Annotation and analysis of the LAST MINUTE corpus

Dietmar Rösner, Rico Andrich, Thomas Bauer, Rafael Friesen, Stephan Günther
Otto-von-Guericke Universität
Institut für Wissens- und Sprachverarbeitung (IWS)
Postfach 4120, D-39016 Magdeburg
roesner@ovgu.de

Abstract

This paper presents and discusses the techniques used for annotation and analysis of the LAST MINUTE corpus. This corpus comprises multimodal recordings (audio, video, bio-psychological data, verbatim transcripts) of Wizard of Oz simulated naturalistic human companion interactions in German.

1 Introduction

How do 'naive' users spontaneously interact with a system that – like companion systems (Wilks, 2010) – allows them to converse in spoken natural language? Can distinct user groups be detected based on observed linguistic behaviour? How do observed linguistic markers correlate with socio-demographic or psychometric data of the users?

These are issues that are highly relevant for the design of companion-like systems that shall flexibly adapt to their users based on personal characteristics and preferences as well as on the current situation. For the investigation of such questions, corpora with recordings of naturalistic interactions between users and (typically Wizard of Oz (WoZ)) simulated systems are indispensable assets (Legát et al., 2008; Webb et al., 2010).

Analysis of corpora of transcribed naturalistic interactions demands for different types of processing: shallow techniques with broad coverage as well as fine-grained analyses of dedicated passages in the textual recordings. Often these approaches are employed in sequence: first, quantitative analyses based on shallow processing (e.g. detection and counting of structures captured by regular expressions) result in statistical distributions for feature values of interest. Then, for a follow up in-depth analysis of the resp. extreme cases or outliers, qualitative approaches need to be employed.

In this paper we report on the role of shallow and 'deep' techniques – in the sense just presented – in evaluating one such corpus for German: the LAST MINUTE corpus (LMC). This corpus is derived from a large scale Wizard of Oz (WoZ) experiment where users had to solve a mundane task with the need for planning, replanning and strategy change (Frommer et al., 2012b; Rösner et al., 2014).

The paper is organised as follows: In section 2 the LAST MINUTE corpus is shortly presented. This is followed by a discussion of the methods used to analyse the transcripts from this corpus (section 3). Section 4 presents and discusses results from the analysis of discourse (4.1), behavior (4.2) and wizard errors and inconsistencies (4.3). In the summary (section 5) we discuss consequences for the design of future companion systems.

2 LAST MINUTE corpus

The experiment that underlies the multimodal recordings in the LAST MINUTE Corpus was designed in such a way, that the dialogs between simulated system and users were on the one hand restricted enough but on the other hand still offered enough opportunities for individual variation (Frommer et al., 2012b; Rösner et al., 2012). The domain chosen – packing a suitcase for a holiday trip of fourteen days – was mundane enough not to require any specialist knowledge as a prerequisite on the side of the subjects. As a key aspect, an inherent need for re-planning (need for unpacking after reaching a weight limit) and for strategy change (from summer to winter items after the delayed weather information about the target location) was built into the WoZ szenario.

The LMC is a valuable resource based on a large number of highly formalised, yet still variable experiments with subjects balanced with respect to gender and age group. In addition to work based on the verbatim transcripts the LMC has as well been employed for research based on other modalities, i.e. in audio analysis (e.g. (Prylipko et al., 2014)), in video analysis and in fusion of analysis

Proceedings of the Int. Conference of the German Society for Computational Linguistics and Language Technology,
pages 112–121, University of Duisburg-Essen, Germany, Sep 30–Oct 2 2015.

results from different modalities (e.g. (Frommer et al., 2012a)).

As a resource the LMC is 'middle ground' between on the one hand data (or a corpus) from a small scale experiment with narrow interactions and a single hypothesis only and on the other hand a corpus based on recordings from virtually unrestricted real life interactions (like e.g. Vera am Mittag (Grimm et al., 2008) with recordings from a German TV talk show). A more detailled presentation of the LAST MINUTE Corpus is given in (Rösner et al., 2014).

