A Fully Annotated Corpus for Studying the Effect of Cognitive Ageing on Users' Interactions with Spoken Dialogue Systems

A Fully Annotated Corpus for Studying the Effect of Cognitive Ageing on Users' Interactions with Spoken Dialogue Systems

Kallirroi Georgila(1), Maria Wolters(1), Vasilis Karaiskos(1), Melissa Kronenthal(1), Robert Logie(2), Neil Mayo(1), Johanna D. Moore(1), Matt Watson(3)

(1) Human Communication Research Centre, University of Edinburgh
(2) Human Cognitive Neuroscience-Psychology, University of Edinburgh
(3) Psychology, University of Sunderland
kgeorgil@inf.ed.ac.uk (K. Georgila), mwolters@inf.ed.ac.uk (M. Wolters)

Abstract

In this paper we present a corpus of interactions of older and younger users with nine different dialogue systems. The corpus has been fully transcribed and annotated with dialogue acts and "Information State Update" (ISU) representations of dialogue context. Users not only underwent a comprehensive battery of cognitive assessments, but they also rated the usability of each dialogue system on a standardised questionnaire. In this paper, we discuss the corpus collection and outline the semi-automatic methods we used for discourse-level annotations. We expect that the corpus will provide a key resource for modelling older people's interaction with spoken dialogue systems.

1. Introduction

In the last decade, spoken dialogue systems have not only been a major research area, but they have also been widely adopted by industry. Most of the work to date focuses on young, healthy users. However, as the average life expectancy increases it will very soon become essential to design dialogue systems in such a way that they can accommodate older people's interaction styles and adapt to a wide range of cognitive abilities.

Despite the growing importance of older users, there is still a dearth of fully annotated corpora of interactions between older people and spoken dialogue systems. This is a particular problem for state-of-the-art statistical approaches to dialogue management (Lemon and Pietquin, 2007), since they crucially rely on adequate training data. Extrapolating from data collected with younger users, who are typically university students, may not be feasible. Firstly, cognitive abilities often deteriorate with age (Baeckman et al., 2001). Thus, an older user may well stumble over aspects of the system that a college student in their prime navigates with barely a glitch. Secondly, older users may well interact differently with spoken dialogue systems than younger users (Möller et al., 2008). Hence, user simulations (Georgila et al., 2005b; Georgila et al., 2006) based on data from younger users may be incapable of covering patterns of behaviour typical of older users.

In this paper, we present the MATCH[1] corpus, which consists of 447 interactions between older and younger users and spoken dialogue systems. The corpus was designed to provide researchers with a solid, extensively annotated data set that will allow them to investigate older users' interactions with spoken dialogue systems in depth. Our corpus is unique in the amount of additional information available for each participant. We include not only a comprehensive range of cognitive measures, but also extensive user satisfaction assessments for each of the 447 dialogues.

[1] http://www.match-project.org.uk

This paper is structured as follows. In Section 2 we outline the design of the corpus, which was collected as part of a cognitive psychology experiment. In Section 3, we discuss our data collection method. Then in Section 4, we present an overview of the manual and automatic techniques used for annotating the corpus. In Section 5, we present an initial comparison of the ways in which older versus younger users interact with our simulated appointment scheduling systems. In section 6, we propose ideas for future work. Finally, in Section 7, we present our conclusions.

2. Corpus Design

The MATCH corpus is one of very few corpora that was specifically designed to include older users. Although a couple of existing dialogue corpora contain data from older speakers, those were included more by accident than by design. There are two notable exceptions:

- the JASMIN-CGN corpus, which contains over 14 hours of interactions between Dutch and Flemish older users and a Wizard-of-Oz (WoZ) system (Cucchiarini et al., 2006),
- the MeMo corpus, which contains 62 interactions between 31 older and younger users and a Smart Home WoZ system (Möller et al., 2008).

While the JASMIN-CGN corpus was designed to cover a wide range of phonetic, phonological, and discourse phenomena, the MeMo corpus illustrates the effect of different types of help prompts on the way older and younger users interact with a spoken dialogue system. The MATCH corpus complements both of these corpora in that it was designed to examine the impact of cognitive ageing on users' interaction with spoken dialogue systems. All data was collected in the context of a cognitive psychology experiment (Wolters et al., 2008), where participants underwent an extensive battery of tests before interacting with the experimental dialogue systems. As a result, we have detailed data

on each user's cognitive abilities that may well be unique in corpora of human-machine interactions prepared for distribution.
In the original cognitive psychology experiment, we systematically varied:

1. the number of options that users were presented with (one option, two options, four options),

2. the confirmation strategy employed (explicit confirmation, implicit confirmation, no confirmation).

