Linguistic analyses of the LAST MINUTE corpus

Dietmar Rösner, Manuela Kunze, Mirko Otto
Otto-von-Guericke Universität,
Institut für Wissens-
und Sprachverarbeitung
Fakultät für Informatik
Postfach 4120, D-39016 Magdeburg
roesner@ovgu.de

Jörg Frommer
Otto-von-Guericke-Universität,
Universitätsklinik für
Psychosomatische Medizin
und Psychotherapie
Leipziger Straße 44,D-39120 Magdeburg
joerg.frommer@med.ovgu.de

Abstract

The LAST MINUTE corpus comprises multimodal records from a Wizard of Oz (WoZ) experiment with naturalistic dialogs between users and a simulated companion system. We report about analysing the transcripts of the user companion dialogs and about insights gained so far from this ongoing empirical research.

1 Introduction

"Really natural language processing" (Cowie and Schröder, 2005), i.e. the possibility that human users speak to machines just as they would speak to another person, is a prerequisite for many future applications and devices. It is especially essential for so called companion systems (Wilks, 2010).

Corpora with naturalistic data from either human to human (e.g. (Oertel et al., 2012)) or human-machine interactions (e.g. (Legát et al., 2008), (Webb et al., 2010)) are an essential resource for research in this area. In the following we report about ongoing work in the linguistic analysis of the transcripts from the LAST MINUTE corpus. This corpus comprises multimodal records from a Wizard of Oz (WoZ) experiment with naturalistic dialogs between users and a simulated companion system.

The paper is organized as follows: In section 2 we give a short overview of the WoZ experiments. This is followed by a description of the LAST MINUTE corpus in section 3. In section 4 we report about analyses and empirical investigations with the transcripts. We sum up with a discussion of ongoing and future work.

2 The WoZ experiments

2.1 Design issues

Our WoZ-scenario is designed in such a way that many aspects of user companion interaction (UCI) that are relevant in mundane situations of planning, re-planning and strategy change (e.g. conflicting goals, time pressure, ...) will be experienced by the subjects (Rösner et al., 2011) .

The overall structure of an experiment is divided into a personalisation module, followed by the 'LAST MINUTE' module. These modules serve quite different purposes and are further substructured in a different manner (for more details cf. (Rösner et al., 2012b)).

2.1.1 Personalisation module

Throughout the whole personalisation module the dominant mode of interaction is system initiative only, i.e. the system asks a question or gives a prompt. In other words this module is a series of dialog turns (or adjacency pairs (Jurafsky and Martin, 2008)) that are made up by a system question or prompt followed by the user's answer or reaction. In some sense this module thus resembles more an investigative questioning than a symmetric dialog.

2.1.2 The LAST MINUTE module

Selection In the bulk of 'LAST MINUTE' the subject is expected to pack a suitcase for a two week holiday trip by choosing items from an online catalogue with twelve different categories that are presented in a fixed order. In a simplified view we thus have an iterative structure made up from twelve repetitions of structurally similar subdialogs each for the selection from a single category. The options of each category are given as menu on the subject's screen.

Normal packing sub dialog In a normal packing subdialog we essentially have a series of adjacency pairs made up of a user request for a number of items (more precisely: a user request for a number of instances from an item type) from the current selection menu (e.g. 'ten t-shirts') followed by a confirmation of the system (e.g. 'ten t-shirts have been added').

An example excerpt from an unproblematic segment of a packing dialog (subject 20110404bcm) [1]:

```
{07:39}  058  P  zwei tops [two tops]
{07:40}  059  (2.46)
{07:43}  060  W  zwei tops wurden hinzugefügt (.)
[two tops have been selected]
sie können fortfahren [you can proceed]
{07:46}  061  (1.13)
{07:47}  062  P  drei tshirts [three tshirts]
{07:49}  063  (2.42)
{07:51}  064  W  drei tshirts wurden hinzugefügt
[three tshirts have been selected]
{07:53}  065  (3.63)
{07:57}  066  P  ich möchte zur nächsten rubrik
[i want to go to the next category]
{07:59}  067  (2.73)
{08:01}  068  W  sie können jetzt aus der rubrik jacken
[you may now choose from category]
und mäntel auswählen [jackets and coats]
{08:05}  069  (4.6)
{08:09}  070  P  eine sommerjacke [a summer jacket]
```

Barriers The normal course of a sequence of repetitive subdialogs is modified for all subjects at specific time points.