3 Methods

3.1 Discourse Annotation

The LAST MINUTE corpus comprises transcripts of all N = 133 experiments performed. On average each experiment took approximately 30 minutes real time, summing up to more than 56 hours of recorded interactions. In order to be able to quantitatively compare and contrast different dialog courses an adequate representation is needed.

The transcripts in the LMC are annotated with labels for the series of subsequent dialog acts of user and system, the so called dialog act representation (DAR, (Rösner et al., 2014)). This level of representation is independent of the domain of discourse, i.e. it is by no means restricted to the very task presented in LAST MINUTE but is applicable to all types of task-oriented user companion dialogs.

An example The DAR example in Table 4 (cf. appendix) is taken from a dialog segment where a subject (S; here: 20110401adh) tries to pack a (winter) coat but the requests for packing (RP) are rejected (RjP) by the wizard (W) several times (SRP WRjP pairs) and therefore the subject has to request the unpacking of several other items (SRU WAU pairs) in order to create sufficient space. Please note the emotional expression of relief (*'gott sei dank'*, engl. *'thank god'*) when the subject finally succeeds.

3.2 Dialog success measures (DSMs)

A LAST MINUTE experiment is made up of two major phases: a personalisation phase followed by the problem solving phase. In personalisation the system prompts the subject for personal data and stipulates narratives, for example about prior experiences with technical items. In the problem solving phase users have the option to express their requests for the various available actions (packing, change of selection category, unpacking, listing of suitcase contents, ...). User requests may be either accepted and confirmed by the system or they may be rejected.

This allows to evaluate the dialog course both locally and globally. Locally accepted requests are evaluated positively and rejections resp. get a negative score.

The relation between subject requests and their acceptance or rejection allows to define measures for the global dialog success in the problem solving phase of LAST MINUTE (Rösner et al., 2014):

- ratio between the accepted subject requests and the total number of subject requests (termed **DSM1**)
- 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: $0 \leq DSM2 \leq DSM1 \leq 1$.

3.3 Discourse analysis

Both dialog success measures are employed in the following analyses. They allow the following types of investigations:

- How do user groups based on socio-demographics differ with respect to global dialog success (cf. 4.1)?
- How do user groups that are defined based on distinct behavior during the experiment differ with respect to global dialog success (cf. 4.2)?

The methods employed in discourse analysis of the LMC are as follows: The LMC comprises full transcripts of N = 133 experiments. The transcripts are available as an XML-based data structure in the FOLKER format (Schmidt and Schütte, 2010). This highly structured format contains not only the transcription of all user and wizard contributions of every experiment in their relative temporal order, but also additional annotations. These annotations range from recorded nonphonological events (e.g. sighing, coughing, ...) to discourse level events (e.g. dialog act labels).

Starting from the FOLKER encoded transcripts we determine – typically with shallow techniques,

often based on regular expressions – features (or markers) either for complete transcripts or for their subparts (e.g. personalisation vs. problem solving or their resp. subphases). Such features are calculated on every level 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 analysis in which we combine quantitative and qualitative approaches from corpus linguistics (Gries, 2009). The quantitative methods start with compiling the empirical distributions of the feature values. These are visualised appropriately and e.g. tested for normality vs. skewness. Transcripts of (extreme) outliers are then additionally checked qualitatively – typically by human interpretation – in order to detect and discuss possible causes for deviation.

A recurring finding for virtually every investigated feature is that the distribution of feature values shows a large variance. This even holds for features that quantify aspects of the overall extent of the highly standardised experiments (cf. table 1).

Analysing the cause of the observed variance is a major issue in the work reported here. The different user groups based on socio-demographic features – i.e. age group (young vs. elderly subjects) and the four combinations of age group with gender – are potentially a primary source of the observed variance. Indeed: for many features the differences between the age groups and for gender-conditioned subgroups prove to be significant (cf. section 4.1).

