Modeling and evaluating dialog success in the LAST MINUTE corpus

Dietmar Rösner, Rafael Friesen, Stephan Günther, Rico Andrich

Otto-von-Guericke Universität, Institut für Wissens- und Sprachverarbeitung
Postfach 4120, D-39016 Magdeburg
{roesner, friesen, guenther, andrich}@ovgu.de

Abstract

The LAST MINUTE corpus comprises records and transcripts of naturalistic problem solving dialogs between N = 130 subjects and a companion system simulated in a Wizard of Oz experiment. Our goal is to detect dialog situations where subjects might break up the dialog with the system which might happen when the subject is unsuccessful. We present a dialog act based representation of the dialog courses in the problem solving phase of the experiment and propose and evaluate measures for dialog success or failure derived from this representation. This dialog act representation refines our previous coarse measure as it enables the correct classification of many dialog sequences that were ambiguous before. The dialog act representation is useful for the identification of different subject groups and the exploration of interesting dialog courses in the corpus. We find young females to be most successful in the challenging last part of the problem solving phase and young subjects to have the initiative in the dialog more often than the elderly.

Keywords: Human-Companion-Interaction, Dialog Success, Wizard-Of-Oz

1. Introduction

The LAST MINUTE corpus comprises records and transcripts of naturalistic problem solving dialogs between N = 130 subjects and a companion system simulated in a Wizard of Oz (WoZ) experiment (Rösner et al., 2012b).
The 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., 2012a).
With 56 hours of recorded and transcribed[1] interactions the LAST MINUTE corpus is an invaluable resource for collaborating groups that employ e.g. the recorded audio or video streams to train classifiers for their analyses. This includes emotion detection from facial expressions or from prosodic aspects of speech signals (Frommer et al., 2012a). Other WoZ experiments in the companion paradigm (Wilks, 2010) include (Legát et al., 2008) and (Webb et al., 2010).

2. LAST MINUTE dialogs

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)).
In the bulk of LAST MINUTE – the problem solving phase – 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.

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 detailed information about the target location Waiuku is available (Waiuku barrier).

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

3. Previous work

A first global measure In order to compare different dialogs we started with the following coarse global measure for the course of interaction of the LAST MINUTE problem solving dialogs: 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 unsuccessful turns in relation to all turns as a measure of the relative faultiness of the dialog as a whole. (Rösner et al., 2012a)
In detail:

- successful turns are those with an explicit wizard confirmation of success,
- the turns that are counted as failed or unsuccessful include those with the following system responses: unprocessable input, item not in suitcase, weight limit reached again, system enforced category change.
- system responses that are neutral or ambiguous are ignored.

[1] The audio recordings were transcribed using with FOLKER. (Schmidt and Schütte, 2010) The transcription followed the the GAT-2 minimal standard. (Selting et al., 2009)

For a cohort of N = 130 subjects the values for this global measure range between 9 % and 73 % (unsuccessful turns) with a mean of approximately 26% and a variance of 10%. The only information source for this measure were the log files of the system utterances, because the full transcription of all experiments was not available at that time. The log files allowed to assume which subject request precedes a system utterance but many situations remained ambiguous and all assumptions were unproven. Now, with the full transcripts and annotation of the packing phase of all subjects, better insight and finer analyses are possible.

Refined measure In this work we refine the segmentation from turns to dialog acts which allows finer analysis of the subjects' intentions in the dialog. We also refine the operationalization of successful dialog in two ways which help to answer the two questions:

- Did the system react as the subject expected it? (cf. section 5.1.)
- Did the user expect what the system did? (cf. section 5.2.)

The questions are quite similar but lead to different perspectives on the corpus and to supplementary results.

4. Annotation

We analyze the task-related success, so the annotation is also focused on the task. In the previous work and in following we only consider the problem solving phase of LAST MINUTE.