The combination of these 3×3 design choices yielded nine different dialogue systems.
Our design choices were motivated by an on-going, unresolved debate in the Human-Computer Interaction literature about the ideal number of options to be presented to older users. While some researchers advocate presenting fewer options (e.g. (Zajicek, 2004)) in order to ease the load on users' working memory, others have found that reducing the number of options either does not help (Huguenard et al., 1997) or is harmful (Commarford, 2006).
Users were asked to schedule a health care appointment with each of the nine systems, yielding a total of nine dialogues per user. We chose appointment scheduling as our domain for three reasons:

1. it is a well-understood example of the slot-filling paradigm,

2. it is a task familiar to both older and younger users,

3. it is highly relevant to telecare, an application domain with a large number of older users.

In order to assess the effect of users' cognitive abilities on their interaction with each of the nine systems, all participants underwent a comprehensive battery of cognitive assessments. This battery covered the two main dimensions of intelligence, fluid intelligence, which is linked to abstract reasoning, and crystallised intelligence, which is linked to acquired knowledge. We also assessed the speed of information processing and the capacity of working memory, the short term store for processing information.
After each interaction, users were asked to rate the system using a 39-item questionnaire. This questionnaire was based on the ITU-T recommendation P.851 as implemented in (Möller et al., 2007), one of the de-facto standards in the field. The questionnaire items included perceived task completion, overall impression, and user satisfaction. On completion of the questionnaire, which took about five minutes, participants were asked to recall four items of information about the appointment; health professional, day, time, and location. The short delay introduced by the questionnaire simulates a momentary distraction between the user hanging up the phone and noting down the appointment in their diary. Due to the length of the experiment, participants only booked one appointment with each system. Correct recall of the appointment was used as an additional measure of task success. Information about the appointments booked and recalled is included in the corpus together with the annotated dialogues.

1 Option (Yes/No):
System: Would you like to see the occupational therapist?
2 Options:
System: Would you like to see the occupational therapist or the community nurse?
4 Options:
System: Would you like to see the occupational therapist, the community nurse, the physiotherapist or the diabetes nurse?

Figure 1: Presentation of options.

3. Corpus Collection

Each of the nine systems was simulated using a WoZ design (Dahlbaeck et al., 1993). The human wizard took over the function of the speech recognition, language understanding, and dialogue management components. Simple templates were used for natural language generation. The resulting output sentences were spoken by the unit selection text-to-speech synthesiser Cerevoice (Aylett et al., 2006), which has been shown to be intelligible to older users (Wolters et al., 2007).
All dialogues followed the same overall structure: First, users arranged to see a specific health care professional, then they arranged a specific half-day, and finally, a specific slot on that half-day was agreed. In all three steps, the system initially presented the user with a fixed number of options: one (yes/no answer), two, or four (see figure 1). The user's choice was either confirmed explicitly through a confirmation dialogue, implicitly by mentioning the user's choice again in the next stage of the dialogue, or not confirmed at all (see figure 2). All dialogues were strictly system-initiative: The WoZ system not only controlled the choice of options presented to the user at each stage of the dialogue, it also did not allow users to skip stages by, say, requesting an appointment on a particular half-day at a particular time. This design ensured that all users were presented with the appropriate number of options and the appropriate confirmation strategy at least three times in each dialogue. Furthermore, system-initiative dialogue systems present fewer problems to the speech recognition component, resulting in better task completion (Black et al., 2005). Speech recognition for older people is known to be challenging compared to younger populations (Anderson et al., 1999; Müller et al., 2003). The reasons are manifold: poorer acoustics, age-related changes to vocal tract and vocal folds (Linville, 2000), unclear conversational content ("what can I say"), and variation due to stress induced by dealing with a computer. In a final step, the wizard confirmed the appointment, giving four pieces of information: the health professional, the day of the appointment, the time of the appointment, and the location of the appointment. All of these items, except for location, had been discussed earlier.
Overall, we recruited 26 older and 24 younger participants. Older participants were aged between 50 and 85, while younger participants were aged between 20 and 30. The older users contributed 232 dialogues, the younger ones 215. Three dialogues were not recorded due to problems with the recording equipment.