When significant differences in the distribution of feature values have been found between socio-demographic groups then the additional question arises if these differences correlate significantly with differences in dialog success (as measured with DSM1 and DSM2).

3.4 Behavioral Analysis

In behavioral analyses errors that users make and problems they face 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 section 4.2, early errors in the personalisation phase could be identified that bear 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.

4 Results

An early result about the socio-demographic subgroups in the cohort of the LAST MINUTE experiment was that the subgroup of young women has significantly higher values of dialog success measures than the other three subgroups (i.e. young men, elderly women, elderly men, cf. (Rösner et al., 2014)). What other significant differences between these subgroups can be detected? For lack of space, we will concentrate on differences between socio-demographic subgroups with respect to verbosity and with respect to usage of politeness particles.

4.1 Discourse analysis: Age and Gender matters

Differences in verbosity: We employ the ratio of Tokens per Turn (TpT) as a verbosity measure for the user contributions in the LMC dialogs. Given the distinct nature of the different phases in the LAST MINUTE experiment, the measure varies between the more narrative oriented phases in personalisation and the phases of problem solving with a preference for usually shorter commands.

Major results for problem solving include (cf. fig. 2, table 2): age group matters, young subjects are significantly less verbose than elderly (Wilcoxon: W = 1722, p = 0.03251), whereas 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).[1]

Similar results hold for TpT values for other parts of the experiment. A point in case is for example the narratives phase in personalisation (cf. table 2).

Politeness particles as indicators for CASA: When humans conversing with a computer system do employ politeness particles this can be seen as indicator for (mindlessly) treating Computers as Social Actors (CASA, (Nass and Moon, 2000))

[1] Unless noted otherwise all statistical tests and calculations have been performed with the R language (R Development

Table 1: Examples of empirical distributions of features based on complete transcripts (N = 133)

marker	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD	total
tokens	266.0	444.0	545.0	602.7	699.0	1601.0	247.34	80160
turns	62.00	81.00	86.00	86.08	91.00	111.00	9.95	11448
TpT	2.804	5.143	6.282	7.060	8.109	19.290	2.95	n.a.

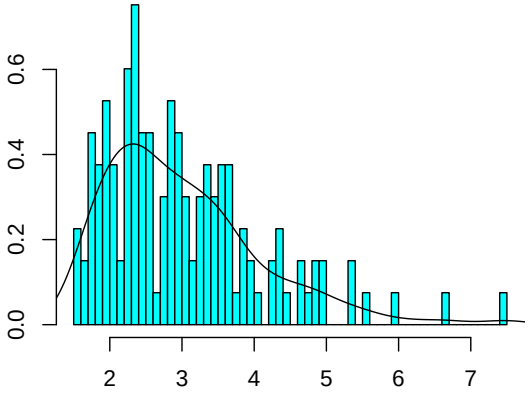

Figure 1: Distribution of Tokens per Turn (TpT) ratios for problem solving ($N = 133$)

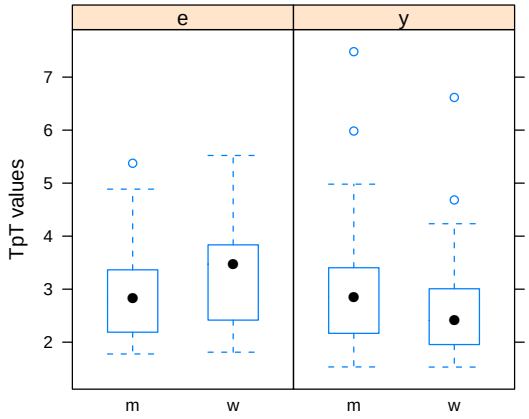

Figure 2: Tokens per turn ratios for problem solving conditioned by age group (e = elderly, y = young) and gender (m = men, w = women) ($N = 133$)

Table 2: Test results (Wilcoxon tests) for differences in mean verbosity between socio-demographic groups (e = elderly, y = young; m = men, w = women)

marker	g1	rel	g2	p-value
TpT probl. solv.	y	<	e	0.02016
TpT probl. solv.	m	<	w	n.s.
TpT pers. narratives	y	<	e	0.00423
TpT pers. narratives	w	<	m	n.s.

