests to all turn ratios (DSM2) (cf. Fig. 13.9). The differences between the mean values for DSM2 before the WLB and those of P2b and P2w are highly significant (Wilcoxon tests: P2s vs. P2b, W = 444.5, $p < 2.2 \times 10^{-16}$, $d_{Cohen} = 2.595$; P2s vs. P2w, W

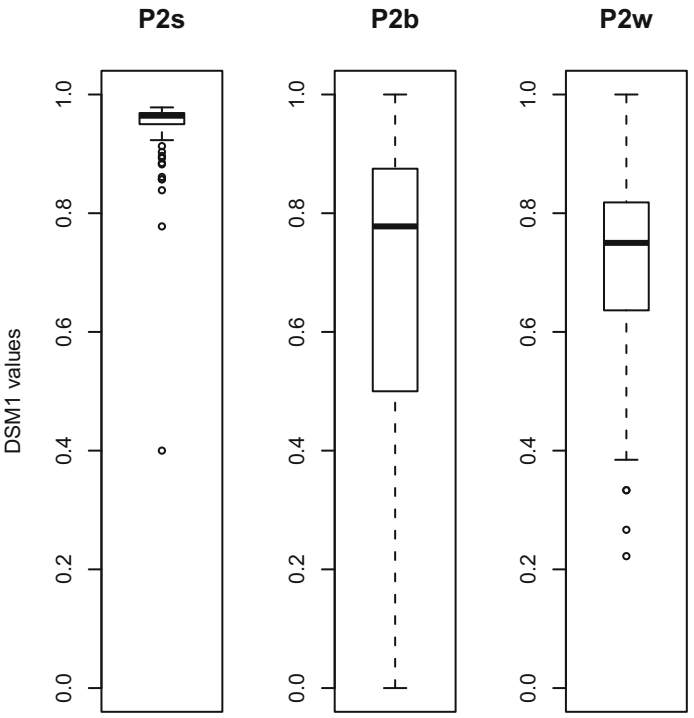

Fig. 13.8 Distributions of DSM1 values in the three subphases of problem solving ($N = 133$)

$= 356.5, p < 2.2 \times 10^{-16}, d_{Cohen} = 3.395$), whereas the slight differences between P2b and P2w are insignificant ($W = 8303, p = 0.3883, d_{Cohen} = 0.15$).

Evaluations with Phasewise DSMs Dialog success measures for subphases of the complete LAST MINUTE dialogs allow us to investigate questions that are raised by findings with the global measures. A point in case for such more fine-grained analyses is questions about differences between user success with respect to the barriers or the intervention.

When we investigate differences in global dialog success between the subgroups of subjects with and without intervention, we find that the differences between the two groups with respect to global DSM2 values are close to significant (Wilcoxon test: $W = 1794, p = 0.06258, d_{Cohen} = 0.40$; the resp. differences for DSM1 are insignificant).

This raises the question of whether the higher global DSM2 values for the group with intervention can be attributed to the intervention. We therefore perform a fine-grained investigation of the phasewise DSM values. We first compare the subgroups of subjects with and without intervention in the dialog phase P2w after the weather information (and thus after the intervention). We get differences in the means for this phase for both DSM values but these are not significant (Wilcoxon rank sum

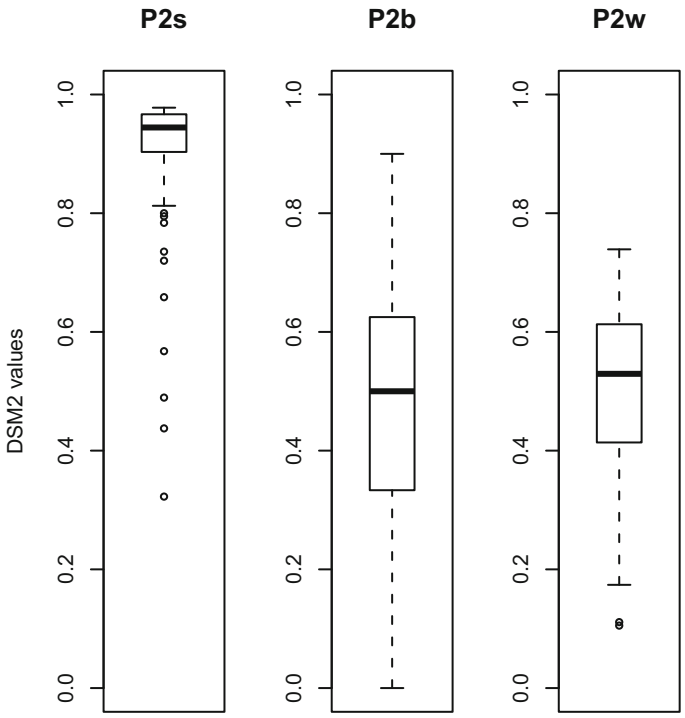

Fig. 13.9 Distributions of DSM2 values in the three subphases of problem solving ($N = 133$)

test, DSM1: $W = 2240.5$, $p = 0.8854$, $d_{Cohen} = 0.03$; DSM2: $W = 1946.5$, $p = 0.2397$, $d_{Cohen} = 0.22$).