Explicit:
User: I would like to see the occupational therapist, please.
System: You would like to see the occupational therapist. Is that correct?
User: Yes.
Implicit:
User: I would like to see the occupational therapist, please.
System: When would you like to see the occupational therapist, on Monday afternoon or on Friday morning?
User: Monday afternoon would be best.
None:
User: I would like to see the occupational therapist, please.
System: When would you like to come, on Monday afternoon or on Friday morning?
User: Monday afternoon would be best.

Figure 2: Confirmation strategies.

The cognitive assessment battery consisted of four tests: the Mill Hill vocabulary test, which assesses crystallised intelligence (Raven et al., 1998), Raven's Progressive Matrices (Raven et al., 1998), which assess fluid intelligence, Digit/Symbol Substitution (Wechsler, 1981), which assesses information processing speed, and a working memory span test (Unsworth and Engle, 2005). Two of the older participants were unable to complete the working memory span test. More information about the assessment battery is given in (Wolters et al., 2008).
The demographics and cognition statistics of the participants are given in table 1. Table 2 shows some statistics about the number of dialogues, turns, and utterances in the corpus. Note that one turn may contain several utterances.

Variable	Older	Younger	Total
# Dialogues	232	215	447
# Turns	3316	2921	6237
# System Turns	1718	1564	3282
# User Turns	1598	1357	2955
# Utterances	4024	3215	7239
# System Utterances	1977	1796	3773
# User Utterances	2047	1419	3466

Table 2: Dialogue statistics.

4. Corpus Annotation

All dialogues were recorded digitally with a sampling frequency of 48 kHz and transcribed orthographically by an experienced human transcriber using the tool Transcriber (`http://trans.sourceforge.net`). The transcriber followed the guidelines developed by the AMI project (`http://www.amiproject.org`) for the creation of the AMI meeting corpus (Carletta, 2007).
All transcriptions and annotations are stored in NXT format (Carletta et al., 2003). Orthographic transcriptions are linked to the corresponding wave files. Information about users' scores on the cognitive tests, about the agreed appointment, about the recalled appointment, and about user satisfaction ratings are also stored in the NXT representation of each interaction.

In particular, our annotations are based on "Information State Update" (ISU) representations of dialogue context (Larsson and Traum, 2000). *Information States* are feature structures intended to record all the information about the preceding portion of the dialogue that is relevant to making dialogue management decisions. To our knowledge, this is the only corpus of older people's interactions with spoken dialogue systems that has been annotated with Information States and we expect that it will prove invaluable for learning dialogue strategies (Lemon and Pietquin, 2007) and user simulations (Georgila et al., 2005b; Georgila et al., 2006) for this type of population.
We have adopted the annotation format described in (Georgila et al., 2005a; Georgila et al., 2008) with a few modifications and improvements. Each user utterance is annotated with dialogue acts and Information States using a modified version of the automatic annotation system described in (Georgila et al., 2005a; Georgila et al., 2008). Modifications include a new parser, adaptation of the set of dialogue acts to the new domain, and extension of the Information State structure.
Figure 3 shows an example Information State. It corresponds to the dialogue state following the user utterance "Monday afternoon please but not at two, better at four", which replies to the system prompt "When would you like an appointment with the physiotherapist, on Monday afternoon or Thursday afternoon?".

4.1. Dialogue Act Annotations

In addition to orthographic transcriptions, the corpus has been annotated with dialogue acts. Although the terms speech act and dialogue act are often used interchangeably in literature, here, dialogue act is a concatenation of the speech act and task, i.e. `accept_halfday` corresponds to ⟨ `accept_info, halfday` ⟩. Using a unique mapping, we associate each dialogue act with a ⟨ speech act, task ⟩ pair where the speech act is task independent and the task corresponds to one of the three stages of the appointment scheduling dialogue. Table 7 depicts the list of user speech acts in the corpus.
In order to calculate the inter-annotator reliability, 3 experienced annotators annotated the same 36 dialogues (18 from older and 18 from younger people, 4 dialogues for each dialogue system) with a simpler version of the dialogue acts shown in Table 7 (some labels were merged). The resulting kappa score was 0.82 (Cohen, 1960).