Table 3: Age group and gender differences in usage of politeness particles

nr pol. parts	em	ew	ym	yw	$\sum$
0 (below median)	7	5	**19**	**24**	55
1 (at median)	3	2	6	4	15
$>$ 1 (above median)	**19**	**25**	11	8	63

Counting the number of occurrences of politeness particles *'bitte'* (engl. *'please'*) and *'danke'* (engl. *'thank you'*) in user utterances per transcript provides distributions for all N = 133 subjects as visualised in figs. 3 and 4. Please note: 55 subjects never uttered one of these politeness particles. The median lies at one occurrence.

Again age matters: The subgroup with counts of used politeness particles above the median is clearly dominated by elderly subjects, whereas the subgroups below and at the median are dominated by young subjects (cf. table 3).

Kruskal tests show significant results for the two age groups (Kruskal-Wallis chi-squared = 24.6171, df = 1, p-value = 6.993e-07) and the four pairings of gender and age group (chi-squared = 26.0632, df = 3, p-value = 9.251e-06), but for gender we get insignificant differences only.

4.2 Behavioral analyses

In the following paragraphs subgroups of subjects with distinct problems are investigated.

Early problems with 'tell and spell': At the very beginning of the personalisation phase every subject is prompted:

```
Bitte nennen und buchstabieren Sie zunächst Ihren
Vor- und Zunamen!
```

Core Team, 2010) (Baayen, 2008)

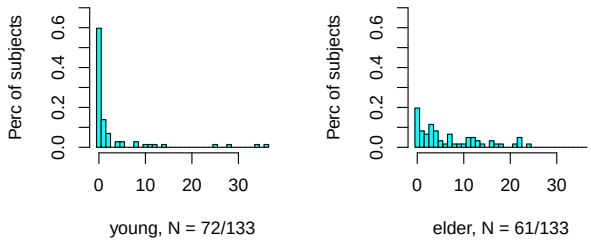

Figure 3: Distributions of number of occurrences of politeness particles in user utterances per transcript, subgroups young vs. elderly (Please note the different relative amount of zero occurrences)

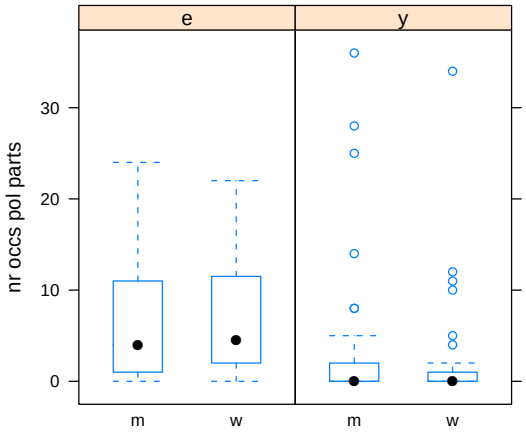

Figure 4: Number of occurrences of politeness particles in user utterances per transcript, conditioned by age group and gender (Please note the zero medians for the young subgroups and the resp. outliers)

```
[Please tell and spell your first name and surname!]
```

Some subjects need several trials, some even completely fail to provide the requested information. An example excerpt: subject 20110131bcl (anonymized) with three unsuccessful trials

```
{00:18} W guten tag und herzlich willkommen (.) ...
          [Hello and welcome. ...]
          bitte nennen und buchstabieren sie
          zunächst ihren vor und zunamen
          [please tell and spell your first name
          and surname]
{00:45} (1.89)
{00:47} P charlotte kurz
{00:48} (5.88)
{00:54} W bitte nennen und buchstabieren sie
          zunächst ihren vor und zunamen
{00:58} (---)
{00:59} P charlotte (.) kurz
{01:00} (8.58)
{01:09} W bitte nennen und buchstabieren sie
          zunächst ihren vor und zunamen
{01:13} (1.08)
{01:14} P charlotte kurz (3.8) ((schnalzt)) (.)
          mein vorname ist charlotte °h (-)
          mein familienname ist kurz
          [charlotte kurz (3.8) ((smacks)) (.)
          my first name is charlotte °h (-)
          my family name is kurz]
```

Please note: the wizards issued no more than three prompts, even when the third trial was still faulty. (Obviously, in such cases a runtime companion system should give a more adequate system response and not simply repeat the partially fulfilled prompt).