These modifications or barriers are:

- after the sixth category, the current contents of the suitcase are listed verbally (listing barrier),
- during the eighth category, the system for the first time refuses to pack selected items because the airline's weight limit for the suitcase is reached (weight limit barrier).
- at the end of the tenth category, the system informs the user that now more detailled information about the target location Waiuku is available (Waiuku barrier).

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

2.2 Challenges for the subjects

In their initial briefing the subjects have been informed that all interaction shall be based on speech only and that neither keyboard or mouse are therefore available to them. Since the briefing does not comprise any detailled information about the natural language processing and the problem solving capabilities or limitations of the system the subjects are more or less forced to actively explore these aspects during the course of interaction.

The challenge for the subjects is twofold: They have to find out how (i.e. with which actions) they can solve problems that they encounter during interaction and they have to find out what linguistic means are available for them to instruct the system to perform the necessary actions. In other words, in order to be successful they have to build up a model of the capacities and limitations of the system based on their experience from successful or unsuccessful interactions. The user's model of the system will of course strongly influence the behavior of the user and the subsequent course of the interaction.

The discussion in (Edlund et al., 2008) leads to the following rephrasing of this challenge: Which metaphor will the subjects use when interacting with the WoZ simulated system? Will they treat the system more like a tool, i.e. choose the *interface metaphor*, or will they prefer the *human metaphor*, i.e. accept the system as an interlocutor and behave more like they would in human-human dialogs?

One approach to these questions is in the qualitative evaluation of the post-hoc in-depth interviews that a subset of ca. half of our subjects underwent after the experiments. In this paper we follow a complimentary approach: The linguistic behavior of the subjects is analysed under the perspective what conclusions it licenses about user assumptions about the speech-based system they experience.

[1] All excerpts from transcript are given - unless otherwise noted - with the GAT 2 minimal coding (cf. below). English glosses added in brackets for convenience.

Table 1: Comparison between corpora with naturalistic human-computer interactions

	SAL	SEMAINE	LAST MINUTE
Participants	4	20	130
Groups	students	students	balanced in age, gender, education
Duration	4:11:00	6:30:41	ca. 57:30:00
Nr of Sensors	2	9	13
Max. Video Bandwidth	352x288; 25Hz	580x780; 50Hz	1388x1038; 25Hz
Audio Bandwidth	20kHz	48kHz	44kHz
Transcripts	yes	yes	yes (GAT 2 minimal)
Biopsychological data	n.a.	n.a.	yes (heart beat, respiration, skin reductance)
Questionnaires	n.a.	n.a.	sociodemographic, psychometric
In depth Interviews	n.a.	n.a.	yes (70 subjects)
Language	English	English	German

3 Data sources

3.1 LAST MINUTE corpus

The LAST MINUTE corpus comprises multimodal recordings from the WoZ experiments with N = 130 participants (audio, video, biopsychological data), the verbatim transcripts and as additional material data from psychological questionnaires and records and transcripts from interviews (for more details cf. (Rösner et al., 2012a)).

In table 1 we summarize various parameters (as reported in (McKeown et al., 2010)) of two widely employed corpora with recordings from naturalistic human-computer dialogs – SAL (Douglas-Cowie et al., 2008) and SEMAINE (McKeown et al., 2010) – and contrast them with the resp. values for the LAST MINUTE corpus.

Sample The LAST MINUTE corpus consists of data sets from 130 subjects. 70 of them are between 18 and 28 years old ('young'; M=23,2; Md=23,0; s=2,9; 35 of them are male, 35 are female) and 60 of them are 60 years old or older ('elderly'; M=68,1; Md=67,0; s=4,8; the oldest subject is 81 years old; 29 of them are male, 31 are female). Within the group of the young 44 subjects have a high school diploma, 26 have none. Within the group of the elderly 35 subjects have a high school diploma or a university degree and 25 subjects have no university degree.

