Chapter 14
The LAST MINUTE Corpus as a Research Resource: From Signal Processing to Behavioral Analyses in User-*Companion* Interactions

Dietmar Rösner, Jörg Frommer, Andreas Wendemuth, Thomas Bauer, Stephan Günther, Matthias Haase, and Ingo Siegert

Abstract The LAST MINUTE Corpus (LMC) is one of the rare examples of a corpus with naturalistic human-computer interactions. It offers richly annotated data from $N_{total} = 130$ experiments in a number of modalities. In this paper we present results from various investigations with data from the LMC using several primary modalities, e.g. transcripts, audio, questionnaire data.

We showed that sociodemographics (age, gender) have an influence on the global dialog success. Furthermore, distinct behavior during the initial phase of the experiment can be used to predict global dialog success during problem solving. Also, the influence of interventions on the dialog course was evaluated.

D. Rösner (✉)
Institut für Wissens- und Sprachverarbeitung (IWS), Otto-von-Guericke Universität, 39016 Magdeburg, Germany

Center for Behavioral Brain Sciences, 39118 Magdeburg, Germany
e-mail: roesner@ovgu.de

J. Frommer • M. Haase
Universitätsklinik für Psychosomatische Medizin und Psychotherapie, Otto-von-Guericke Universität, 39120 Magdeburg, Germany
e-mail: joerg.frommer@med.ovgu.de; matthias.haase@med.ovgu.de

A. Wendemuth
Institut für Informations- und Kommunikationstechnik (IIKT), Otto-von-Guericke Universität, 39016 Magdeburg, Germany

Center for Behavioral Brain Sciences, 39118 Magdeburg, Germany
e-mail: andreas.wendemuth@ovgu.de

T. Bauer • S. Günther
Institut für Wissens- und Sprachverarbeitung (IWS), Otto-von-Guericke Universität, 39016 Magdeburg, Germany
e-mail: tbauer@iws.cs.uni-magdeburg.de; stguenth@iws.cs.uni-magdeburg.de

I. Siegert
Institut für Informations- und Kommunikationstechnik (IIKT), Otto-von-Guericke Universität, 39016 Magdeburg, Germany
e-mail: ingo.siegert@ovgu.de

S. Biundo, A. Wendemuth (eds.), *Companion Technology*, Cognitive Technologies,
DOI 10.1007/978-3-319-43665-4_14

Additionally, the importance of discourse particles as prosodic markers could be shown. Especially during critical dialog situations, the use of these markers is increasing. These markers are furthermore influenced by user characteristics.

Thus, to enable future *Companion*-Systems to react appropriately to the user, these systems have to observe and monitor acoustic and dialogic markers and have to take into account the user's characteristics, such as age, gender and personality traits.

14.1 Introduction

Wizard of Oz (WoZ) experiments are a well-established approach in research about *Companion* technology (cf. [24, 43]) and human-computer interaction (HCI) in general. WoZ experiments allow us to investigate many of the open issues in user-*Companion* interaction (UCI) without the need to actually have to implement the functionalities of the envisaged *Companion*-System [44]. Examples of such questions that demand empirical answers include the following: How do 'naive' users spontaneously interact with a *Companion*-System if it allows them to converse in spoken natural language? Can distinct user groups be detected based on observed behavior? How do observed linguistic markers correlate with sociodemographic or psychometric data of the users?

Multimodal recordings from such WoZ experiments are valuable assets, but their impact remains limited if they are not prepared for and not made available for third-party usage within the research community in the form of a corpus. To convert raw data records from an experiment into a corpus usable as research resource is by no means a trivial and easy task. On the contrary, this task is both challenging conceptually and expensive with respect to time and effort needed. This is one of the reasons why publicly accessible corpora with naturalistic human-computer interactions are still rare exceptions (cf. [10]).

Converting raw data recorded in experiments into a corpus demands at least two major steps: *transcription* and *annotation*. In transcription audio records from the interactions need to be converted into written records that are then amenable to analysis by methods of computational linguistics and corpus linguistics. Annotation is the process of adding interpretative labels to the recorded data. Annotations may serve multiple purposes. They may be used in further analyses of the data or they may serve as input to machine learning procedures that are, for example, employed in training and testing of respective classifiers.

Research Questions The LAST MINUTE Corpus (LMC) is one of the rare examples of a corpus with naturalistic human-computer interactions. It offers richly annotated data from $N_{total} = 133$ experiments in a number of modalities (cf. Chap. 13). The LMC thus allows researchers from many disciplines to investigate research questions from a multitude of perspectives and with a plethora of approaches and methods. In this paper we exemplify these options with example

investigations and their results from three independent, yet cooperating, groups. The following research questions will be addressed:

- How do user groups based on sociodemographics (age, gender) differ with respect to linguistic aspects of the interaction and especially in global dialog success (cf. Sect. 14.4.1)?
- How do user groups that are defined based on distinct behavior during the experiment differ in global dialog success (cf. Sect. 14.4.2)?
- Does the intervention show effects in the course of the dialogs (cf. Sect. 14.4.3)?
- How do human raters annotate the emotional content of selected audio and video excerpts from the corpus (cf. Sect. 14.3.3)?
- What is the extent of improvement of classifier performance when classifiers are trained separately for the four age- and gender-based subgroups (cf. Sect. 14.4.4)?
- How do the age- and gender-based subgroups differ with respect to the use of discourse particles before and after the weight limit barrier (cf. Sect. 14.4.4)?
- Do personality traits of subjects influence the use or non-use of discourse particles before and after the weight limit barrier (cf Sect. 14.4.5)?

The investigations differ not only in the methods and perspectives but also in the primary modalities that are employed (e.g. transcripts, audio, questionnaire data) and in the size of the subcohorts of subjects included. The latter ranges from a small sample of just 13 sets of excerpts in an investigative study with human raters (cf. Sect. 14.3.3) to the full set of $N_{total} = 133$ verbatim transcripts (cf. Sects. 14.4.1 and 14.4.2).

14.2 Material: LAST MINUTE Corpus

The experiment that is underlying the multimodal recordings in the LAST MINUTE Corpus (LMC) was designed in such a way that the dialogs between the simulated system and the users were on the one hand restricted enough but on the other hand still offered enough room for individual variation [8, 31]. The domain chosen was mundane enough not to demand any specialist knowledge as a prerequisite. On the other hand an inherent need for re-planning (unpacking after weight limit) and for strategy change (from summer to winter items after the weather information barrier WIB) was built into the scenario.

After having completed about two thirds of the experiment, participants received additional information from the computer system, calling into question the way they handled the task so far. This so-called "weather information barrier" (`WIB`) represented a complex set of problems [9], because participants had to consider a large number of interacting variables and had no insights regarding the dynamics of the course of the experiment. As a result of undergoing the `WIB`, participants had to adapt their strategy to the new circumstances. Subsequent to the `WIB`, a randomly selected part of the participants received an affect-oriented intervention, the design

of which was based on general factors of psychotherapy (resource activation, problem actualization, accomplishment and clarification) [11]. Prior studies have already shown that empathic interventions initialized by computer systems can alter affective states that interfere with processes of communication (cf. [4, 16, 21]).

The WoZ scenario of LAST MINUTE is described in Chap. 13 in detail and with transcripts of example interactions.

The investment into the careful design of the scenario of the LAST MINUTE experiments pays off now. The resulting LMC is a valuable resource based on a large number of highly formalized, yet still variable, experiments with subjects balanced with respect to gender and age group.