From N = 133 subjects the answer to the prompt 'Please tell and spell . . . ' is accepted after the first response 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, 2 more leave the first name out. (Please note: wizards did not react to these latter types of incomplete answers). In sum: from $N = 133$ subjects the answer to the prompt 'Please tell and spell . . . ' is wrong or incomplete in at least 34 cases (i.e. 25.6%), full task completion is reached only in 74.4% of the cases.

The age groups differ with respect to task completion: exactly 2 spelling request are needed by 5 elderly and 7 young subjects, whereas the 8 subjects with exactly 3 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' (Dalton and Fraenkel, 2012) or to effects of cognitive aging (Wolters et al., 2009) in general.

This leeds 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 do 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 means for DSM2 (no problem: 0.7075; problem: 0.6612) is significant as a Wilcoxon test reveals (W = 773, p-value = 0.02482; 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-value = 0.02382).

In sum: problems with the very first task in personalisation are an early predictor for later problems in the problem solving dialog of LAST MINUTE.

Early predictor: problems in 'data acquisition'. In the personalisation phase initiative lies primarily with the system. Here a typical adjacency pair (Jurafsky and Martin, 2008) is made up of a wizard prompt or question followed by a user narrative or answer.

Already the third wizard prompt demands for quite a number of personal data thus challenging not only the subjects's hearing understanding and comprehension abilities but as well her/his short term memory capacity.

```
Damit sich das Computerprogramm individuell an Sie
anpassen kann, sind einige konkrete Informationen
zu Ihrer Person erforderlich. Können Sie bitte zu
folgenden Punkten Angaben machen: Ihr Name, Ihr
Alter, Ihr Wohnort, Ihr Beruf, Ihr Arbeitsort, Ihre
Familie, Ihre Körpergröße, Ihre Konfektionsgröße,
Ihre Schuhgröße?

[in order to adapt to you the computer programme
needs some specific pieces of information. Could you
please give the following details: your name, your
age, your place of residence, your profession, your
place of work, your family, your body height, your
clothing size, your shoe size?]
```

If subjects do not give all of the requested data they are reprompted for missing data with questions of the type *'bitte ergänzen sie angaben zu . . . '* (engl. : *'please complete information about . . . '*).

Thus, in cases of a normal dialog course in personalisation the sources of variation are reprompts (e.g. 'tell and spell'), the number of questions of the type 'please complete information about . . . ' and the number of prompts for 'more detail'. Sources of variation in unforeseen courses are user questions, e.g. caused by hearing and/or understanding problems like in the following excerpt (subject 20110131bcl):

```
{03:09} W bitte ergänzen sie angaben zu ihrer körpergröße
         [please complete information about your body height]
{03:13} P ihrer was ihrer welcher größe
         [your what your which height]
{03:18} W bitte ergänzen sie angaben zu ihrer körpergröße
         [please complete information about your body height]
{03:22} P das weiß ich nicht (--) welche größe denn das weiß
          verstehe ich jetzt nicht so richtig
         [I do not know (--) which height
         I do not really know understand this]
```

Please note: a higher number of adjacency pairs in personalisation thus in general indicates problems on the subject's side. The empirical distributions of the total number of user turns in data acquisition, conditioned by age group and gender, are visualised in fig. 6.