3.2 Wizard logs

The wizards have been trained and their behaviour has been anticipated and prescribed as nonambiguous as possible in a manual (Frommer et al., 2012) .

All dialog contributions from the system (i.e. wizard) were pronounced by a text-to-speech system (TTS). The input for the TTS either was generated dynamically from the knowledge base (e.g. verbalisations of the current contents of the suitcase) or was chosen by the wizards from menus with prepared stock phrases. As a last option for unforeseen situations wizards could – supported by autocompletion – type in text to be uttered by the TTS. In the course of more than 130 experiments with on average approx. 90 dialog turns each (in sum a total of ca. 11800 turns) only in one single turn – during the very first experiments – the wizards had to resort to this last option.

After a WoZ session all wizard contributions together with their timings are available as additional log file.

Evaluation of the wizard log files already allows to classify the overall interaction of different subjects with respect to a number of aspects (cf. 4.2).

3.3 Transcripts

All experiments and interviews were transcribed by trained personnel following the GAT 2 minimal standard (Selting et al., 2009). This standard captures the spoken text, pauses, breathing and allows to include comments describing other nonlinguistic sounds.

In order to simplify the production of the GAT 2 transcripts, we started from the logged wizard statements which were converted into GAT2 transcripts of the wizard. These transcripts were used as template for full transcripts of the interaction, into which only transcriptions of the utterances of the subject had to be inserted.

All transcripts were made using FOLKER (Schmidt and Schütte, 2010). Some transcripts were created by more than one transcriber, the parts are connected using exmeralda (Schmidt and Schütte, 2010). Own software was employed to support the transcribers in detecting and correcting possible misspellings.

The transcripts try to be as close as possible to the actual pronunciation of the subjects. They therefore include as well nonstandard writings, e.g. for dialect (e.g. 'jejebn' instead of 'gegeben', engl. 'given'). In addition the utterances of the subjects exhibit phenomena that are typical for spontaneous spoken language, e.g. repairs, restarts, incongruencies.

3.3.1 Corpus size

Counting only the user contributions the total number of tokens in the corpus of transcripts sums up to 79611. The number of tokens per transcript ranges from 252 till 1730 with mean value 612,39 (variance 186,34). The total size may seem small when compared to the size of large available corpora, but one should keep in mind that in spite of their differences all 130 dialogs are focussed and thus allow for in depth comparisons and analyses.

3.3.2 Processing of transcripts

For linguistic processing of the Folker based transcripts we employ the UIMA framework.[2]

The first step is to transform Folker format into UIMA based annotations. After this, we initiate a number of linguistic and dialogue based analyses. For these analyses, we used internal and external tools and resources. For example, we integrated resources of GermaNet[3], LIWC (Wolf et al., 2008) and of the project Wortschatz Leipzig[4].

4 Linguistic analyses

Linguistic analyses of the LAST MINUTE transcripts are an essential prerequisite for an in depth investigation of the dialog and problem solving behavior of the subjects in the WoZ experiments. A long term goal is to correlate findings from these analyses with sociodemographic and psychometric data from the questionnaires.

4.1 Linguistic structures employed by subjects?

4.1.1 Motivation

First inspections of transcripts revealed that there are many variations in lexicalisation but only a small number of linguistic constructs that subjects employed *during the packing and unpacking subdialogues* of the LAST MINUTE experiments.

For issuing packing (or unpacking) commands these structural options comprise:

- full sentences with a variety of verbs or verb phrases and variations in constituent ordering,
- elliptical structures without verbs in a (properly inflected) form like `<number> <item(s)>`,
- 'telegrammatic structures' in a (technically sounding and mostly uninflected) form with an inverted order of head and modifier like `<item> <number>`.

As a first quantitative analysis of the 'LAST MINUTE' phase, the absolute and relative numbers for the usage of these constructs have been calculated from the full set of transcripts.

4.1.2 Results

Based on the analyses of the packing/unpacking phase of N = 130 transcripts we get the following figures:

We have a total of 8622 user utterances. If we perform POS tagging (with a slightly modified version of STTS[5]) and then count the varying POS tag patterns, we find 2041 different patterns for the 8622 utterances. The distribution is strongly skewed (cf. table 2): A small number of (regular) POS patterns captures a large fraction of utterances.