4.1. Dialog acts of the subject

The utterances of the subjects show a large variance with respect to many linguistic features: lexical choice, syntactic patterns, well- vs. ill-formedness, etc. (Rösner et al., 2012c) Fully automatic correct classification of all dialog acts is not possible up to now, so all dialog acts from the packing phase of the subjects were annotated by a human rater.
During the annotation an initially tiny tagset was adjusted to distinguish many domain specific dialog acts but to contain mostly frequent tags.
The annotation process was computer assisted: The annotations of ca. 1/2 of the segments which are easy to detect correctly (e.g. "two jeans", "next category") were retrieved automatically (rule-based) – the rater verified the automatic annotations and annotated the remaining segments. The annotator normally annotated the written transcript, but whenever the transcript was ambiguous or the annotator needed more information he was free to listen to the audio recordings.
During transcription very long subject contributions were split into smaller segments, so sometimes a single dialog act spreads over several segments. Such segments were marked as associated and were glued automatically for the analyses presented here. Some transcript segments contain more than one dialog act, so they could be annotated with several annotations (e.g. a segment with the utterance "ein pullover (.) weiter" - engl. "a pullover (.) go on" contains a *packing request* and a *request for category change* and was annotated with both – preserving the temporal order). Thus the segmentation of the transcribers does not interfere with the dialog act annotation.

4.2. Dialog acts of the wizard

Most of the Wizard contributions were preformulated and could be annotated automatically by using regular expressions.

4.3. Dialog act representation (DAR)

The first measure – based on unambiguously classifiable wizard utterances only – obviously is quite coarse. We now work with measures that are based on sequences of successive subject and wizard utterances.
The dialog act representation is hierarchical (x stands for 'unknown'): The first capital letter indicates the speaker (S for subject, W for wizard). The second capital letter stands for the dialog act:

Sx: R = request, O = offtalk, Np = non-phonological and pauses, A = answer

Wx: A = accept, Rj = reject, I = information, Q = question/request

A third capital letter may refine a dialog act subtype:

SRx: P = packing, U = unpacking, E = exchange, F = finalization, L = listing, C = category change

SOx: T = offtalk, Q = question

WAx: P = packing, U = unpacking, F = finalization

WRjx: P = packing, U = unpacking, Np = non-processable

WIx: C = category change, W = weather, F = finalization, L = listing

WQx: F = finalization, I = intervention, C = comment, E = elaboration

the lowercase letters are inserted for readability.
The hierarchical organization of the DARs is shown in figure 1.

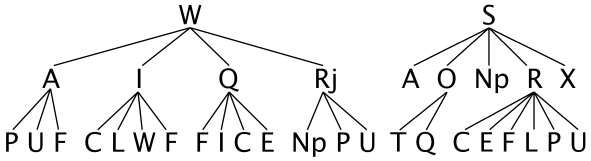

Figure 1: Hierarchy of the DAR for wizard and subject actions

The distribution of the different DARs can be seen in figure 2.
Several subsequent non-phonological events might be annotated with a single SNp – if they were transcribed within one segment. Sometimes – e.g. when there was a pause between two non-phonological utterances – they might both be transcribed separately. In this case the pause transcription was then filled in by the transcription software and all three segments were later annotated as SNp, leading to the

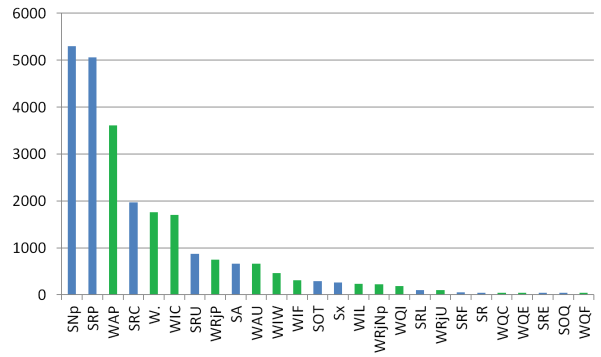

Figure 2: Frequency of the different DAR annotations. Wizard = green, subject = blue. Sum of all DAR annotations is 24963.