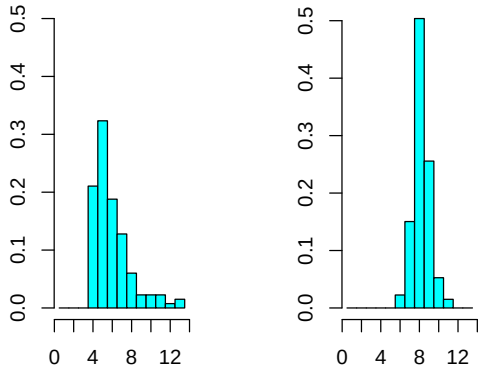

Figure 5: Total number of user turns in subphases of personalisation per transcript ($N = 133$): data acquisition (left), narratives (right)

In the following we perform a median split with respect to the total number of turns (i.e. adjacency pairs) in the subphase 'data acquisition'. 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-value = 0.04035; DSM2: W = 1604.5, p-value = 0.00718).

Issues of control: 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,
- wizard induced category change (WICC): the wizard enforces a category change.

More than half of the subjects are 'in control' in this sense. They have either zero or only one or at most two wizard induced category changes (from a total of 14 in a complete experiment). The complement of this group ('poor control') has between three and up to 10 WICCs.

Poor control of category changes (i.e. `WICCs > 2`) predicts poor global dialog success. The two subgroups – at and below the

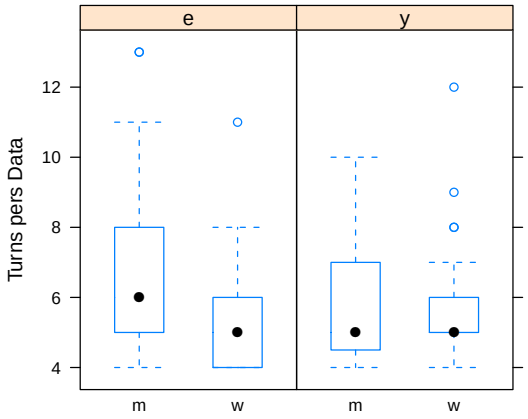

Figure 6: Total number of user turns in data acquisition per transcript, conditioned by age group and gender ($N = 133$)

WICC median of 2 or above the WICC median, resp. – show significant differences in both global dialog success measures (DSM1: Wilcoxon test, p-value = 0.02614; DSM2: Wilcoxon test, p-value = 3.610e-6).

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

Long pauses as indicators of helplessness: There is a subgroup of the subjects with poor control that – after some choices in a category – passively wait without any further action, sometimes for 40 seconds or more, until the system finally enforces a category change (WICC).

Not surprisingly the occurrence of such a type of long pause is again a predictor for global dialog failure. The subgroup of ten subjects that have at least one occurrence of a pause longer than ten seconds before a WICC has significantly poorer dialog success measures when compared to the complementary group of 123 subjects without such pauses (Wilcoxon tests, DSM1: W = 268, p-value = 0.0031, DSM2: W = 138.5, p-value = 4.87e-05).

4.3 Wizard problems: errors and inconsistencies

The LAST MINUTE experiment is a carefully designed and highly standardised experiment, based on a detailled manual (Frommer et al., 2012b), performed by intensively trained personnel (wizards) with elaborated computer support. In spite of intensive training and the detailled manual the wizards did not always operate consistently and accurately. This is not surprising given the large number of subjects and the time span of nearly a year for the completion of all N = 133 experiments. Fine-grained investigation of wizard behavior is necessary for quality assurance: it helps to avoid erroneously attributing a problematic dialog course to the subject when actually the wizard caused the problem.

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.

We also found wizard errors, characterised as 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 cannot be processed' (WRjNp) when indeed the intention of the subject was clearly recognizable and the intended action was performable.