4.2. Information State Annotations

Several of the features depicted in figure 3 simply specify the annotations for the current utterance. Others are various book-keeping features such as turn and utterance numbers. The most difficult problem in annotating dialogue context for slot-filling applications is determining which slots have been filled, confirmed, grounded, or even emptied, by a user utterance. In our ISU annotations we keep track of all these changes in the status of slots. We define a piece of information as "confirmed" only if it has been positively confirmed (after the system has explicitly or implicitly attempted to confirm it). There is no need to have a separate field for the value of the confirmed slot because the value which is con-

	Demographics			Cognition			
	# Users	Age	% female	MillHill	DSST	Ravens	WMS
Younger	24	22 ± 3	71%	42	75	54	37
Older	26	66 ± 9	61.5%	52	51	49	28
Δ sig.	n/a	**	n.s.	**	**	**	.

Table 1: Participant statistics .: $p<0.05$, **: $p<0.001$ or better.

Group	Speech Acts
Confirmation	confirm_pos, confirmimplicit_pos
Grounding	confirm_pos, confirmimplicit_pos
	reprovide_info_overall

Table 3: List of user speech acts associated with confirmations and grounding.

firmed must be the same as the value with which the slot has been filled. In the same way, a slot is "grounded" if it is either confirmed or if the system and the user have reached a mutual agreement regarding the status of this slot, indicated by the fact that the dialogue has moved to the next stage. Table 3 lists the speech acts which are associated with confirmations and grounding. Furthermore, the Information State contains fields about the slots that have been marked as unavailable by the user ("blocked") and their values.

Note also in figure 3 the difference between the groups of Information State fields { FilledSlotsHist, FilledSlotsValuesHist, BlockedSlotsHist, BlockedSlotsValuesHist, ConfirmedSlotsHist, GroundedSlotsHist } and { FilledSlotsStatus, FilledSlotsValuesStatus, BlockedSlotsStatus, BlockedSlotsValuesStatus, ConfirmedSlotsStatus, GroundedSlotsStatus }. The former give us information about the exact order in which slots have been filled, blocked, confirmed or grounded and may contain several instances of the same slot, e.g. the slot "hp" could be confirmed twice. The latter ("FilledSlotsStatus", etc.) inform us about the current status of the slots and thus may only contain one instance per slot. This distinction is very important because, for example, if a confirmed slot is refilled with a new value it will remain in the "ConfirmedSlotsHist" field even though its new value has not been confirmed yet. The history of dialogue acts, speech acts, and tasks is also included in our annotations.

Initially, the complete corpus was automatically annotated with dialogue acts and ISU representations of dialogue context. These annotations were then processed manually by an experienced human annotator. The annotator did not have to annotate the dialogues from scratch but only correct the automatic annotations, in particular, the dialogue acts, filled slots, filled slots values, blocked slots, blocked slots values, confirmed slots, and grounded slots. From the hand-corrected annotations, the automatic annotation tool then computed the list of ⟨ speech act, task ⟩ pairs that corresponded to each dialogue act and also dialogue history-level annotations, such as the current status of each of the slots required by the task, the history of speech acts, etc.

DIALOGUE LEVEL
Turn: user
TotalTurnNumber: 4
TurnNumber: 2
Speaker: user
TotalUtteranceNumber: 5
UtteranceNumber: 2
DialogueAct: [accept_halfday,social_polite,block_slot,provide_slot]
SpeechAct: [accept_info,social,block_info,provide_info]
TransInput: Monday afternoon please but not at two, better at four.
SystemOutput:

TASK LEVEL
Task: [halfday,polite,slot,slot]
FilledSlot: [halfday,slot]
FilledSlotValue: [monday pm,four pm]
BlockedSlot: [slot]
BlockedSlotValue: [two pm]
ConfirmedSlot: [hp]
GroundedSlot: [hp]

LOW LEVEL
Segmentation: [monday afternoon],[please],[but not at two,],[better at four]

HISTORY LEVEL
FilledSlotsStatus: [hp],[halfday],[slot]
FilledSlotsValuesStatus: [physiotherapist],[monday pm],[four pm]
BlockedSlotsStatus: [slot]
BlockedSlotsValuesStatus: [two pm]
ConfirmedSlotsStatus: [hp]
GroundedSlotsStatus: [hp]
DialogueActsHist: greeting,suggest_hp_2,[accept_hp,social_polite], suggest_halfday_2_implicit,[accept_halfday,social_polite, block_slot,provide_slot]
SpeechActsHist: opening_closing,suggest_2,[accept_info,social], suggest_2_implicit,[accept_info,social,block_info,provide_info]
TasksHist: greeting,hp,[hp,polite],halfday,[halfday,polite,slot,slot]
FilledSlotsHist: [hp],[halfday,slot]
FilledSlotsValuesHist: [physiotherapist],[monday pm,four pm]
BlockedSlotsHist: [slot]
BlockedSlotsValuesHist: [two pm]
ConfirmedSlotsHist: [hp]
GroundedSlotsHist: [hp]

Figure 3: Example dialogue context/Information State in text format, simplified from the NXT format. User-provided information appears between [ ] brackets.