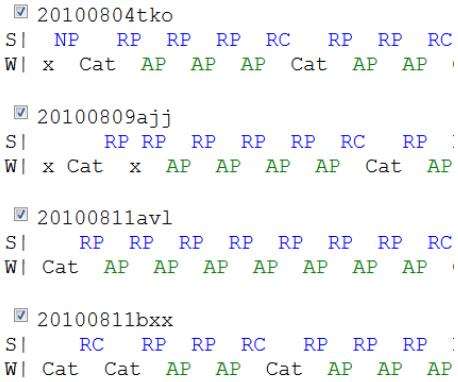

Figure 3: Beginning of the packing phase of four subjects in the dialog act sequence representation (two line version, where the first letter is ommitted because each speaker is represented in a separate line).

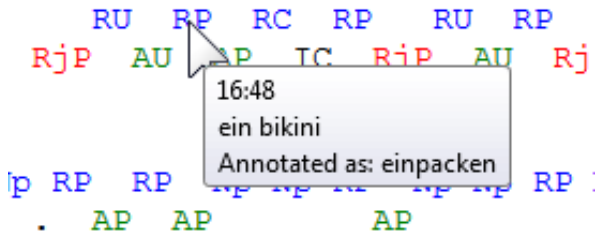

Figure 4: The mouseover text comprises time, text and annotation of the underlying transcript segment.

sequence: SNp SNp SNp. So the length of a SNp and the number of SNp annotations in a row may vary for similar cases.
Subject requests (SRx) may either be accepted (and performed, WAx) or rejected (WRjx = subject exceeded the weight limit, WRjNp = utterance could not be processed) by the system/wizard. Some subjects requested 'exchange X for Y' (SRE) – the system did not provide this functionality, but could either unpack (WAU) or pack (WAP), which was half of the task and considered as success here.
Accepted requests are further described in section 5.
We are interested in finding common and uncommon interaction patterns by means of exploring dialog act sequences. The DAR allows the quantitative analysis of many dialog act combinations – e.g. the system's reaction to offtalk and questions, the subjects' reaction to system questions (SQx) and information messages (SIx) – and even longer subsequences, but this is not in the focus of this paper and will be addressed in other publications.

4.4. LAST MINUTE workbench

The result of dialog act tagging of the corpus is available as an HTML file with the packing phase sequences for all subjects. The DAR is part of the workbench – a collection of tools for the exploration of the LAST MINUTE corpus. There is a one-lined version to search for sequences and a two-lined version for better readability where the first letter is ommitted and each speaker is represented in a separate line A part of the two-lined version can be seen in figure 3. Wizard dialog acts that are a clear rejection of a subject's action are colored red, confirming utterances are green, neutral and unclear ones are black. This ensures a quick overview of a subject's task success. The three barriers of the packing phase are marked by | symbols (cf. 2.).
All dialog act tags feature a mouseover text with the timepoint and text of the underlying transcript segment and the long form annotation. With this information every dialog act can be found in the original transcript. An example can be seen in figure 4.
Some interaction patterns are 'interrupted' (depending on what is looked for) by nonphonological segments (`SNp`). For all exploration tasks that are disturbed by `SNp` segments both representations, the one-lined and the two-lined, were also computed without the nonphonological segments. An example: There may be sequences like `SRC SNp WIC` in the original version. If a successful category change is searched (`SRC WIC`) it can only be found in the version without `SNp`.
In the following we consider the 5294 SNp as non-functional and work with a version without SNp.
This sequence visualisation allows a quick overview and the search for abstract interaction patterns. The HTML document also comprises a script for the calculation of dialog act bigram statistics. The bigram statistic can be calculated for specific user groups and for specific phases of the experiment with the script.

4.5. Bigram statistic

In figure 2 we can see, that there are many packing requests (SRP) but less packing confirmations (WAP). What happens after subject requests can be seen in a bigram statistic. In table 1 we show the bigram statistic for the packing phase. This is useful to find common bigrams or to analyze what other tag followed a certain tag (lines) or what happened before a certain tag (columns).
The bottom left and the top right of the four large boxes show the numbers for bigrams of one speaker. The top right large box shows bigram counts for subject dialog acts followed by a system response.