An example (subject 20110329aus):

```
{15:04} W ... bevor weitere artikel ausgewählt wer-
          den können (.) müssen sie für genügend
          platz im koffer sorgen (.) hierfür können
          bereits eingepackte artikel wieder ausge-
          packt werden (.) auf nach frage erhalten
          sie eine aufzählung der bereits ausge-
          wählten artikel [... before more items
          can be chosen (.) you have to create
          enough space in the suitcase (.) for this
          purpose already packed items can be un-
          packed (.) upon request you can get a
          listing of the already chosen items]
{15:27} (2.81)
{15:30} P ja bitte [yes please]
{15:31} (3.85)
{15:35} W ihre aussage kann nicht verarbeitet werden
          [your statement cannot be processed]
```

In sum: a manual like (Frommer et al., 2012b) is *necessary*, but by *no means sufficient* for successful experiments. A manual defines the overall structure of experiments, but for non-trivial interactions nearly necessarily many questions will arise.

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 inconsistencies is the acceptance or rejection of synonyms. In some cases wizards accepted synonyms of packed items, in other cases they did not. In the following excerpt from 20101220bmh the use of a synonym is accepted

```
{18:51} P vier paar socken auspacken hh°
          [unpack four pairs of socks]
{18:58} P vier paar strümpfe wurden entfernt (.)
          [four pairs of socks were removed]
          sie können fortfahren [please proceed]
{19:02} P ein badeanzug [a bathing suit]
```

In contrast, the same synonym is rejected for subject 20100901amb:

```
{15:23} P °h entferne (--) drei socken
          [remove three socks]
{15:30} W ihre aussage kann nicht verarbeitet werden
          [your statement cannot be processed]
{15:33} P drei socken [three socks]
{15:38} W der gewünschte artikel ist nicht im koffer
          enthalten [requested item is not contained
          in suitcase]
{15:42} P na ick hab vierzehn socken reinjepackt
          (werden ja wohl) drei drin sein (.)
          ((schmatzt)) °hhh (4.04) welche artikel
          sind im koffer enthalten
          [well I have packed fourteen socks
          then three should very well be there
          (.) ((smacks)) °hhh (4.04) which
          items are contained in the suitcase]
```

Please note: refused unpacking requests of this type are very irritating for subjects. This is underlined by the protesting reaction of the subject.

5 Discussion and Outlook

We have presented examples of analyses of transcripts in the LAST MINUTE corpus of naturalistic human companion interactions and we have illustrated the interplay of shallow, quantitative and broad coverage approaches with qualitative human interpretations. In the following we will summarise major insights from these analyses and discuss their consequences for the design of future companion systems.

5.1 Major insights from analyses

Major insights from the analyses can be summarised as follows:

User groups based on socio-demographics matter. 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 of the sources of communication problems seem to be difficulties in 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 for global failure of the (independent) later problem solving dialog (cf. 4.2).

A strong indicator for a potential user problem is an overly long pause when the user actually has the turn, i.e. the right to give the next utterance (cf. 4.2).

The analysis of wizard errors and inconsistencies and the analysis of resp. user reactions (cf. 4.3) clearly demonstrates the dominance of semantic and pragmatic expectations of subjects in user companion interaction. Users are obviously puzzled when the system tries to enforce lexical or syntactic constraints that are in conflict with the user's expectations.

5.2 Consequences for the design of companion systems

The findings from the analyses of the dialogs in the LAST MINUTE corpus have consequences for the design of companion systems that are based on speech interaction with their users.

On the one hand differences between socio-demographic 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 (cf. table 1 or (Wolters et al., 2009)), demands for personalised calibration of dialog management strategies. Tests that are easy to perform and evaluate and that have strong predictive power for potential problems in the subsequent global dialog course – cf. section 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 section 4.2 such overly long pauses are strong indicators for problems and helplessness on the user's side and demand for an adequate response by the system.

Acknowledgments

The presented study is performed in the framework of the Transregional Collaborative Research Centre SFB/TRR 62 "A Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG). The responsibility for the content of this paper lies with the authors.

References

R.H. Baayen. 2008. *Analyzing Linguistic Data – A Practical Introduction to Statistics using R*. Cambridge University Press.

P. Dalton and N. Fraenkel. 2012. Gorillas we have missed: Sustained inattentional deafness for dynamic events. *Cognition*, 124(3):367–372.