As a resource the LMC is 'middle ground' between data (or a corpus) from a (small)scale experiment with a single hypothesis only and a corpus based on recordings from virtually unrestricted real-life interactions (e.g. Vera am Mittag [13], with records from a German TV talk show).

14.3 Methods

14.3.1 Analysis of Transcripts

Discourse Analysis The LAST MINUTE corpus comprises transcripts of all N = 133 experiments performed. On average, each experiment takes approximately 30 min real time. In order to be able to quantitatively compare and contrast different dialog courses, an adequate representation is needed [32, 33].

We employ a dialog representation based on the series of subsequent dialog acts of user and system, the so-called dialog act representation (DAR, [32]). This level of representation is independent of the domain of discourse, i.e. it is by no means restricted to the task in LAST MINUTE but is applicable to all types of task-oriented user-*Companion* dialogs.

In the following we use the dialog success measures (DSMs) as defined in Chap. 13. They allow the following types of investigations: How do user groups based on sociodemographics differ in global dialog success (cf. Sect. 14.4.1)? How do user groups that are defined based on distinct behavior during the experiment differ in global dialog success (cf. Sect. 14.4.2)?

The methods employed in discourse analysis of the LMC are as follows: The transcripts are available as an XML-based data structure in the FOLKER format [36]. This highly structured format contains not only the transcription of all user and wizard contributions during each experiment plus their relative temporal order, it comprises also additional annotations ranging from recorded nonphonological events (e.g. sighing, coughing) up to annotations on the discourse level (e.g. dialog act labels). For further details, cf. Chap. 13 or [32].

Starting from the FOLKER-encoded transcripts we determine features (or markers) either for complete transcripts or for their subparts (e.g. personalization

Table 14.1 Examples of empirical distributions of features calculated for complete transcripts (N = 133)

Marker	Min.	First Qu.	Median	Mean	Third Qu.	Max.	SD	Total sum
Tokens	266.0	444.0	545.0	602.7	699.0	1601.0	247.34	80,160
Turns	62	81	86	86.08	91	111	9.95	11,448
Tokens per turn	2.804	5.143	6.282	7.060	8.109	19.290	2.95	n.a.

vs. problem solving or their resp. subphases). Such features are calculated on all levels of the linguistic system, i.e. from the lexical level (e.g. occurrence counts for classes of lexical items) via syntax (e.g. preferred syntactic style in user commands) to semantic classifications (e.g. local meaning of user utterances) and pragmatic concerns (e.g. can the user's current intention be detected?).

The feature sets derived in this way then undergo a thorough data analysis in which we combine quantitative and qualitative approaches from corpus linguistics [12]. The quantitative methods start with the empirical distributions of the feature values. These are visualized appropriately and tested with respect to normality vs. skewness. Transcripts of (extreme) outliers are additionally checked qualitatively in order to detect possible reasons for the deviations.

A repeating finding for virtually all investigated features is that the distributions of feature values show a large variance. This even holds for features that quantify aspects of the overall extent of the highly standardized experiments (cf. Table 14.1).

Analyzing the reasons for the observed variance is a major issue in the work reported here. The different user groups based on sociodemographic features—i.e. age group (young subjects vs. elderly subjects) and the four combinations of age group with gender—are a primary potential source for the observed variance. Indeed, for many features the differences between the age groups and for subgroups based on subconditioning with gender prove to be significant (cf. Sect. 14.4.1).

When significant differences in the distribution of feature values have been found between sociodemographic groups, then the additional question arises about whether these differences correlate with significant differences in dialog success (as measured with DSM1 and DSM2).

Behavioral Analysis In behavioral analyses, errors that users make and problems they run into are valuable assets. This holds especially when early occurrences of problematic user behavior prove to be predictive for later global dialog success or failure. As will be elaborated in Sect. 14.4.2, early errors in the personalization phase can be detected that have this predictive power. The data analysis methods employed in evaluating observed differences in user behavior are the same as presented above. The only difference is that user groups are now defined on *observed differences in behavior in the course of the dialogs* and no longer on a priori differences between subjects like age group or gender.

14.3.2 Analysis of Psychometric Data

The authors examined the effectiveness of an affect-oriented intervention, which was given to participants after a confrontation with a complex set of problems, and in addition investigated its influence on participants' interaction behavior. In contrast to the rest of the experiment, where the interaction was guided by the ideational metafunction [14], the intervention was designed to address the participant on a conversant interactional level (interpersonal and textual metafunction). Therefore, the authors analyzed the influence of interpersonal problems on the effectiveness of the intervention using questionnaires and a self-developed criterion ('the dialog exchange'). Psychological questionnaires are broadly used research instruments for data acquisition. Generally, traits are measured via single questions (items) and potential answers are differentiated via Likert scales. Single items are summarized in subscales.

The **Inventory of Interpersonal Problems (IIP-C)** [17] measures problems which occur within interpersonal relationships. Applying the interpersonal circumplex model helps to assess behavior that is problematic for the test person as well as behavior he or she tends to show excessively. Eight scales are used for evaluation: domineering/controlling, vindictive/self-centered, cold/distant, socially inhibited, nonassertive, overly accommodating, self-sacrificing, intrusive/needy. These eight scales are in accordance with the octants of an circumplex model of interpersonal behavior, traits, and motives. The IIP-C was used for analyses presented in Sects. 14.4.3 and 14.4.5.

The **dialog exchange**. Analysis of the linguistic interactions in the WoZ experiment is possible by focusing more closely either on content- or on conversation-related aspects. Considering the research questions, an analysis of the conversation dynamics as such seems reasonable. Thus, the dialog exchange criterion was conceptualized in reference to the dialogism of interpersonal interactions. In conversation analysis, dialog is characterized by the 'boundaries of utterances', which are determined by aspects like 'change of speakers' (which is a fundamental characteristic of spoken language) [3] as well as the internal closure of single speaker contributions. In the WoZ experiment, the number of verbal contributions (so-called 'logs') given by the simulated system was recorded automatically [31]. The system was designed to respond to the participant. With the help of recorded logs we were able to determine the number of changes in the dialog between the system and the participant (dialog exchange criterion). Neither the content nor the length of utterances was considered.

The authors investigated the **following questions**. What is the impact of an affect-oriented intervention on participants' interaction behavior (dialog exchange) after a complex problem situation (barrier)? How does the extent of interpersonal problems influence the effectiveness of the affect-oriented intervention?

To answer these questions, the authors applied a range of different **methods**. According to the standardized experimental scenario, all subjects had to pass the identical procedure, which allowed for an exact definition of the course of events

before and after the barrier. Initially, the experiment was divided into four parts. In the baseline condition (BSL), the participants who accomplished the task proceeded without any further limitations. The first momentous limitation was the weight limit barrier (WLB), which the participants did not expect to occur. Later in the experiment, the system provided weather information (WIB) prompting participants to change their strategy. In the revision stage (RES), subjects got the opportunity to repack their suitcase under increasing time pressure [8]. In order to gather further information on the effect of the intervention on the dialog exchange, the authors compared the intervention and control group before and after the barriers. For statistical analysis, repeated measures ANOVA's has been used to test the effects of different independent variables on the dialog exchange. We conducted one within subjects ANOVA to test only the effects of the different conditions over time (BSL, WLB, WIB and RES) and used the Greenhouse Geisser correction of degrees of freedom when a significant Mauchly Test indicated lack of sphericity.

14.3.3 Analysis of Audio Records

For a realistic scenario, the development of the interaction is important and the users' reaction within critical events has to be assessed. Therefore, it has to be confirmed that the users show emotional reactions after the experimental barriers and that this reaction is different for the different kinds of barriers. This is later used to assess the type of barrier a user is allocated by using his acoustics (cf. Sect. 14.4.4).