4.6. Ambiguous interpretation

For the proper differentiation of some dialog acts the context is needed, e.g. Wizard utterances announcing category

Tag	#	SA	SOQ	SOT	SR	SRC	SRE	SRF	SRL	SRP	SRU	Sx	W.	WAP	WAU	WIC	WIF	WIL	WIW	WQC	WQE	WQF	WQI	WRjNp	WRjP	WRjU	Sum
#													17			113											130
SA	29	161				3			1	13	1	1	86		1	4	1	2	188	1	45	4	121	2	2		666
SOQ	1		4	5		2				2	1			1	1		2	6	2	1				10	3	2	43
SOT	10		2	19		22		3	1	40	14	3	64	7	1	21	12	1	22	9			2	26	9	4	292
SR	1		1	1	3							3	15	2		6	1		1				2	8	2	1	47
SRC	41		1	4		59				16	2		230	14		1342	20	119	43			15		52	12	1	1971
SRE		1		1			6				1		4	2	11	3								4	9	2	44
SRF	11					1		3	1		1		8			3	2		2			22		2			56
SRL	1								3				2				3	89						3			101
SRP	20	7	1	8	2	27	1		2	282	14	36	165	3551	29	72	31	6	11	1		1	5	71	710	4	5057
SRU	1			4		1	1		3	14	48	10	29	1	615	4	17	1	1					31	6	81	868
Sx	4	1		3		7	2		1	21	3	10	70	49	6	23	31	3	2					16	10	2	264
W.	5	32	4	56	9	225	5	14	5	222	47	79	8			11			7								729
WAP	2	2	3	43	10	1164	1	7		2260	25	35	2	4		6	67	2	1								3634
WAU		2	2	4	2	28	7		2	369	173	17	3	1			54										664
WIC		4	6	22		192	6	4	3	1370	108	11				1									1		1728
WIF		5		16		41		4		35	4	19	1			118		4	66								313
WIL			1	3	3	1	3	1		137	67	3	1	1			11		1								233
WIW		157	1	20		15				37		3	22			1			32	34			58				380
WQC		40		3	1					1	2								1								48
WQE		45								1		1															47
WQF	1	34					1	5				1															42
WQI		168			2	2				14											2						188
WRjNp	2		7	18	5	59	2	2	6	70	25	12	1				16										225
WRjP	1	7	9	54	10	108	8	12	68	136	289	19	1	1			39			2							764
WRjU			1	8		14	1	1	5	17	43	1					6										97
Sum	130	666	43	292	47	1971	44	56	101	5057	868	264	729	3634	664	1728	313	233	380	48	47	42	188	225	764	97	18631

Table 1: Bigram statistic for the tags of the packing phase. The first tag of a bigram is on the left, the second on the top. As an example: The most common bigram is SRP WAP (successful packing) with 3551 occurrences. # stands for beginning or end of the packing phase.

changes. Here the same wording is used for any category change, be it system initiated or requested by the subject.

4.7. An example

The following DAR example is taken from a dialog segment where a subject (20110401adh) tries to pack a (winter) coat but the packing attempt is rejected several times (SRP WRjP pairs) and therefore the subject has to unpack several other items (SRU WAU) in order to create sufficient space. SNPs stand for nonphonological utterances of the subject and can be interpreted as expressions of the experienced efforts.

```
... SRP WRjP SNp SNp SRU WAU SRP WRjP SRU WAU SRP WRjP
SOT SNp SOQ SNp SNp SNp SRU WAU SRP WAP SOT ...
```