In classifying POS tag sequences we distinguish four categories: full sentences and sentence

[2] uima.apache.org/

[3] http://www.sfs.uni-tuebingen.de/lsd/

[4] wortschatz.uni-leipzig.de/

[5] www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html

Table 2: Most frequent POS patterns for elliptical structures

class	sem	POS pattern	nr of occs
E	P	ART NN	1020
E	P	CARD NN	657
E	P	NN	537
E	C	ADJ NN	355
E	C	ADJ,ADV	349
E	C	NN	148

like structures (S; with an obligatory verb), elliptical constructs without a verb (E), telegrammatic constructs (T, cf. above) and meaningful pauses, i.e. user utterances, that more or less consist of interjections only (DP).

In descending order of occurrences we have the following counts:

- 5069 user utterances or 58.79 % (realised with 223 patterns) are classified as E,
- 807 user utterances or 9.36 % (realised with 135 patterns) as S,
- 551 user utterances or 6.39 % (realised with 21 patterns) as T, and finally
- 178 user utterances or 2.06 % (realised with 8 patterns) as DP.

At the time of writing 2017 utterances realised in 1654 different patterns can not uniquely be classified. In many cases this is due to the typical phenomena of spontaneous spoken language, e.g. repairs, restarts and the use of interjections.

4.1.3 Discussion and remarks

The use of elliptical structures is a typical aspect of efficient communication in naturally occuring dialogs (Jurafsky and Martin, 2008). Thus the dominance of elliptical structures in the user contributions of the LAST MINUTE corpus can be seen as a clear indicator that most subjects have experienced the dialog with the system in a way that licensed their natural dialog behavior.

The empirical analysis of the structure of user utterances has fed as well into the implementation of an experimental system that allows to replace the wizard with an automated system based on the commercial speech recogniser Nuance.

4.2 Success and failure of dialog turns?

4.2.1 Motivation

Dialog turns are either successful or they may fail for a variety of reasons. We will first discuss failures during the LAST MINUTE phase. Failed turns can easily be detected in the so called wizard logs (cf. 3.2) because in the case of failure no confirmation is given by the system but some different utterance.

Success A dialog turn starting with a user request to pack or to unpack some items is successful when the situation allows to perform the requested action. In the wizard log file this can easily detected by the respective confirmative response of the system ('... wurd.* hinzugef.* ...', '... wurd.* entfernt ...').

Error messages The least specific 'error message' of the system tells the user that his utterance can not be processed ('ihre aussage kann nicht verarbeitet werden'). There are a number of reasons for using this 'catch all' system response. These include:

- The wizards conjure that the voice quality of the user's utterance is too poor for current automated speech recognition (ASR) technology.
- The content of the user's utterance is beyond the allowed scope of the current subdialog.
- The syntactic or semantic complexity of the user's utterance is judged to be beyond the limits of current NLP technology.

The following system reactions are more specific:

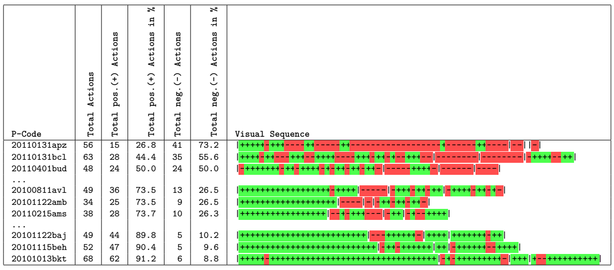
- When a user tries to unpack items that have not been packed into the suitcase he gets the response that these items are not contained in the suitcase ('... nicht im Koffer enthalten').
- When a user reaches the weight limit for the suitcase again then a packing command is responded to by the system with the message that the chosen item(s) can not be packed due to the weight limit (' ...k.*nn.* nicht hinzugef.* werden ... ').

- When the time for a category is over (local time limit) then the system tells this to the user and enforces a change of the category (' . . . muss jetzt beendet werden . . . ').

The excerpt in table 3 illustrates a problematic dialog situation: the simulated system does not accept a collective term ('tauchausrüstung', engl. 'diving equipment') employed by subject 20110224awh in an unpacking request.