To **evaluate the emotional content** right after the barriers, we created short excerpts for all four events containing video and audio utilizing a subset of $N_{labeling} = 13$ speakers from the LMC.

These clips are given to the labelers, who should rate each clip. The used labels are inspired by a previous experiment, as described in [38]: *surprise*, *interest*, *relief*, *joy*, *contempt*, *confusion*, *sadness*, *hope*, and *helplessness*. The labelers can choose between one of these predefined labels, but are also allowed to not give any label to a clip or to give a self-defined label. Six labelers, all not familiar with the corpus, conducted that labeling task. This results in the distribution of labels given in Fig. 14.1. It reveals that the dialog phase after each barrier has its own distribution of several emotional states. This shows that the experimental barriers evoke different reactions by the users. The distribution of emotional states after each barrier confirms the expected reaction. To select the barriers worth for later automatic analyzes, the amount of the user's speech data has to be taken into account. As for CLB and WIB, the user is hardly involved as only information is presented; further experiments are conducted between BSL and WLB. Further details can be found in [37, 39].

Furthermore, we analyze **discourse particles (DPs) as an interaction pattern**. During Human–Human interaction (HHI) several semantic and prosodic cues are exchanged between the interaction partners and used to signalize the progress of

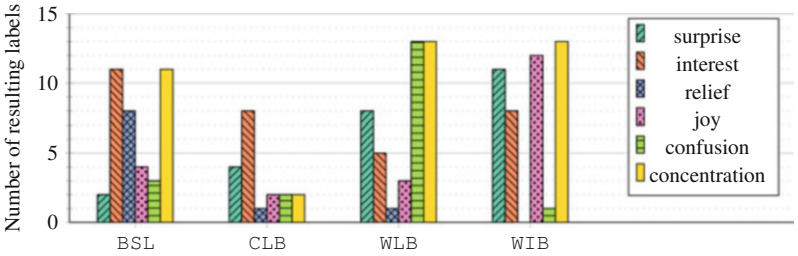

Fig. 14.1 Distribution of emotions over the dialog barrier phases of LMC gathered by manual labeling on a subset of 13 subjects (cf. [39]). BSL denotes the baseline, CLB the listing barrier, WLB the weight limit barrier, and WIB the weather information barrier

the dialog [1]. Especially, the intonation of utterances transmits the communicative relation of the speakers and also their attitude towards the current dialog. Thus, DPs can be seen as pattern exposing information about the current interaction.

Furthermore, it is assumed that these short feedback signals are uttered in situations of a higher cognitive load [5] where a more articulated answer cannot be given. As, for instance, stated in [23, 35], specific monosyllabic verbalizations, the DPs, have a specific intonation. In [35] it is stated that DPs like "hm" or "uhm" cannot be inflected but can be emphasized and are occurring at crucial communicative points. The DP "hm" is seen as a "neutral consonant" whereas "uh" and "uhm" can be seen as "neutral vocals". The intonation of these particles is largely free of lexical and grammatical influences. Schmidt called that a "pure intonation" [35].

Additionally, an empirical study of German is presented in [35] determining seven form-function relations of the DP "hm" due to listening experiments. Several studies confirmed the form-function relation for HHI; cf. [20, 27]. In Sects. 14.4.4 and 14.4.5 it is investigated whether these cues are also used within HCI and can serve an indicator for critical parts of the dialog. Furthermore, influences on the usage of DPs are analyzed.

14.4 Results

14.4.1 Discourse Analysis: Age and Gender Matters

Differences in Verbosity The ratio of tokens per turn (TpT) is an adequate verbosity measure for dialogs. Given the different nature of the different phases in the experiment, the measure varies between more narrative-oriented phases and phases with a preference for usually shorter commands (Fig. 14.2).

As a major result for problem solving (without intervention) we get that age group matters and that young subjects are significantly less verbose than elderly one

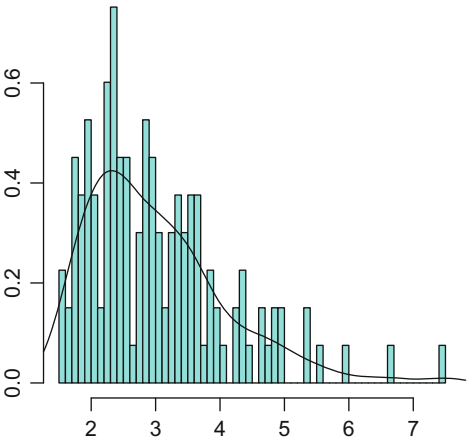

Fig. 14.2 Distribution of tokens per turn (TpT) ratios for problem solving without intervention ($N = 133$)

(cf. Fig. 14.3,[1] Table 14.2; Wilcoxon: $W = 1722$, $p = 0.03251$*, $d_{Cohen} = 0.24$). In contrast, gender gives insignificant differences only. In addition, the pairings of age group and gender result in significant differences as well (Kruskal-Wallis chi-squared = 8.375, $df = 3$, $p = 0.03886$*).[2] Similar results hold for TpT values for other parts of the experiment. A case in point is, for example, the narratives phase in personalization (cf. Table 14.2).

Politeness Particles as Indicators for CASA When humans conversing with a computer system do employ politeness particles when they address the system, this can be seen as an indicator for (mindlessly) treating Computers as Social Actors (CASA, [26]).

Counting the number of occurrences of politeness particles *'bitte'* (Engl. *'please'*) and *'danke'* (Engl. *'thank you'*) in user utterances per transcript, we get distributions for all N = 133 subjects as depicted in Figs. 14.4 and 14.5. Note: 55 subjects have *not a single occurrence* of one of these politeness particles and the median for all subjects lies at one occurrence.

Again, age matters. The subgroup above the median of counts of used politeness particles is clearly dominated by elderly subjects, whereas the subgroups at the median and below the median are dominated by young subjects (cf. Table 14.3).

Tests show significant differences for the two age groups (Wilcoxon W = 1138, $p = 7.078e-07$***, $d_{Cohen} = 0.51$) and the four pairings of gender and age group

[1]The distributions are visualized—here and in other figures—as trellis box plots: 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. [2]).

[2]Unless noted otherwise, all statistical tests and calculations have been performed with the R language [30] [2]. Significance levels are denoted by *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

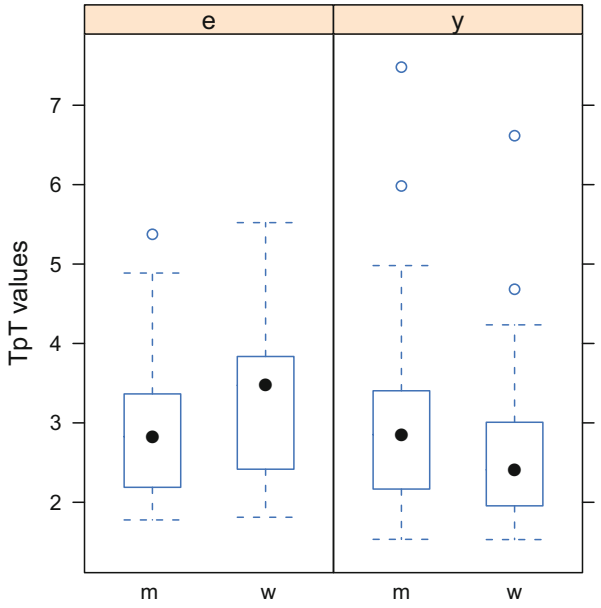

Fig. 14.3 Distributions of tokens per turn (TpT) ratios for problem solving conditioned by gender and age group ($N = 133$; m = men, w = women; e = elder, y = young)

Table 14.2 Differences in mean TpT values for sociodemographic groups

Marker	g1	Rel	g2	p value	Test	d_{Cohen}
TpT problem solving	y	lt	e	0.03251*	Wilcoxon	0.24
TpT problem solving	m	lt	w	n.s.	Wilcoxon	n.a.
TpT pers. narratives	y	lt	e	0.00369**	Wilcoxon	0.47
TpT pers. narratives	w	lt	m	n.s.	Wilcoxon	n.a.