Below is an excerpt of the corresponding transcript. Following the GAT-2 minimal standard (Selting et al., 2009) short pauses are noted as (.) and (-), longer pauses with their duration in brackets, e.g. (1.77). English glosses are added for convenience. Please note the emotional expression of relief ('gott sei dank', engl. 'thank god') when the subject finally succeeds.

```
...
SRP     ein mantel
        [a coat]
  WRjP  der artikel mantel kann nicht hinzugefügt werden
        (.) anderenfalls würde die maximale
        gewichtsgrenze des koffers überschritten werden
        [the item coat cannot be added (.) otherwise the
        weight limit of your suitcase will be exceeded]
SNp     ((raschelt)) ((schmatzt))
        [rustles, smacks]
SNp     (-)
SRU     ein buch raus
        [one book out]
   WAU  ein buch wurde entfernt
        [a book has been removed]
SRP     ein mantel
        [a coat]
  WRjP  der artikel mantel kann nicht hinzugefügt werden
        (.) anderenfalls würde die maximale
        gewichtsgrenze des koffers überschritten werden
        [the item coat cannot be added (.) otherwise the
        weight limit of your suitcase will be exceeded]
SRU     badelatschen raus
        [beach slippers out]
   WAU  ein paar badelatschen wurden entfernt
        [a pair of beach slippers has been removed]
SRP     ein mantel
        [a coat]
  WRjP  der artikel mantel kann nicht hinzugefügt werden
        (.) anderenfalls würde die maximale
        gewichtsgrenze des koffers überschritten werden
        [the item coat cannot be added (.) otherwise the
        weight limit of your suitcase will be exceeded]
SOT     tja
        [well]
SNp     (1.77)
SOQ     was kann man denn noch rausnehmen
        [well what else can be removed]
SNp     (1.48)
SNp     pf pf pf pf pf pf pf
        [pf pf pf pf pf pf pf]
SNp     (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]
...
```

4.8. Relation to other tagsets

Our analyses are focused on the task specific semantic contents of a user's or wizard's utterance and do not take other information like length and number of pauses, breathing, non-task specific offtalk, questions etc. into account. Therefore most of the features of larger, standardized tagsets DAMSL (Core and Allen, 1997) or DiAML (Bunt et al., 2012) are not needed and we chose to construct our own DAR tailored specifically to the problem domain. This DAR can easily be mapped to DiAML or DAMSL annotations though. As an example DARs of the form SRx would be mapped to a *Directive* annotation in DAMSL while WAx/WRjx would be mapped to *Agreement Accept/Reject*, WRjNp to *Signal-Non-Understanding*, WIx to

Assert and so on.
We plan to investigate measures utilizing other tagsets than the DAR. These measures will incoporate information from e.g. the analysis of questions and answers, offtalk and the non-task-solving phase etc.

4.9. Errors and inconsistencies

In spite of intensive training and a detailed manual (Frommer et al., 2012b) the wizards did not always operate consistent and accurate. This is not surprising given the large number of subjects and the time span of nearly a year for the completion of all N = 130 experiments.
We found 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 some wizard errors, meaning thereby situations, where a wizard did not operate according to the guidelines of the manual. One type of such wizard error is the rejection of a subject request with 'your input could not be processed' (WRjNp) when indeed the intention of the subject was clearly recognizable and the intended action was performable.
DAR analysis uncovers candidates for such wizard errors and inconsistencies: at least 55 candidates for an erroneously rejected packing request, 31 for category change requests and 23 for unpacking, resp. These numbers show that the wizards did not work perfect but made only few mistakes which do not influence the following calculations significantly.

5. DAR based measures

The dialog act based representation (DAR) characterizes dialog courses with sequences of dialog act labels. This allows to define a variety of measures for dialog success or failure.
A first approach is to calculate the relative frequency of successful subject commands (SRx). Here we count the following bigrams as successful requests (numbers in brackets: number of successful action / number of action. Calculated for all subjects, cf. table 1):

packing:	SRP WAP	(3551/5057)
category change:	SRC WIC	(1342/1971)
unpacking:	SRU WAU	(615/868)
exchange:	SRE WAU/WAP	(13/44)
listing:	SRL WIL	(89/101)
finalization:	SRF WQF	(22/56)

For single subjects, the number of successful actions ranges from 19 to 74, the relative frequency of successful actions ranges from 39.6% to 87.5% with a mean of 69.6% and a standard deviation of 9.5%.
Table 2 shows the correlation of absolute and relative frequency of successful actions with age and gender calculated using Kendall's tau. Missing values where not significant on a $p < 0.05$ level.