J. Frommer, B Michaelis, D. Rösner, A. Wendemuth, R. Friesen, M. Haase, M. Kunze, R. Andrich, J. Lange, A. Panning, and I. Siegert. 2012a. Towards emotion and affect detection in the multimodal LAST MINUTE corpus. In *LREC 2012 Conf. Abstracts*, pages 3064–3069.

J. Frommer, D. Rösner, M. Haase, J. Lange, R. Friesen, and M. Otto. 2012b. *Teilprojekt A3 – Früherkennung und Verhinderung negativer Dialogverläufe – Operatormanual für das Wizard of Oz-Experiment; Arbeitspapier des Sonderforschungsbereichs - Transregio 62 'Eine Companion-Technologie für Kognitive Technische Systeme' = Project A3 - Detection and avoidance of failures in dialogues*. Pabst Science Publishers, Lengerich.

S. T. Gries. 2009. *Quantitative Corpus Linguistics with R: A Practical Introduction*. Routledge.

M. Grimm, K. Kroschel, and S. Narayanan. 2008. The Vera am Mittag German audio-visual emotional speech database. In *Multimedia and Expo, 2008 IEEE Int. Conf. on*, pages 865–868.

D. Jurafsky and J. H. Martin. 2008. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall, 2nd edition.

M. Legát, M. Grůber, and P. Ircing. 2008. Wizard of oz data collection for the czech senior companion dialogue system. In *Fourth Int. Workshop on Human-Computer Conversation*, pages 1 – 4, University of Sheffield.

C. Nass and Y. Moon. 2000. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, 56(1):81–103.

D. Prylipko, D. Rösner, I. Siegert, S. Günther, R. Friesen, M. Haase, B. Vlasenko, and A. Wendemuth. 2014. Analysis of significant dialog events in realistic human-computer interaction. *Journal on Multimodal User Interfaces*, 8(1):75–86.

R Development Core Team, 2010. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

D. Rösner, R. Friesen, S. Günther, and R. Andrich. 2014. Modeling and Evaluating Dialog Success in the LAST MINUTE Corpus. In *Proc. of LREC'14*, Reykjavik, Iceland. ELRA.

D. Rösner, J. Frommer, R. Friesen, M. Haase, J. Lange, and M. Otto. 2012. LAST MINUTE: a Multimodal Corpus of Speech-based User-Companion Interactions. In *Proc. of the 8th LREC*, pages 2559–2566, Istanbul, Turkey.

T. Schmidt and W. Schütte. 2010. Folker: An annotation tool for efficient transcription of natural, multi-party interaction. In N. Calzolari (Conference Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

M. Selting, P. Auer, D. Barth-Weingarten, J. R Bergmann, P. Bergmann, K. Birkner, E. Couper-Kuhlen, A. Deppermann, P. Gilles, S. Günthner, et al. 2009. Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung-Online-Zeitschrift zur verbalen Interaktion*, 10.

Nick Webb, David Benyon, Jay Bradley, Preben Hansen, and Oli Mival. 2010. Wizard of Oz Experiments for a Companion Dialogue System: Eliciting Companionable Conversation. In *Proc. of LREC'10*. ELRA.

Y. Wilks. 2010. *Close engagements with artificial companions: key social, psychological, ethical and design issues*, volume 8. John Benjamins Publishing.

M. Wolters, K. Georgila, J.D. Moore, and S.E. MacPherson. 2009. 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.

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 maximal 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 maximal 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 maximal 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
	...	...

Table 4: Excerpt from transcript with DAR labels: S indicates a subject and W a wizard contribution. Dialog acts include requests (R), rejections (Rj), accepts (A) for actions like packing (P) or unpacking (U). SNp stands for nonphonological utterances, SOT and SOQ for offtalk and questions resp. According to the GAT-2 minimal standard (Selting et al., 2009) short pauses are noted as (.) and (-), longer pauses with their duration (in seconds) in brackets, e.g. (1.77).