(Kruskal chi-squared = 26.0632, df = 3, $p = 9.251e-06^{***}$), but for gender we get insignificant differences only. The most significant and largest pairwise difference between subgroups with respect to the use or non-use of politeness particles is between young women and elderly women (Wilcoxon W = 204, $p = 1.688e{-}06^{***}$, $d_{Cohen} = 0.72$; cf. Fig. 14.5).

The Impact of Age and Gender Young subjects on average use none or significantly fewer politeness particles, they do employ significantly fewer tokens per turn (TpT) in personalization narratives and in problem solving than elderly subjects. In addition, young subjects on average are significantly more successful than elderly subjects with respect to the different dialog success measures (cf. Chap. 13).

When all women vs. all men are contrasted, gender makes no global significant differences with respect to use of politeness particles, tokens per turn (TpT) and differences in control (i.e. wizard-induced category changes; cf. Sect. 14.4.2). The

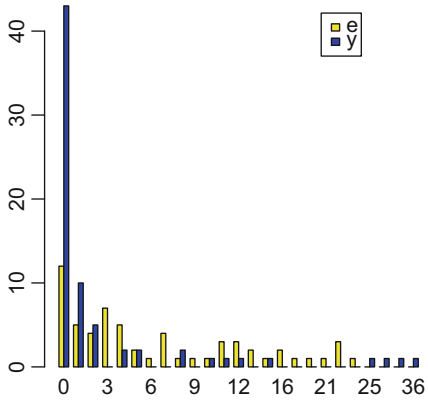

Fig. 14.4 Subgroup comparison for young (*dark bars*; N = 72) vs. elderly subjects (*light bars*; N = 61): Distributions of number of occurrences of politeness particles in user utterances per transcript (please note the number of zero occurrences)

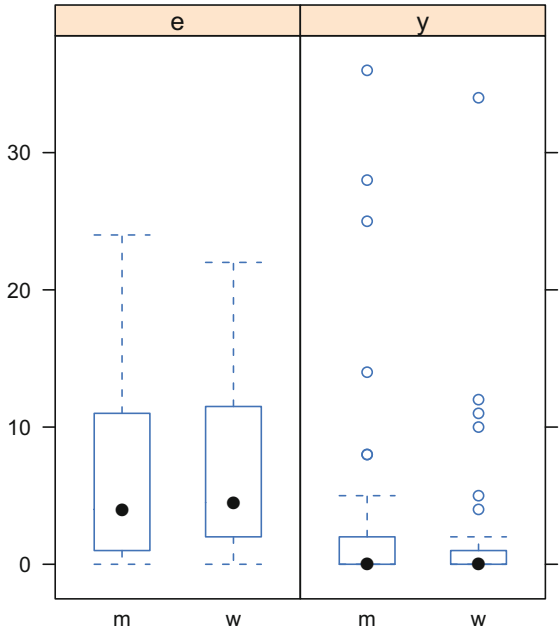

Fig. 14.5 Distributions of number of occurrences of politeness particles in user utterances per transcript, conditioned by age group and gender ($N = 133$; m = men, w = women; e = elder, y = young. Please note the zero medians for the young subgroups and the outliers)

Table 14.3 Distributions of gender and age group pairings for subgroups of subjects in relation to median of used politeness particles

Subgroup	*em*	*ew*	*ym*	*yw*	Sum
Above median	**19**	**25**	11	8	63
AT median	3	2	**6**	**4**	15
Below median	7	5	**19**	**24**	55

Boldface is used to emphasize the two largest values in each row of this table

differences in means with respect to the dialog success measures (with larger mean values for women) are not significant. But there are significant differences between some of the age group and gender-defined subgroups of subjects. Young women, for example, are significantly more successful (in both dialog success measures) than young men, elderly men and elderly women. Pairwise differences between the latter three subgroups are not significant.

14.4.2 Behavioral Analyses

Early Problems with 'Tell and Spell' At the very beginning of the personalization phase all subjects are prompted:

```
Bitte nennen und buchstabieren Sie zunächst Ihren Vor- und Zunamen!
Please tell and spell your first name and surname!
```

Some subjects need several trials; some even completely fail to provide the requested information. From $N = 133$ subjects the answer to the prompt 'Tell and spell ...' is accepted after the first answer for 113 subjects, after the second trial for 12 subjects and after the third trial for 8 subjects. Actually the task completion ratio is even worse: 20 subjects only *spell* but do not tell their name; two more leave the first name out. (Note: wizards did not react on these latter types of incomplete answers.) In sum: from $N = 133$ subjects the answer to the prompt 'Tell and spell ...' is wrong or incomplete in at least 34 cases (i.e. 25.6%); full task completion is in only 74.4% of the cases.

The age groups differ with respect to task completion: exactly two spelling request are needed by five elderly and seven young subjects, whereas the eight subjects with exactly three trials are all elderly.

Why should 'tell and spell ...' be a problem? The failure of subjects with respect to this task may be attributed to 'inattentional deafness' [7] or to effects of cognitive aging [44] in general. This leads to the following **hypothesis**: Subjects with problems with the 'tell and spell ...' task will have problems with other parts of the experiment as well and will have lower values in the dialog success measures.

To test this hypothesis we contrast the distribution of dialog success measures for the no problem group (i.e. exactly one trial) and the complementary problem group (i.e. with two or more trials).

The difference in mean for DSM2 (no problem: 0.7075; problem: 0.6612) is significant as a Wilcoxon test reveals (W = 773, $p = 0.02482^*$, $d_{Cohen} = 0.56$; the distribution of the no problem group clearly differs from a normal distribution).

Similar results hold for DSM1: the problem group has poorer dialog success values and—again—these differences between the no problem and the problem group are significant (Wilcoxon: W = 770.5, $p = 0.02382^*$, $d_{Cohen} = 0.4993$).

In sum, problems with the very first task in personalization are an early predictor of later problems in the problem solving dialog of LAST MINUTE proper.