5.1. Successful packing and unpacking actions

The variety of actions changes during the experiment. At the beginning the subjects (normally) only pack items and change the category – which is usually accepted by the system (SRP WAP). [2] After the weight limit barrier other actions become necessary: The subjects have to unpack items in order to pack others.
A more refined measure is e.g. given by the number of successful packing actions after the weight limit barrier in relation to the needed unpacking actions. This quantitative measure uncovers e.g. a subgroup of 8 (of 130) subjects that had not a single successful packing action after the weight limit at all in sharp contrast to a top performer group of 12 subjects with more than ten and up to 14 successful packings. Why users fail in such a drastic way is then a matter of qualitative in-depth analyzes of the transcripts.
Taking successful packing actions after the weight limit as dialog success measure: Top performers are subjects with 10 or more packing successes after weight limit. From a total of 24 subjects in this subgroup we find 13 that are young and female (cf. table 3), far more than would be expected in a random sample.
These measures operationalize how successful the packing/unpacking task was solved by the subjects.

Kendall's tau	age	gender
successes (abs.)		0.215^{**}
successes (rel.)	-0.197^{**}	0.168^{*}

Table 2: Tau statistics for sociodemographic features and successful actions. *$p < 0.01$; **$p < 0.001$.

age/sex	male	female
young	5	13
elderly	5	1

Table 3: Top performers after weight limit.

5.2. Initiative in LAST MINUTE dialogs

An important aspect of dialogs is the initiative: Which participant has initiative and is thus driving the dialog? Losing the initiative leads to losing time to solve the task. The subjects solve the task under time pressure so we assume losing initiative and time 'feels' less successful and driving the dialog 'feels' more successful.
While the system generally has the initiative during the personalization phase, the problem solving phase is primarily characterized by user initiative: the subject expresses a request (R) for a system action and, as a response, the system either confirms (A) or rejects (Rj) the request. An action may be rejected based on aspects of the user's utterance ('your request could not be processed') or – although the utterance was 'understood' and accepted – the action could be not performed for task related reasons. In sum: the pairs of successive dialog acts are SRx WAx, SRx WRjx or SRx WRjNp, but there are exceptions to this general rule.
One example is category change. If the allocated processing time of approximately one minute for a category is over the system takes initiative and performs an announced change of category (WIC).

[2] Few subjects e.g. try to unpack items before it is necessary, but this are only single cases.

Such system initiated category changes come in different structures:
Either the latest user request remains unprocessed (or ignored) resulting in a pair SRx WIC, or the request is performed and the change of category is announced immediately following the acceptance info resulting in a dialog act triple SRx WAx WIC.
There are also situations where a subject does not give further commands (e.g. subject makes a thinking pause or does not know how to proceed) and the system changes the category. The subjects did not initiate such category change (even when they might approve it).
Wizard induced category changes come in variety of patterns:

- SRP WAC
- SRU WAC
- SRP WAP WAC
- SRU WAU WAC
- ...

A user initiated category change follows the dialog act pattern SRC WIC.

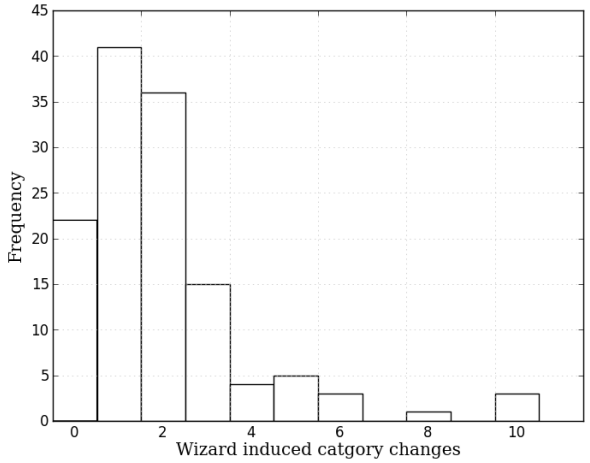

Figure 5: Number of wizard induced category changes.