A global measure In order to compare different dialogs we start with the following coarse global measure: We distinguish turns that are - based on the logged system response – judged as successful from those that are judged as unsuccessful or faulty. We then use the ratio of unsuccesful turns in relation to all turns as measure of the relative faultiness of the dialog as a whole.

Figure 1 visualises different dialog courses of subjects from the experiments (green: successful turn, red: unsuccessful turn).



P-Code	Total Actions	Total pos.(+) Actions	Total pos.(+) Actions in %	Total neg.(-) Actions	Total neg.(-) Actions in %	Visual Sequence
20110131apz	56	15	26.8	41	73.2	|+++++-+++----++-----++------------------+------++----|--| |-|
20110131bcl	63	28	44.4	35	55.6	|++++-++---+++--++++----+++-++-+--+++--|--------|--------|-++++--++|
20110401bud	48	24	50.0	24	50.0	|-+++++++-++-++++-+-++-++-++-|-----++++-|------|----|
...						
20100811avl	49	36	73.5	13	26.5	|++++++++++++++++++-++++|-----|-+++-++-++|-++++-++-+-|
20101122amb	34	25	73.5	9	26.5	|+++++++++++++++++++|----|-|-++-++-++-|
20110215ams	38	28	73.7	10	26.3	|+++++++++++++++++|--+-+++---|-++|-+--++++|
...						
20101122baj	49	44	89.8	5	10.2	|+++++++++++++++++++++++++|---++++++-|++++|+++++++-++|
20101115beh	52	47	90.4	5	9.6	|+++++++++++++++++++++++++++|-++-++++++|++|-++++++--++++|
20101013bkt	68	62	91.2	6	8.8	|+++++-+++++++++++++++++++++++++++++++++|-+-+++++++++-|+++|+--++++++++++|

Figure 1: Variations in dialog courses

4.2.2 Questions

Are there correlations of dialog success with sociodemographic variables and with personality traits measured with the psychometric questionnaires?

4.2.3 Results

For a cohort of N = 130 subjects the values for this global measure range between 9 % and 73 % with a mean of approximately 26 % and variance 10.

In fig. 2 the result of a contrastive analysis of the dialog courses of all N = 130 subjects, divided into the subcohorts of elderly vs. young subjects, is given. The chart illustrates that more than half of the elderly have significantly more negative dialog turns than the young subjects.

The differences are significant: a t-test yields a t-value of -3.779595 and a p-value of 0.000240.

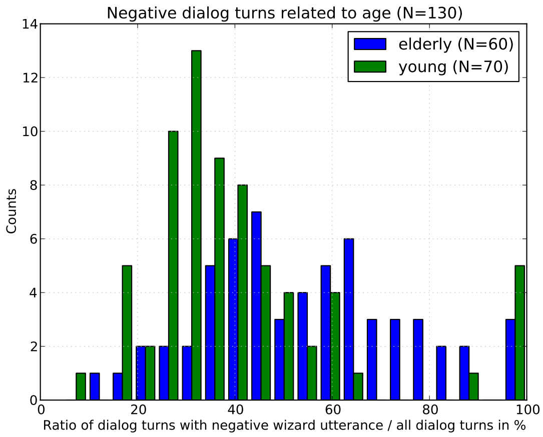


Figure 2: Contrastive evaluation of dialog courses: elderly vs. young subjects

4.3 Do users adapt to the system?

4.3.1 Motivation

The subjects in our WoZ experiments are only informed that they will interact in spoken language with a new type of system that shall be personalised and that therefore will ask some questions and pose some tasks. They do not get explicit instructions about the linguistic constructions and interaction patterns that are possible or not possible to use in the interaction.

How do people with differing technical background interact in such a situation?

One issue to explore is the question of rapport: Do users mirror the language of the system, e.g. on the lexical or syntax level? The system e.g. uses the general and somewhat uncolloquial term 'hinzufügen' (engl. to add) in its feedback for selection operations. Similarly the system always uses 'wurde entfernt' (engl. approx. 'was removed') when confirming unpacking requests. Do users mirror this usage?