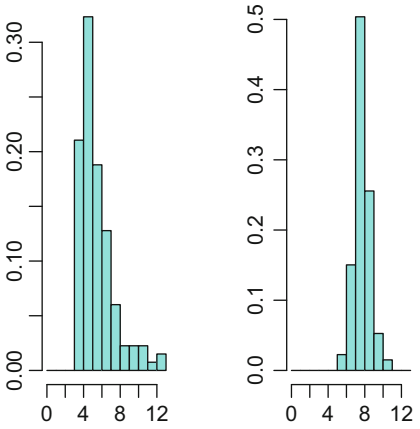

Fig. 14.6 Distributions of total number of user turns in subphases of personalization per transcript ($N = 133$): data acquisition (*left*), narratives (*right*)

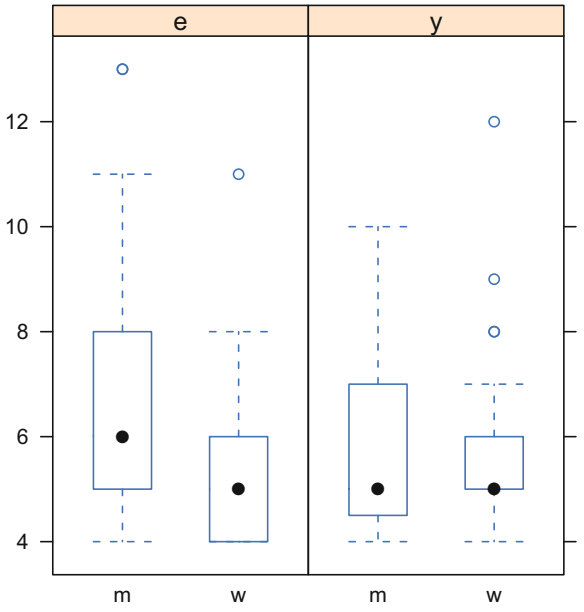

Fig. 14.7 Distributions of total number of user turns in data acquisition per transcript, conditioned by age group and gender ($N = 133$; m = men, w = women; e = elder, y = young)

Early Predictor: Data Acquisition in Personalization In the personalization phase, initiative lies primarily with the system. Here a typical adjacency pair is made up of a wizard prompt or question followed by a user narrative or answer.

In cases of a normal dialog course the sources of variation are reprompts (e.g. 'tell and spell'), the number of questions of the 'bitte ergänzen sie angaben zu …' type and the number of prompts for 'more detail'. Sources of variation in unforeseen courses are user questions, e.g. caused by understanding problems (Fig. 14.6).

Note: more adjacency pairs in personalization, thus in general, indicate problems. The empirical distributions of the total number of user turns in data acquisition, conditioned by age group and gender, are depicted in Fig. 14.7.

In the following we perform a median split with respect to the total number of turns (i.e. adjacency pairs) in the subphase 'data'. The overall result: the subgroup of subjects below (and at) the median (of 5) has significantly better values for both dialog success measures in problem solving. For both dialog measures, Wilcoxon tests judge the differences between the groups as significant (DSM1: W = 1746, $p = 0.04035^{*}$, $d_{Cohen} = 0.265$; DSM2: W = 1604.5, $p = 0.00718^{**}$, $d_{Cohen} = 0.435$).

Issues of Control: Pauses as Indicators of Helplessness Being in control or not is an important issue in a dialog. In the LM experiments the issue of control is underlying the distinction between two types of category change: subject-induced category change (SICC, the subject explicitly utters a request for category change) vs. wizard-induced category change (WICC, the wizard enforces a category change).

More than 50% of the subjects are 'in control' in this sense. They have either zero or only one or two wizard-induced category changes (from a total of 14 category changes in a complete experiment). The complement of this group ('poor control') has between three and ten WICCs. Poor control of category changes (i.e. `WICCs > 2`) predicts poor global dialog success. The two subgroups—at and below the WICC median of 2 or above the WICC median, resp.—show significant differences in both global dialog success measures (DSM1: Wilcoxon test, W = 1667.5, $p = 0.02614^{*}$, $d_{Cohen} = 0.45$; DSM2: Wilcoxon test, W = 1139, $p = 3.610 \times 10^{-06***}$, $d_{Cohen} = 0.96$).

Again: age group makes a major difference between the two subgroups whereas gender differences are only of minor relevance.

Long Pauses There is a subgroup of subjects with poor control that—after some choices in a category—passively wait without any further action, sometimes for 40 s or longer, until the system finally enforces a category change (WICC).

Not surprisingly, the occurrence of such a type of long pause is again a predictor of global dialog failure. The subgroup of subjects that have at least one occurrence of a pause longer than 10 s before a WICC has significantly poorer dialog success measures when compared to the complementary group of subjects without such pauses. (DSM1: Wilcoxon test, W = 268, $p = 0.003109^{**}$, $d_{Cohen} = 1.71$; DSM2: Wilcoxon test, W = 138.5, $p = 4.87 \times 10^{-05***}$, $d_{Cohen} = 2.36$).

14.4.3 Results from Analysis of Psychometric Data

First of all, the influence of the intervention on the general experimental course was examined. The authors used repeated-measures ANOVA for data analysis to analyze the effects of interpersonal problems (IIP) on dialog exchange. One ANOVA was conducted with intervention (intervention $N = 62$ vs. control group $N = 68$) as

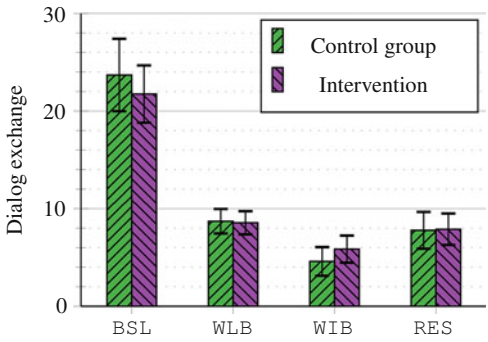

Fig. 14.8 Dialog exchange (control and intervention group) during the experimental course

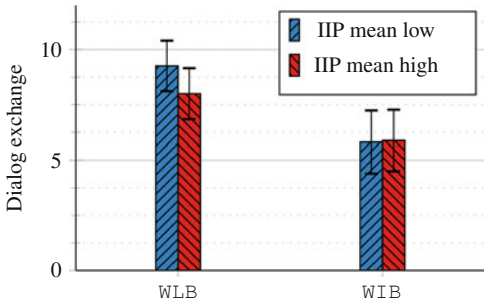

Fig. 14.9 Dialog exchange considering the level of interpersonal problems (high vs. low) between WLB and WIB

between subject factor. This revealed a significant interaction of intervention on dialog exchange over time ($F(1.89, 242.36) = 3.39$, $p < 0.038$), indicating that subjects who had received an intervention had a higher level of dialog exchange. An observation of dialog exchange over time revealed the difference between intervention and control group at the BSL already ($d_{Cohen} = 0,29$); see Fig. 14.8. At this point, the course did not differ in both groups yet.

The next step was to examine the impact of interpersonal problems on the effectiveness of a system-initiated intervention. Averaged over all test intervals, this revealed no statistical significant difference ($F(1, 60) = 2.02$, $p < 0.160$) between the extent of interpersonal problems (dichotomization by means of a median split led to groups with high vs. low levels of interpersonal problems) and the response behavior of a system-initiated intervention. However, the descriptive account indicates that participants with pronounced interpersonal problems show a lower dialog exchange prior to WIB and intervention ($d_{Cohen} = 0,55$). This difference can't be identified after the intervention anymore ($d_{Cohen} = 0,03$). Both groups (IIP level high vs. low) show no statistically significant difference regarding the amount of dialog exchange. Figure 14.9 shows how the intervention seems to be more effective for participants with a high degree of interpersonal problems.

14.4.4 Results from Analysis of Audio Records

In this section, we present our acoustic analysis to identify whether a user has no problems within the interaction, or is experiencing a barrier. Therefore, we define a two-class problem and try to distinguish the dialog phases after the `BSL` and the `WLB` events, where the user should be set into a certain clearly defined condition. As we have seen from discourse analysis and psychometric data analysis and already shown for acted and spontaneous emotions, we can identify different speaker groups that behave differently during this naturalistic interaction. We therefore investigate whether the incorporation of user characteristics can improve the speech-based recognition. To apply our methods of Speaker Group Dependent (SGD)-modeling on LMC, we utilize the same age and gender groupings as in the previous sections: young vs. elderly subjects and men vs. women. The combination of both grouping factors led to four sub-groups: (`ym`, `em`, `yw`, and `ew`).