When we then compare phasewise DSM values in the phases from the beginning of problem solving to the weight limit (P2s) and from there to the weather info (P2b), we see additionally that already in those phases (long before the intervention) the subgroup with intervention has higher DSM values than the subgroup without. Some differences between the subgroups are even close to significant (e.g. Wilcoxon for DSM2 in P2s: W = 1780.5, $p = 0.0544$, $d_{Cohen} = 0.42$).

In other words, the dialog behavior of the groups with and without intervention already differed long before the intervention took place with the former group being more successful on average than the latter. This analysis result is in line with a general observation: As soon as user–*Companion* dialogs show enough degrees of freedom and approximate realistic dialogs—as is the case with the LAST MINUTE dialogs—their course under varying constraints is hard to predict in advance and fine-grained analyses are needed.

13.4.2 *"Okay Now It Is a Computer Voice Talking with Me": Subjective Experience of the Artificial Computer Voice in the LMC*

One of the main objectives in investigating the user interviews conducted subsequent to the WoZ experiment was to reconstruct the participants' subjective experience of the Initial Dialogue (personalization module). Therefore, analyses focused on participants' individual views on the system, including their ascriptions regarding the system's interaction behavior and essential features as well as their experiences of themselves during this module (see [16] for some of the main results). Thereby, participants' reports explicitly dealing with the system's artificial voice occurred repeatedly. The interview guide did not contain any exmanent question regarding this issue, which means that the participants came up with it by themselves. Because of its subjective relevance, it was decided to focus on the subjective experiences of the system's voice in an in-depth analysis.

Sample and Material In order to investigate the subjective experience of the Initial Dialogue, the initial narratives of 31 participants were analyzed.[4] The aim of this study was to collect and analyze reports on the subjective experience of the system's voice in the interviews of these 31 participants. Up to now, collecting and analyzing reports from the initial narratives of 12 of these participants has been finished (five young men and six young women, one elderly woman).

Analyses The interview sequences were broken up into so-called meaning units and then were condensed, summarized and interpreted using methods of qualitative content analysis [19]. Throughout this abstraction process categories were developed which arrange the summarized meaning units.

Results Three categories were developed out of the material. These represent the variance of participants' experiences regarding the artificial computer voice.

1. *Characterization of the voice—from weird to agreeable:* The system's voice is mainly characterized as *"monotone" (e.g. KK*[5]*)*, *"non-human" (e.g. UK)*, *"choppy" (e.g. CT)*, *"weird" (e.g. AK)* and *"mechanical" (e.g. SS)*. It is clearly associated with a technical device and differentiated from a human voice *("a very technical voice which has so little of a human face", SR)*. There are participants who ascribe to it that it is unable to convey emotions *("affectionately cold", SR)*. Taken together with the absence of dialect and prosody, this leads to criticism regarding the voice's lack of individuality as well as regarding the system's strange- and differentness experienced by the participants *("well it was like*

[4] In order to maximize variance, interviews were chosen randomly one by one considering a balanced allocation of groups according to the qualitative sample plan (cf. Sect. 13.3.4). When no more increase of variance could be detected, no further material was added (criterion of theoretical saturation).

[5] Participant's initials.

talking to someone who is not German and just started to learn the language and pronounces the words differently (.) it's really like yeah almost like an accent so to say", UK). There are only few participants who generally perceive the voice as *"agreeable" (AK)* or experience variability in the voice during the interaction *(MS)*.