We measure the wizard induced category changes for each subject. Figure 5 shows the distribution of this measure. The bulk (88%) of the subjects have between three and zero wizard induced category changes. The mean for this measure for all N = 130 subjects is 1.97, the standard deviation 1.96. After any category change the initiative is always again on the user side.
Self initiated changes of the category for selection are an indicator for the degree of control of the dialog flow and for the taken initiative that a subject exhibits. An in-depth analysis of self vs. system initiated category changes reveals a highly significant correlation between age and number of successful self initiated category changes after the weight limit barrier: calculating Kendall's tau for these two quantities reveals that, with a tau statistic of -0.23 and a p-value smaller than 10^{-4}, younger subjects have on average a higher number of such self initiated category changes than the elderly.

6. Summary

Both measures – successful actions and initiative – indicate how successful subjects were in the whole packing dialog or in parts of the interaction. This findings can be used for the correlation with personality and linguistic parameters to find indicators for critical dialog courses.
We have presented a dialog act based representation (DAR) for the problem solving dialogs in the LAST MINUTE corpus.
This representation is on the one hand abstract enough to generalize over the large variety of linguistic expressions in the user contributions. It is on the other hand specific enough to capture essentials of these contributions, i.e. the core dialog act and its underlying intention.
We have presented measures for success and failure of dialogs based on this representation. These measures allow to cluster transcripts into meaningful groups that deserve further qualitative analysis. In addition the relevant segments in the dialogs are identified. In other words: quantitative investigations guide subsequent qualitative in-depth interpretation of transcripts. Findings from such scrutiny are expected to inform the design of elaborated interaction strategies for future companion systems.
In addition the identification of those segments of the LAST MINUTE dialogs where problems accumulate or even escalate is an invaluable support for those collaborating groups that exploit audio or video records to train classifiers for their analyzes, be it emotion detection from facial expressions or from prosodic aspects of speech signals (Frommer et al., 2012a).
The identification of successful and unsuccessful bigrams can be used to calculate a task success value as in the PARADISE framework (Walker et al., 1997). We focus on explorative analyses and sequential properties here but plan to compare our measures to the PARADISE framework.

6.1. Future work

Ongoing and future work with the LAST MINUTE corpus includes:

- automate dialog act classification of user utterances,
- in depth analysis and classification of recorded offtalk,
- automatic detection of negative dialog situations,
- fine grained analyzes of failed dialogs in order to develop design guidelines for companion systems,
- investigate role of breathing, sighing and other non-phonological user contributions,
- finer grained analysis of pauses and hesitations in dialog course,
- compare to measures from the PARADISE framework (Walker et al., 1997),
- annotation of the dialog acts with more raters.

Acknowledgment

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. For the reviewers a sample from the corpus with anonymized data is available from the following URL `http://iws.cs.uni-magdeburg.de/a3/confs/index.htm` with loginname `reviewer` and password `review`.

7. References

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David R Traum. 2012. Iso 24617-2: A semantically-based standard for dialogue annotation. In *LREC*, pages 430–437.

Mark Core and James Allen. 1997. Coding dialogs with the damsl annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, pages 28–35.

Jörg Frommer, Bernd Michaelis, Dietmar Rösner, Andreas Wendemuth, Rafael Friesen, Matthias Haase, Manuela Kunze, Rico Andrich, Julia Lange, Axel Panning, and Ingo Siegert. 2012a. Towards emotion and affect detection in the multimodal LAST MINUTE corpus. In *LREC 2012 Conference Abstracts*, page 110. accepted.

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

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.

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 Laurence Devillers, Björn Schuller, Anton Batliner, Paolo Rosso, Ellen Douglas-Cowie, Roddy Cowie, and Catherine Pelachaud, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12) - Workshop Abstracts*, page 171, Istanbul, Turkey, may. European Language Resources Association (ELRA). Workshop 93: Corpora for Research on Emotion Sentiment and Social Signals (ES3).

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.

Dietmar Rösner, Manuela Kunze, Mirko Otto, and Jörg Frommer. 2012c. Linguistic analyses of the LAST MINUTE corpus. In Jeremy Jancsary, editor, *Proceedings of KONVENS 2012*, pages 145–154. ÖGAI, September.

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.

Website of the Transregional Collaborative Research Centre SFB/TRR 62. http://www.sfb-trr-62.de/.

Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 271–280. Association for Computational Linguistics.

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.