4.3.2 Results

Within the total of N = 130 transcripts only in $N_1 = 25$ transcripts at least one occurrence of a form of the verb 'hinzufügen' (engl. 'to add')

Table 3: A problematic dialog situation

```
{23:05}  208  P  also (1.5) zahnpflegeset [dental care kit]
{23:09}  209  (4.0)
{23:13} 210 W ein zahnpflegeset kann nicht hinzugefügt werden (.) [a dental care kit cannot be added]
anderenfalls würde die von der fluggesellschaft vorgeschriebene [otherwise the weight limit prescribed by the airline]
maximale gewichtsgrenze des koffers überschritten werden (.) bevor [will be exceeded (.) before selecting]
weitere artikel ausgewählt werden können (.) müssen sie für genügend [any further items (.) you will have to make room for them]
platz im koffer sorgen (.) hierfür können bereits eingepackte artikel [in your suitcase (.) therefore already selected items]
wieder ausgepackt werden (.) auf nachfrage erhalten sie eine [can be unpacked (.) you can get a list of the selected items]
aufzählung der bereits ausgewählten artikel [upon request]
{23:37}  211  (3.4)
{23:41} 212 P ((fingertrommeln)) was nehm_wer_n da wieder raus (.) [((finger drumming)) well what shall we take out again]
dann nehm_wer die tauchausrüstung raus [let us take out the diving equipment]
{23:48}  213  (2.51)
{23:51} 214 W der gewünschte artikel ist nicht im koffer enthalten [the selected item is not contained in the suitcase]
{23:54}  215  (--)
{23:55} 216 P öh ((gedehnt)) (.) die taucherbrille schnorchel (-) ist [the diving goggles snorkel are ]
enthalten (1.9) ((fingertrommeln)) [contained (1.9) ((finger drumming))]
{24:02} 217 W ihre aussage kann nicht verarbeitet werden [your statement cannot be processed]
{24:05} 218 (2.69)
{24:08} 219 P so (.) °h °h h° (1.7) ein rasierset (-) brauche ich [so (.) i do need a beard trimmer]
{24:14} 220 (3.5) {24:18} 221 W der artikel rasierset kann nicht [item beard trimmer cannot be added]
hinzugefügt werden (.) anderenfalls würde die maximale gewichtsgrenze [otherwise the weight limit]
des koffers überschritten werden (--) ... [of the suitcase will be exceeded (--) ...]
```

could be found in user utterances of the packing/unpacking phase. For these $N_1 = 25$ transcripts we have a range from 1 to maximally 8 occurrences with mean: 2.36, std: 1.85 and median: 2.0.

Within $N_2 = 68$ transcripts at least one occurrence of a form of the verb 'entfernen' (engl. 'to remove') could be found in user utterances of the packing/unpacking phase. For these $N_2 = 68$ transcripts we have a range from 1 to maximally 13 occurrences with mean: 4.22, std: 3.34 and median: 3.0.

In the intersection of both groups, i.e. at least one occurrence each of a form of the verbs 'entfernen' and 'hinzufügen', we have $N_3 = 20$ transcripts. For these $N_3 = 20$ transcripts we have a range from 2 to maximally 19 combined occurrences with mean: 8.85, std: 4.64 and median: 8.0.

4.3.3 Discussion and remarks

That users mirror the lexical items of the system is thus rather the exception than the rule. Nevertheless it seems worth to be explored if - and if so how - the subgroup of subjects that do so differs from those subjects that do not.

4.4 Politeness in user utterances?

4.4.1 Motivation

In all its utterances the system uses the polite version, the German 'Sie' (polite, formal German version of 'you') when addressing the user. In requests the system employs the politeness particle 'bitte' (engl. 'please'). How polite are users in their utterances?