2. *The voice's impact on the participant—from adaption to distancing:* The voice contributes to experiencing the interaction as a man-machine interaction. It is perceived as unusual, and it is strange to talk to an artificial voice. Especially at the beginning, the voice evokes feelings of scariness, irritation or anxiety *("I felt uneasy about it", MS)*. Furthermore, it can inhibit the development of trust in the system (*"if it had been a nice and lovely voice one would of course have had more trust even though it is a computer", KK)*. Many participants refrain from using familiar communication principles from human-human interaction (HHI) *("one don't even try to speak with it like with a human being instead one switches to the mode okay now it is a computer voice speaking with me", UK)*. Based on experiencing the voice as monotone and non-human, the participants ascribe to the system that it is demanding, does not tolerate contradictions or allows breaking its flow of words. These issues create a feeling of losing control. Furthermore, participants' initiative reduces, because they feel hampered in influencing the interaction *("because of this strange voice it was like it asks me something and I react", FW)*. The voice is experienced as keeping the participant at a distance *("it appeared like rejecting", SS)* or otherwise evoking the desire for distancing in the participant. Some participants continuously struggle with accepting the voice, whereas others make lots of substantial adjustments to adapt to it. On the one hand, they make cognitive efforts *("really strain oneself", MH)*. On the other hand, they adapt their behavior to the system's abilities, which they anticipate as being deficient because of the system's voice quality *("that I have to speak clearly and can't use slang because maybe it will not understand that", UK)*.
3. *Wishes regarding the voice—from somehow different to more human-like:* Besides participants who would simply appreciate a change of the voice, there are participants who express more specific ideas regarding this change, e.g. the wish for a more modulated voice *("more fluently (. . .) like a wave while talking", CT)* or the preference for a female voice. Furthermore, some would appreciate a more individual voice using, e.g. idioms or proverbs, to be enabled to develop a *"personal counterpart" (FW)*, which appears less strange and more likeable. This all culminates in the wish for a more human-like voice, which features all of these characteristics and is imagined as much more pleasant. All in all, the idea of individualization of the voice runs through the participants' reports, e.g. the possibility to choose out of a set of predefined voices.

Discussion Participants' descriptions of the computer voice are comparable to those reported in other studies (e.g. [36]). Its artificial sound triggers associations with a technical counterpart. The participants require a lot of cognitive effort to understand it and adapt their behaviour to anticipated system's abilities. They

want to understand the system and want to be understood by it. Regarding the emotional level, they are frightened, scared and alienated, feel like they are losing control and wish to distance themselves from the voice—feelings which can be discussed regarding the uncanny valley phenomenon [22]. Other studies [13, 36] underline the dissatisfaction with artificial systems' voices in HCI. In our analysis, participants consequently withdraw from the interaction by reducing their initiative. This is a critical finding when considering the aim of *Companion*-Systems to be experienced as trustworthy attendants and relational partners [3, 37]. Besides the desire to generate mutual understanding, the need to relate to the system could be conveyed by the wish for a more pleasant and human-like voice, which is in line with users' ideas for improving HCI by referring to experiences from HHI [36]. Maybe participants imagine that they could enter a deeper and more trustworthy relationship with such a voice. According to [15] this could be explained by the humans' need to belong [2], which appears even in interactions with simple artifacts showing only basic social cues like speech.

It has to be considered that the presented results are preliminary, because they are based on material of mainly young participants, which is currently being expanded regarding sample size and considered interview sequences. Nevertheless, first meaning structures occurred, which will be verified and extended in further analyses.

13.5 Summary

The LMC stands out from other available naturalistic corpora considering sample size, sample heterogeneity, total length of each interaction, number and quality of the recordings as well as further user information from questionnaires and interviews (cf. [21, 29]). It marks a cornerstone for work in the SFB/TRR 62 [3]. Working groups in the SFB consortium focus on research questions regarding the automatic detection and classification of markers for users' emotions during the interaction in different modalities (e.g. [6, 34], see also Chap. 14). In this chapter, exemplary results were presented regarding the users' observable dialog behavior and dialog success as well as results regarding their unobservable inner processes, feelings and evaluations during the interaction. All in all, the LMC provides potential for investigations from various perspectives, and researchers are invited to explore it according to their particular research interests.

Contributions This paper reports on joint work—i.e. the design and execution of the LAST MINUTE experiments—as well as on distinct contributions of the two involved groups. The responsibility for the dialog act annotation of the LMC and for the discourse analysis of transcripts (cf. Sects. 13.3.3 and 13.4.1) lies with D. Rösner, R. Andrich, R. Friesen, and S. Günther. The responsibility for sample and post-hoc interviews (cf. Sects. 13.3.1, 13.3.4, 13.4.2) lies with J. Frommer, M. Haase and J. Krüger. For the discussion and the conclusions in Sect. 13.5 the responsibility is shared by all authors.

Availability The LAST MINUTE corpus is available for research purposes upon written request (via email or mail) from project A3 of SFB TRR 62 (heads: Prof. Frommer and Prof. Rösner).