The emotional assessment of the dialog phases is described in Sect. 14.3.3. Due to the quite time-consuming work to generate the transcripts, these experiments could only be carried out on a subset of the LMC, containing just $N_{acoustic-HSDP2} = 79$ participants. As classification baseline, the Speaker Group Independent (SGI) set is used. It contains all 79 subjects. Age and gender of the speakers are known a priori on the basis of the subjects' transcripts. Different age-gender groupings together with the number of corresponding subjects are depicted in Fig. 14.10. Training and testing is based on the subjects' utterances of the two dialog phases after the `BSL` and `WLB`. These utterances are extracted automatically on the basis of the transcripts. This results in 2301 utterances with a total length of 31 min. Furthermore, the following acoustic characteristics are utilized as features: 12 Mel-Frequency Cepstral Coefficients, Zeroth cepstral coefficient, Fundamental frequency, and Energy. The Δ and $\Delta\Delta$ regression coefficients of all features are used to include contextual information. As channel normalization technique, RelAtive SpecTrAl (RASTA)-filtering is applied. Gaussian Mixture Models (GMMs) with 120 mixture components utilizing four iteration steps are used as classifiers. For validation we use a Leave-One-Speaker-Out (LOSO) strategy. As performance measure, the unweighted average recall (UAR) is applied.

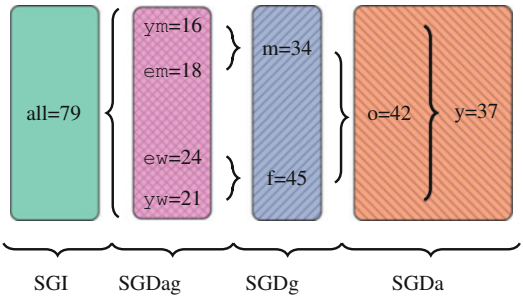

Fig. 14.10 Subjects' distribution into speaker groups (SG) on LMC. Abbreviations: I = independent set, D = dependent on a = age, g = gender

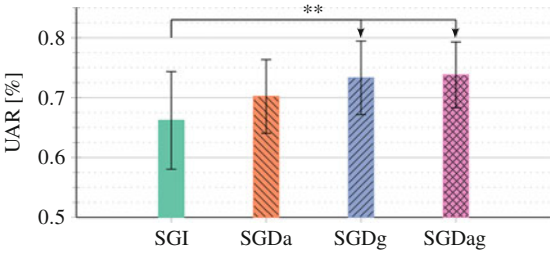

Fig. 14.11 Recognition results for SGI and different SGD configurations. The *stars* denote the significance level: **($p < 0.01$) using ANOVA

Afterwards, we performed the experiments on the SGI set as well as in sets grouped according to age (a) and gender (g), training the corresponding classifiers in an LOSO manner. To allow a comparison, we combined the different results of all speaker groupings. For instance, the results for each male and female speaker are put together to get the overall result for the SGDg set. This result can then be directly compared with results gained on the SGI set. The outcome is shown in Fig. 14.11.

The classification achieved with LMC shows that SGDag grouping can significantly outperform the SGI results with a rate of 73.3%. The improvement is significant for SGI to SGDg ($F = 8.706$, $p = 0.0032$, $d_{Cohen} = 0.492$) and SGI to SGDag ($F = 10.358$, $p = 0.0013$ $d_{Cohen} = 0.526$). Both comparisons are within the zone of desired effects, after Hattie [15]. When comparing the achieved UARs utilizing either age or gender groups, it can be seen that the gender grouping outperforms the age grouping. Further details can be found in [37, 41].

Discourse Particles as Interaction-Patterns in HCI In the following, we analyze whether discourse particles (DPs) can be seen as interaction patterns occurring at critical situations within an HCI. We start by using the whole session and analyzing global differences in DP usage. Afterwards, the local usage within significant situations is analyzed, referring to the `WLB` barrier. All investigations are performed utilizing $N_{acoustic-SH66} = 90$ subjects of LMC. Based on the transcripts, all DPs are automatically aligned and extracted, utilizing a manual correction phase. The extraction results in a total number of 2063 DPs, with a mean of 23.18 DP per conversation and a standard deviation of 21.58. This result shows that DPs are used in our HCI experiments, although the conversational partner, the technical system, was not enabled to express them or react to them. The average DP length is approx. 1 s±0.4 s. Only 2600 tokens from all 82,000 tokens represent DPs, illustrating the small number of uttered DPs. As a statistical test, an one-way non-parametric ANOVA is used to compare the means of our two median-split samples [22].

To provide valid statements on the DP usage in a naturalistic HCI within the different SGD groups, two aspects have to be taken into account. The first aspect is the verbosity, denoting the number of tokens a speaker has made during the experiment. We analyze both verbosity and DP usage over the dialog-phases starting after specified barrier-events. As a second aspect, the usage of DP depending on age and gender of the subjects is analyzed, analogously to our previous approach in affect recognition. We again use the same speaker grouping; see Fig. 14.10. From the

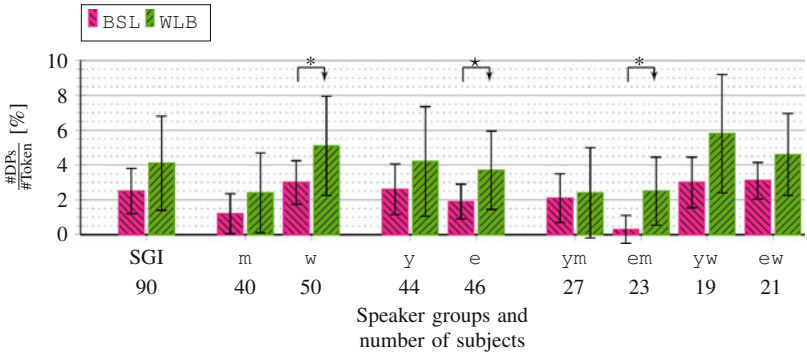

Fig. 14.12 Mean and standard deviation for the DP usage distinguishing the dialog phases after BSL and WLB regarding different speaker groups in LMC. For comparison, the group-independent frequency (SGI) is given. The significance level with ($p < 0.05$) is denoted with *, *star* denotes close proximity to significance level. The effect size is within the zone of desired effect according to [15], using d_{Cohen}

barriers' description, we assume that a higher cognitive load due to the re-planning task WLB increases the DP usage, since DPs are known to indicate a high cognitive load (cf. [5]). For the analysis of this assumption, we calculated the relation of uttered DP and verbosity within the dialog phases after both dialog barriers and distinguished this from the previously used speaker groupings. Both comparisons are within the zone of desired effects, reaching from 0.4 to 1.0, after Hattie [15]. The results are depicted in Fig. 14.12.

Regarding the DP usage between the two dialog barriers BSL and WLB, it is apparent that for all speaker groupings the average number of DPs for WLB is higher than for BSL. This is significant for the speaker group w, and near the significance level in the speaker group e. This observation supports the statement from [29] that male users and young users tend to have less problems to overcome the experimental barriers, confirming the findings of Sect. 14.4.1 that the age and the gender of the subjects matter. Considering the combined age-gender grouping, only for the em grouping can a significant difference between BSL and WLB be observed. Thus, it can be summarized that DPs are capable of serving as interaction patterns indicating situations where the user is confronted with a critical situation in the dialog [40]. This investigation reveals the need to detect and interpret these signals (cf. [25, 28]).