4.4.2 Results

Pronouns The following counts are all (unless otherwise noted) taken from the packung/unpacking phase of the transcripts: Within a total of N = 130 transcripts only in $N_1 = 21$ transcripts at least one occurrence of 'sie' als formal personal pronoun in addressing the system is used. Only within $N_2 = 4$ transcripts the informal 'du' (or one of its inflected forms) is used to adress the system (other uses of 'du' are within idiomatic versions of swear words like 'ach du lieber gott', engl. 'oh god'). Within $N_3 = 18$ transcripts subjects employ the plural personal pronoun 'wir' (engl. 'we'). Some occurrences of 'wir' in offtalk can be seen as more or less fixed phrasal usages (like 20101115beh 'ach das schaffen wir locker', engl. '... we will make this with ease' or 20110401adh 'wo waren wir', engl. 'where have we been'), but when used in commands (packing, unpacking, ...) then this pronoun can be given an inclusive collective reading as referring to both subject and system as a joint group. Please note: The pronoun 'wir' thus allows users to avoid to explicitly approach the system.

Example of this latter usage:

```
20110307bss:ja dann nehm wir eine jacke raus
[engl.: yeah then we take a jackett off]
20110315agw: dann streichen wir ein hemd
[engl.: then we cancel a shirt]
```

In sum: How users approach the system differs significantly. Most subjects avoid any personal pronouns when adressing the system, some employ the German 'Sie' (formal German version of 'you') and only very seldom the informal German 'du' is used.

Politeness particles From N = 130 subjects $N_1 = 67$ use one of the politeness particles 'bitte' or 'danke' at least once within the packing/unpacking phase. The maximum number of uses is 34, with a mean of 7.57, standard deviation of 7.89 and median of 4.0. If we neglect those subjects at and below the median as only occasional users of these particles we get $N_2 = 32$ subjects that use these particles much more frequent.

Intersecting the group of subjects with at least one occurrence of 'sie' (cf. above) with users of the politeness particles 'bitte' or 'danke' results in a subgroup of $N_3 = 19$ subjects. These have combined numbers of occurrences ranging from 2 till 32 with mean: 8,60, std: 8,38 and median: 4,00. In other words: most user of 'sie' are as well users of the politeness particles.

4.4.3 Discussion and remarks

Politeness is one of a number of indicators of the way how subjects experience the system.

The difference in using personal pronouns and politeness particles is another example that most users do not try to build rapport with the system on the level of lexical choices.

As with other subgroups of subjects (as e.g. detected in 4.3) the following questions have to be further investigated:

Are there differences in the overall dialog success or failure between 'normal' and 'polite' users?

Are there correlations between user politeness and sociodemographic data and personality traits measured with the psychometric questionnaires?

4.5 Conclusion

In the light of the metaphor discussion (cf. 2.2) we can summarize and re-interpret our results as follows: Subjects whose linguistic behavior gives a strong indication for the dominance of one of these metaphors are minorities within our sample. This holds for the minority group of those that prefer technically sounding 'telegrammatic structures' (cf. 4.1) and thus obviously prefer the interface metaphor. It holds as well - on the other extreme - for the group of those that heavily employ interpersonal signals such as formal pronouns and politeness particles thus indicating a human metaphor at work. Although further investigations are necessary, the majority of our subjects seems to work with 'a metaphor that lies between a human and machine - the android metaphor' (Edlund et al., 2008).

5 Future work

We report here about on going work. More issues have been or are still investigated with the LAST MINUTE corpus that can - due to limited space - only be mentioned here. Issues to be further explored include:

Effects of reinforcement learning We have found indications that subjects strongly tend to reuse linguistic constructs that have resulted in successful dialog turns (an effect that can be interpreted as a form of reinforcement learning).

Verbosity vs. sparseness of lingustic expression As already noted above, subjects strongly differ in their verbosity. This is of course more obvious in the narratives of the personalisation phase, but it is measurable even in the LAST MINUTE phase.

Detection and analysis of offtalk Linguistic analysis is essential for the detection of offtalk. Many questions arise: How often does offtalk occur? How can offtalk utterances be further classified (e.g. thinking aloud, expressing emotions, ...)? Is there a correlation between the degree and nature of offtalk usage and sociodemographic data and personality traits?

Emotional contents In the experiments reported here we have three sources of utterances with emotional contents: self reporting about past emotions in the personalisation phase for all subjects, self reporting about current emotions in the intervention phase for the randomly chosen subjects with an intervention and spontaneous expression of emotions (e.g. swear words, offtalk, self accusations, etc.) especially at the barriers or when problems occur during the interaction.