Acknowledgements This work was done within the Transregional Collaborative Research Centre SFB/TRR 62 "*Companion*-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

References

1. Baayen, R.H.: Analyzing Linguistic Data – A Practical Introduction to Statistics Using R. Cambridge University Press, Cambridge (2008)
2. Baumeister, R.F., Leary, M.R.: The need to belong: desire for interpersonal attachments as a fundamental human motivation. Psychol. Bull. **117**(3), 497 (1995)
3. Biundo, S., Wendemuth, A.: *Companion*-technology for cognitive technical systems. Künstl. Intell. (2016). doi:10.1007/s13218-015-0414-8
4. Dennett, D.C.: The Intentional Stance. MIT, Cambridge (1987)
5. Frommer, J., Rösner, D., Haase, M., Lange, J., Friesen, R., Otto, M.: Project A3 - Detection and Avoidance of Failures in Dialogs. Pabst Science Publisher, Lengerich (2012)
6. Frommer, J., Michaelis, B., Rösner, D., Wendemuth, A., Friesen, R., Haase, M., Kunze, M., Andrich, R., Lange, J., Panning, A., Siegert, I.: Towards emotion and affect detection in the multimodal last minute corpus. In: Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Paris (2012)
7. Galletta, A.: Mastering the Semi-Structured Interview and Beyond: From Research Design to Analysis and Publication. NYU, New York (2013)
8. Hassenzahl, M., Burmester, M., Koller, F.: AttrakDiff: ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In: Ziegler, J., Szwillus, G. (eds.) Mensch & Computer 2003. Berichte des German Chapter of the ACM, vol. 57, pp. 187–196. Vieweg+Teubner Verlag, Stuttgart (2003)
9. Honold, F., Schüssel, F., Weber, M., Nothdurft, F., Bertrand, G., Minker, W.: Context models for adaptive dialogs and multimodal interaction. In: 2013 9th International Conference on Intelligent Environments (IE), pp. 57–64. IEEE, New York (2013)
10. Horowitz, L.M., Alden, L.E., Wiggins, J.S., Pincus, A.L.: Inventory of Interpersonal Problems Manual. Psychological Cooperation, San Antonio (2000)
11. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2nd edn. Prentice Hall, Englewood Cliffs (2008)
12. Karrer, K., Glaser, C., Clemens, C., Bruder, C.: Technikaffinität erfassen – der Fragebogen TA-EG. In: Lichtenstein, A., Stößel, C., Clemens, C. (eds.) Der Mensch im Mittelpunkt technischer Systeme, ZMMS Spektrum, Reihe 22, vol. 29, pp. 196–201. VDI Verlag GmbH, Düsseldorf (2008)
13. Kastner, M., Stangl, B.: Exploring a text-to-speech feature by describing learning experience, enjoyment, learning styles, and values–a basis for future studies. In: 2013 46th Hawaii International Conference on System Sciences (HICSS), pp. 3–12. IEEE, New York (2013)
14. King, J., Bond, T., Blandford, S.: An investigation of computer anxiety by gender and grade. Comput. Hum. Behav. **18**(1), 69–84 (2002)
15. Krämer, N.C., Eimler, S., von der Pütten, A., Payr, S.: Theory of companions: what can theoretical models contribute to applications and understanding of human-robot interaction? Appl. Artif. Intell. **25**(6), 474–502 (2011)