14.4.5 Combined Analyses of the Impact of Personality Traits and Discourse Particles

As one can see from the previous section, particularly for the two groups yw and ew, the standard deviation for the DP usage after the WLB is quite high. This also indicates that other factors influence the individual DP usage. Therefore, we analyze

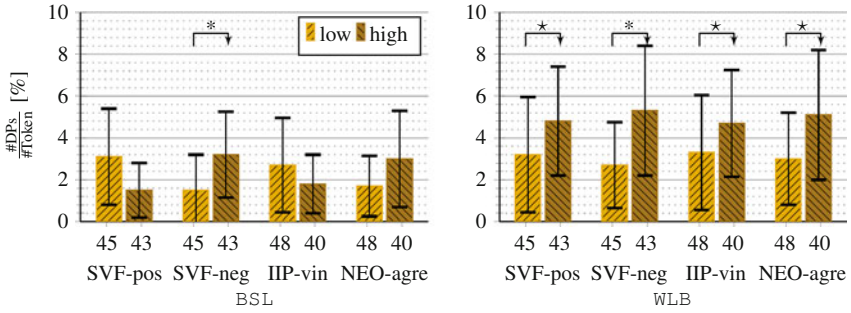

Fig. 14.13 Mean and standard deviation for the frequency of the DPs of the two barriers regarding different groups of user characteristics. The significance level with ($p < 0.05$) is denoted with *, *star* denotes close proximity to significance level using ANOVA

personality traits as additional kinds of factors. Among the gathered personality traits (cf. Chap. 13), we chose those which are in connection with stress coping. To analyze whether a specific personality trait influences DP usage, we utilized the same sample of $N_{acoustic-SH66} = 88$ and used a median split to distinguish between users with low traits (below median) and those with high traits (above median). The resulting numbers of subjects for each group are given in Fig. 14.13. We tested all personality traits available for the LMC, but report only those factors close to the significance level. These factors are determined by the following personality questionnaires: NEO Five-Factor Inventory (NEO-FFI) [6], Inventory of Interpersonal Problems (IIP-C) [18], and Stress-coping questionnaire (SVF) [19].

Considering each psychological trait, no significant differences are noticeable between the two dialog styles "personalization" and "problem solving". In addition to the analysis based on the two experimental phases, we also investigated the different usage of DPs between the dialog phases following the experimental events `BSL` and `WLB` regarding the personality trait factors `SVF-pos`, `SVF-neg`, `IIP-vin`, and `NEO-agre` (cf. Fig. 14.13). In this case, the `SVF-neg` factor shows significant results to distinguish between the low and high groups for both `BSL` ($H = 6.340$, $p = 0.012$, $d_{Cohen} = 0.452$) and `WLB` ($H = 4.617$, $p = 0.032$, $d_{Cohen} = 0.497$), whereas for `SVF-pos`, `IIP-vin` and `NEO-agre` users belonging to the high group show an increased DP usage after the `WLB` barrier that is close to the significance level (cf. [42]).

Thus, it can be assumed that "negative" psychological characteristics stimulate the usage of DPs. A person having a bad stress regulation capability will be more likely to use DPs in a situation of higher cognitive load [5] than a person having a good stress regulation capability.

The investigations presented in this section reveal that the occurrences of DPs can provide hints about specific situations of the interaction. We show that not just the mere occurrence of the DPs is essential, but also the context in which they are used. DPs are occurring more frequently in situations of a higher cognitive load and thus are an important interaction pattern. For the automatic usage of this phenomenon,

described in this chapter, obviously further steps, e.g. automatic DP allocation, are necessary. To this end, we developed a classifier to automatically detect occurrences of the DPs “äh” and “ähm” [28]. The next step will be the automatic assessment of the functional meaning of DPs, which will allow a detailed assessment of the context in which the DPs are used. Therefore, in [25] a rule-based algorithm is presented.

14.5 Discussion

We have reported on a number of investigations of three different research groups into data from the LMC and on their respective results. Some results have cross-fertilized other work. For example, the findings about the significant differences between age group- and gender-based subgroups of the whole cohort of subjects (cf. Sect. 14.4.1) have inspired the experiments with training of subgroup-specific classifiers (cf. Sect. 14.4.4) and the investigation of differences between these subgroups with respect to use of discourse particles (cf. Sect. 14.4.4). Some investigations yielded negative results. For example, no significant effect of the intervention has been found in the data of the LMC (cf. Sect. 14.4.3 and Chap. 13). The strongest results—both with respect to significance levels as well as effect sizes—were achieved from the in-depth behavioral analyses of the transcripts (cf. Sect. 14.4.2).

Major Insights from Analyses User groups based on sociodemographics matter in the LMC. This holds especially for the differences between young and elderly subjects, with the former being more successful *on average*. On the other hand, gender matters only when taken into account as a further subcondition after an age group-based primary grouping.

One reason for communication problems in the LMC is that some subjects have difficulties comprehending and memorizing information that was given as spoken language utterances by the system. Such problems occur significantly more often with elderly subjects. Early occurrences of such problems in speech understanding are a strong predictor of global failure of the (independent) later problem solving dialog (cf. Sect. 14.4.2). Another strong indicator of a potential user problem is an overly long pause when the user actually has the turn, i.e. the right for the next utterance (cf. Sect. 14.4.2).

Design Considerations for *Companion*-Systems The findings from the analyses of the dialogs in the LAST MINUTE corpus may contribute to design considerations for *Companion*-Systems that are based on speech interaction with their users [34]. On the one hand, differences between sociodemographic groups—especially differences between age groups—have to be taken into account by the dialog management of *Companion*-Systems. On the other hand, the broad variance between individuals [44] demands personalized calibration of dialog management strategies. Tests that are easy to perform and evaluate and that have large predictive power for potential problems in the subsequent global dialog course—cf. Sect. 14.4.2—may be employed for this purpose. In addition the dialog history of the user-*Companion*

interactions needs to be monitored continuously. Special emphasis shall be given to situations where the user has the turn but does not take it within a certain time span. As discussed in Sect. 14.4.2, such overly long pauses are strong indicators of problems and possibly helplessness on the user's side, and demand an adequate response by the system.

Contributions: The results reported in this paper have been contributed by three different groups. The responsibility for discourse analysis and behavioral analysis of transcripts (e.g. Sects. 14.4.1 and 14.4.2) lies with D. Rösner, Th. Bauer and St. Günther; for analysis of psychometric data (e.g. Sect. 14.4.3) responsibility lies with J. Frommer and M. Haase; and for analysis of audio records (e.g. Sect. 14.4.4) responsibility lies with A. Wendemuth and I. Siegert. For the discussion and the conclusions in Sect. 14.5, the responsibility is shared by all authors.