A detailled linguistic analysis of these various forms of emotional contents in the LAST MINUTE transcripts is on the agenda.

6 Summary

We have presented the current state of the linguistic analyses of the LAST MINUTE corpus. This corpus of recordings from naturalistic interactions between humans and a WoZ simulated companion system excels available corpora with respect to cohort size, volume and quality of data and comes with accompanying data from psychometric questionnaires and from post hoc in depth interviews with participants. The material is a cornerstone for work in the SFB TRR 62 but is as well available for research in affective computing in general.

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.

Availability

The LAST MINUTE corpus is available for research purposes upon written request from the authors.

References

Roddy Cowie and Marc Schröder. 2005. Piecing together the emotion jigsaw. *Machine Learning for Multimodal Interaction*, pages 305–317.

E. Douglas-Cowie, R. Cowie, C. Cox, N. Amier, and D. K. J. Heylen. 2008. The Sensitive Artificial Listener: an induction technique for generating emotionally coloured conversation. In L. Devillers, J-C. Martin, R. Cowie, E. Douglas-Cowie, and A. Batliner, editors, *LREC Workshop on Corpora for Research on Emotion and Affect, Marrakech, Marokko*, pages 1–4, Paris, France. ELRA.

Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech Communication*, 50(8-9):630 – 645.

Jörg Frommer, Dietmar Rösner, Matthias Haase, Julia Lange, Rafael Friesen, and Mirko Otto. 2012. *Verhinderung negativer Dialogverläufe – Operatormanual für das Wizard of Oz-Experiment*. Pabst Science Publishers.

Daniel Jurafsky and James H. Martin. 2008. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall, 2nd (May 26, 2008) edition.

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

G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic. 2010. The SEMAINE corpus of emotionally coloured character interactions. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 1079–1084, July.

Catharine Oertel, Fred Cummins, Jens Edlund, Petra Wagner, and Nick Campbell. 2012. D64: a corpus of richly recorded conversational interaction. *Journal of Multimodal User Interfaces*, in press.

Dietmar Rösner, Rafael Friesen, Mirko Otto, Julia Lange, Matthias Haase, and Jörg Frommer. 2011. Intentionality in interacting with companion systems – an empirical approach. In Julie Jacko, editor, *Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments*, volume 6763 of *Lecture Notes in Computer Science*, pages 593–602. Springer Berlin / Heidelberg.

Dietmar Rösner, Jörg Frommer, Rico Andrich, Rafael Friesen, Matthias Haase, Manuela Kunze, Julia Lange, and Mirko Otto. 2012a. LAST MINUTE: a novel corpus to support emotion, sentiment and social signal processing. In *LREC 2012 Workshop Abstracts*, pages 171–171.

Dietmar Rösner, Jörg Frommer, Rafael Friesen, Matthias Haase, Julia Lange, and Mirko Otto. 2012b. LAST MINUTE: a multimodal corpus of speech-based user-companion interactions. In *LREC 2012 Conference Abstracts*, page 96.

Thomas Schmidt and Wilfried Schütte. 2010. Folker: An annotation tool for efficient transcription of natural, multi-party interaction. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Margret Selting, Peter Auer, Dagmar Barth-Weingarten, Jörg Bergmann, Pia Bergmann, Karin Birkner, Elizabeth Couper-Kuhlen, Arnulf Deppermann, Peter Gilles, Susanne Günthner, Martin Hartung, Friederike Kern, Christine Mertzlufft, Christian Meyer, Miriam Morek, Frank Oberzaucher, Jörg Peters, Uta Quasthoff, Wilfried Schütte, Anja Stukenbrock, and Susanne Uhmann, 2009. *Gesprächsanalytisches Transkriptionssystem 2 (GAT 2)*. Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion, 10 edition.

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 *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA).

Y. Wilks. 2010. *Close Engagements with Artificial Companions: Key Social, Psychological and Design issues*. John Benjamins, Amsterdam.

Markus Wolf, Andrea B. Horn, Matthias R. Mehl, Severin Haug, James W. Pennebraker, and Hans Kordy. 2008. Computergestützte quantitative Textanalyse. In *Diagnostica*, volume Vol. 54, Number 2/2008. Hogrefe, Göttingen.