16. Krüger, J., Wahl, M., Frommer, J.: Making the system a relational partner: users' ascriptions in individualization-focused interactions with companion-systems. In: Proceedings of the 8th International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services (CENTRIC 2015), pp. 48–54. IARIA XPS Press (2015). http://www.thinkmind.org/index.php?view=article&articleid=centric_2015_3_20_30079
17. Lange, J., Frommer, J.: Subjektives Erleben und intentionale Einstellung in Interviews zur Nutzer-Companion-Interaktion. In: Informatik 2011. Lecture Notes in Informatics, vol. 192, p. 240. Köllen, Bonn (2011). http://www.user.tu-berlin.de/komm/CD/paper/060332.pdf
18. Levenstein, S., Prantera, C., Varvo, V., Scribano, M.L., Berto, E., Luzi, C., Andreoli, A.: Development of the Perceived Stress Questionnaire: a new tool for psychosomatic research. J. Psychosom. Res. **37**(1), 19–32 (1993)
19. Mayring, P.: Qualitative content analysis. Theoretical foundations, basic procedures and software solution. n.p., Klagenfurth (2014). http://nbn-resolving.de/urn:nbn:de:0168-ssoar-395173[retrieved:09,2015]
20. McCrae, R.R., Costa, P.T.: A contemplated revision of the NEO Five-Factor Inventory. Personal. Individ. Differ. **36**(3), 587–596 (2004)
21. McKeown, G., Valstar, M.F., Cowie, R., Pantic, M.: The SEMAINE corpus of emotionally coloured character interactions. In: 2010 IEEE International Conference on Multimedia and Expo (ICME), pp. 1079–1084. IEEE, New York (2010)
22. Mori, M., MacDorman, K.F., Kageki, N.: The uncanny valley [from the field]. IEEE Robot. Autom. Mag. **19**(2), 98–100 (2012)
23. Naumann, A., Hermann, F., Peissner, M., Henke, K.: Interaktion mit Informations-und Kommunikationstechnologie: Eine Klassifikation von Benutzertypen. In: Herczeg, M., Kindsmüller, M.C. (eds.) Mensch & Computer 2008: Viel Mehr Interaktion, pp. 37–45. Oldenbourg Verlag, München (2008)
24. Peterson, C., Semmel, A., Baeyer, C.v., Abramson, L.Y., Metalsky, G.I., Seligman, M.E.P.: The attributional Style Questionnaire. Cogn. Ther. Res. **6**(3), 287–299 (1982)
25. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2010)
26. Rogers, C.: A theory of therapy, personality and interpersonal relationships as developed in the client-centered framework. In: Koch, S. (ed.) Psychology: A Study of a Science. Formulations of the Person and the Social Context, vol. 3. McGraw Hill, New York (1959)
27. Rösner, D., Friesen, R., Otto, M., Lange, J., Haase, M., Frommer, J.: Intentionality in interacting with companion systems – an empirical approach. In: Jacko, J. (ed.) Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments. Lecture Notes in Computer Science, vol. 6763, pp. 593–602. Springer, Berlin/Heidelberg (2011)
28. Rösner, D., Frommer, J., Andrich, R., Friesen, R., Haase, M., Kunze, M., Lange, J., Otto, M.: LAST MINUTE: a novel corpus to support emotion, sentiment and social signal processing. In: Devillers, L., Schuller, B., Batliner, A., Rosso, P., Douglas-Cowie, E., Cowie, R., Pelachaud, C. (eds.) Proceedings of LREC'12 - Workshop Abstracts, Istanbul, p. 171 (2012)
29. Rösner, D., Frommer, J., Friesen, R., Haase, M., Lange, J., Otto, M.: LAST MINUTE: a multimodal corpus of speech-based user-companion interactions. In: LREC, pp. 2559–2566 (2012)
30. Rösner, D., Friesen, R., Günther, S., Andrich, R.: Modeling and evaluating dialog success in the LAST MINUTE corpus. In: Chair, N.C.C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of LREC'14, Reykjavik (2014)
31. Rösner, D., Haase, M., Bauer, T., Günther, S., Krüger, J., Frommer, J.: Desiderata for the design of companion systems – insights from a large scale wizard of Oz experiment. Künstl. Intell. **30**(1), 53–61 (2016). Online first: Oct 28, 2015; doi:10.1007/s13218-015-0410-z
32. Schmidt, T., Schütte, W.: Folker: an annotation tool for efficient transcription of natural, multi-party interaction. In: Chair, N.C.C, Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta (2010)

33. Selting, M., Auer, P., Barth-Weingarten, D., Bergmann, J.R., Bergmann, P., Birkner, K., Couper-Kuhlen, E., Deppermann, A., Gilles, P., Günthner, S., et al.: Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). Gesprächsforschung-Online-Zeitschrift zur verbalen Interaktion **10** (2009)
34. Siegert, I., Philippou-Hübner, D., Hartmann, K., Böck, R., Wendemuth, A.: Investigation of speaker group-dependent modelling for recognition of affective states from speech. Cogn. Comput. **6**(4), 892–913 (2014). doi:10.1007/s12559-014-9296-6
35. Sieverding, M., Koch, S.C.: (Self-) evaluation of computer competence: how gender matters. Comput. Educ. **52**(3), 696–701 (2009)
36. Veletsianos, G.: How do learners respond to pedagogical agents that deliver social-oriented non-task messages? Impact on student learning, perceptions, and experiences. Comput. Hum. Behav. **28**(1), 275–283 (2012)
37. Wilks, Y.: Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues, vol. 8. John Benjamins Publishing, Amsterdam (2010)