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. Allwood, J., Nivre, J., Ahlsén, E.: On the semantics and pragmatics of linguistic feedback. J. Semant. **9**, 1–26 (1992)
2. Baayen, R.: Analyzing Linguistic Data – A Practical Introduction to Statistics Using R. Cambridge University Press, Cambridge (2008)
3. Bakhtin, M.M., Holquist, M., McGee, V., Emerson, C.: Speech Genres and other Late Essays, vol. 8. University of Texas Press, Austin (1986)
4. Bickmore, T., Gruber, A., Picard, R.: Establishing the computer–patient working alliance in automated health behavior change interventions. Patient Educ. Couns. **59**(1), 21–30 (2005)
5. Corley, M., Stewart, O.W.: Hesitation Disfluencies in Spontaneous Speech: the Meaning of *um*. Lang. Ling. Compass. **2**, 589–602 (2008)
6. Costa, P.T., McCrae, R.R.: Domains and facets: hierarchical personality assessment using the revised NEO personality inventory. J. Pers. Assess. **64**, 21–50 (1995)
7. Dalton, P., Fraenkel, N.: Gorillas we have missed: sustained inattentional deafness for dynamic events. Cognition **124**(3), 367–372 (2012)
8. 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)
9. Funke, J.: Complex problem solving: a case for complex cognition? Cogn. Process. **11**(2), 133–142 (2010)
10. Georgila, K., Wolters, M., Moore, J., Logie, R.: The MATCH corpus: a corpus of older and younger users' interactions with spoken dialogue systems. Lang. Resour. Eval. **44**(3), 221–261 (2010)
11. Grawe, K.: Grundriß einer Allgemeinen Psychotherapie. Psychotherapeut **40**, 130–145 (1995)
12. Gries, S.T.: Quantitative Corpus Linguistics with R: A Practical Introduction. Routledge, Abingdon (2009)
13. Grimm, M., Kroschel, K., Narayanan, S.: The Vera am Mittag German audio-visual emotional speech database. In: Proceedings of the 2008 IEEE ICME, pp. 865–868 (2008)
14. Halliday, M.A.K: Language as a Social Semiotic: The Social Interpretation of Language and Meaning. Hodder & Stoughton Educational, London (1976)
15. Hattie, J.: Visible Learning. A Bradford Book. Routledge, London (2009)

16. Hone, K.: Empathic agents to reduce user frustration: the effects of varying agent characteristics. Interacting Comput. **18**(2), 227–245 (2006)
17. Horowitz, L.M., Alden, L.E., Wiggins, J.S., Pincus, A.L.: Inventory of interpersonal problems manual. Psychological Cooperation, San Antonio (2000)
18. Horowitz, L.M., Strauß, B., Kordy, H.: Inventar zur Erfassung interpersonaler Probleme (IIPD), 2nd edn. Beltz, Weinheim (2000)
19. Jahnke, W., Erdmann, G., Kallus, K.: Stressverarbeitungsfragebogen mit SVF 120 und SVF 78, 3rd edn. Hogrefe, Göttingen (2002)
20. Kehrein, R., Rabanus, S.: Ein Modell zur funktionalen Beschreibung von Diskurspartikeln. In: Neue Wege der Intonationsforschung, Germanistische Linguistik, vol. 157–158, pp. 33–50. Georg Olms, Hildesheim (2001)
21. Klein, J., Moon, Y., Picard, R.W.: This computer responds to user frustration: theory, design and results. Interacting Comput. **14**, 119–140 (2002)
22. Kruskal, W., Wallis, W.A.: Use of ranks in one-criterion variance analysis. J. Am. Stat. Assoc. **47**, 583–621 (1952)
23. Ladd, R.D.: Intonational Phonology. In: Studies in Linguistics, vol. 79. Cambridge University Press, Cambridge (1996)
24. Legát, M., Grůber, M., Ircing, P.: Wizard of oz data collection for the czech senior companion dialogue system. In: Fourth International Workshop on Human-Computer Conversation, pp. 1–4. University of Sheffield (2008)
25. Lotz, A.F., Siegert, I., Wendemuth, A.: Automatic differentiation of form-function-relations of the discourse particle "hm" in a naturalistic human-computer interaction. In: Proceedings of the 26th ESSV, Eichstätt (2015)
26. Nass, C., Moon, Y.: Machines and Mindlessness: Social Responses to Computers. J. Soc. Issues **56**(1), 81–103 (2000)
27. Paschen, H.: Die Funktion der Diskurspartikel HM. Master's thesis, University Mainz (1995)
28. Prylipko, D., Egorow, O., Siegert, I., Wendemuth, A.: Application of image processing methods to filled pauses detection. In: Proceedings of INTERSPEECH'14, Singapore (2014)
29. Prylipko, D., Rösner, D., Siegert, I., Günther, S., Friesen, R., Haase, M., Vlasenko, B., Wendemuth, A.: Analysis of significant dialog events in realistic human-computer interaction. J. Multimodal User Interfaces **8**(1), 75–86 (2014)
30. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2010). http://www.R-project.org
31. Rösner, D., Frommer, J., Friesen, R., Haase, M., Lange, J., Otto, M.: LAST MINUTE: a multimodal corpus of speech-based user-companion interactions. In: Proceedings of the 8th LREC, Istanbul, pp. 96–103 (2012)
32. Rösner, D., Friesen, R., Günther, S., Andrich, R.: Modeling and evaluating dialog success in the LAST MINUTE corpus. In: Proceedings of the 9th LREC, Reykjavik (2014)
33. Rösner, D., Andrich, R., Bauer, T., Friesen, R., Günther, S.: Annotation and Analysis of the LAST MINUTE corpus. In: Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology, pp. 112–121 (2015)
34. 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. Künstliche Intelligenz **30**(1), 53–61 (2016). Online first: Oct 28 (2015). doi:10.1007/s13218-015-0410-z
35. Schmidt, J.E.: Bausteine der Intonation. In: Neue Wege der Intonationsforschung, Germanistische Linguistik, vol. 157–158, pp. 9–32. Georg Olms, Hildesheim, Germany (2001)
36. Schmidt, T., Schütte, W.: Folker: An annotation tool for efficient transcription of natural, multi-party interaction. In: Proceedings of the 7th LREC, Valletta (2010)
37. Siegert, I.: Emotional and user-specific cues for improved analysis of naturalistic interactions. Ph.D. thesis, Otto von Guericke University Magdeburg (2015)
38. Siegert, I., Böck, R., Philippou-Hübner, D., Vlasenko, B., Wendemuth, A.: Appropriate Emotional Labeling of Non-acted Speech Using Basic Emotions, Geneva Emotion Wheel and Self Assessment Manikins. In: Proceedings of the 2011 IEEE ICME, Barcelona (2011)

39. Siegert, I., Böck, R., Wendemuth, A.: The influence of context knowledge for multimodal annotation on natural material. In: Joint Proceedings of the IVA 2012 Workshops, pp. 25–32. Santa Cruz (2012)
40. Siegert, I., Hartmann, K., Philippou-Hübner, D., Wendemuth, A.: Human behaviour in HCI: complex emotion detection through sparse speech features. In: Salah, A., Hung, H., Aran, O., Gunes, H. (eds.) Human Behavior Understanding. Lecture Notes on Computer Science, vol. 8212, pp. 246–257. Springer, Berlin (2013)
41. Siegert, I., Böck, R., Wendemuth, A.: Inter-Rater Reliability for Emotion Annotation in Human-Computer Interaction – Comparison and Methodological Improvements. J. Multimodal User Interfaces **8**, 17–28 (2014)
42. Siegert, I., Haase, M., Prylipko, D., Wendemuth, A.: Discourse particles and user characteristics in naturalistic human-computer interaction. In: Kurosu, M. (ed.) Human-Computer Interaction. Advanced Interaction Modalities and Techniques. Lecture Notes on Computer Science, vol. 8511, pp. 492–501. Springer, Berlin (2014)
43. Webb, N., Benyon, D., Bradley, J., Hansen, P., Mival, O.: Wizard of Oz experiments for a companion dialogue system: eliciting companionable conversation. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Paris (2010)
44. Wolters, M., Georgila, K., Moore, J., MacPherson, S.: Being old doesn't mean acting old: how older users interact with spoken dialog systems. ACM Trans. Access. Comput. **2**(1), 2:1–2:39 (2009)