5. Older Versus Younger Users

In this section, we present an initial comparison of the ways in which older versus younger users interact with our simulated appointment scheduling systems. Since we do

Variable	Older	Younger	Sig.
# Turns	79	59	***
# Word Types	81	30	***
# Word Tokens	312	102	***
# Speech Act Types	14	9	***
# Speech Act Tokens	126	73	***

Table 4: Overall dialogue-level differences. Numbers are summed over all dialogues and divided by the number of users ***:p<0.0001 or better.

not have part-of-speech tagging and parsing at present, our analysis concentrates on the lexical level and the dialogue act level. All significance tests were Wilcoxon tests conducted using R (R Development Core Team, 2006). Due to the large number of tests, comparisons that are significant at p<0.05 are treated as only barely significant.

Looking at overall dialogue statistics, we see that older users produce longer dialogues than younger users (table 4). They also have a richer vocabulary and use a larger variety of speech acts. While the three most frequent speech acts always account for more than half of younger users' total speech acts, the proportion can vary between 30% and 70% for older users (figure 4). The difference in vocabulary is even more drastic: 30%–50% of all words spoken by younger users are instances of the three most frequent lexical items (figure 5), whereas the three most frequent lexical items may only cover as little as 10%–30% of all words spoken by older users. Figures 4 and 5 also demonstrate that there is no such thing as a stereotypical older user. In both cases, the variation observed in the older users subsumes most of the variation seen in the younger users.

Table 5 illustrates some typical speech act patterns. Younger users tend to restrict themselves to speech acts that are of immediate relevance to the task. 73.6% of all speech acts produced by younger users are variations of `accept_*`, where users accept options presented by the system, and `confirm_*`, where users confirm a slot. For older users, that proportion falls by nearly a third to 50.4%. The additional speech acts come from two main groups:

- instances of social interaction with the system, such as bidding the system goodbye or thanking it for providing information,
- instances of the user taking the initiative, such as users giving details about the slots that they can or cannot make.

The relative frequencies of selected lexical items, which are summarised in table 6, show a very similar pattern. A third of all words uttered by younger users are "yes" and "no" (category `YesNo`). This percentage drops dramatically to 13.0% for our older users. Moreover, when older users express approval or disapproval, they are more likely to use expressions other than "yes", such as "fine" (category `PosNeg`). As we would expect from our speech act analysis, older users are also more likely to use expressions that are more appropriate in human/human interactions (category `SocWords`), such as forms of "goodbye" (category `Bye`) or "thank you" (category `Thanks`). When comparing our statistics to the word-level analyses of the MeMo corpus (Gödde et al., 2008), we see that the social interaction words that distinguish between older and younger users appear to be task-specific. While older people were significantly more likely to use forms of "please" in the MeMo command-and-control task, we did not find a significant difference in the appointment scheduling context.

Speech Act	Older	Younger	Sig.
Accept_*	22.1	32.1	***
Confirm_*	28.3	41.5	***
Social	17.9	5.3	***
Acknowledge	0.8	0.0	***
Provide_*	7.8	3.4	*
Reprovide_*	1.8	0.2	**
Block	0.5	0.0	*
Garbage	3.2	0.5	***

Table 5: Differences in relative frequencies of speech acts *:p<0.01, **:p<0.001, ***:p<0.0001 or better.

Lexical Cat.	Older	Younger	Sig.
YesNo	13.0	33.8	***
PosNeg	4.1	1.8	*
SocWords	7.7	3.6	*
Thanks	2.3	0.3	***
Bye	1.2	0.2	***
Please	4.0	2.9	n.s.
Sorry	0.2	0.1	n.s.

Table 6: Differences in relative frequencies of lexical categories *:p<0.01, **:p<0.001, ***:p<0.0001 or better.

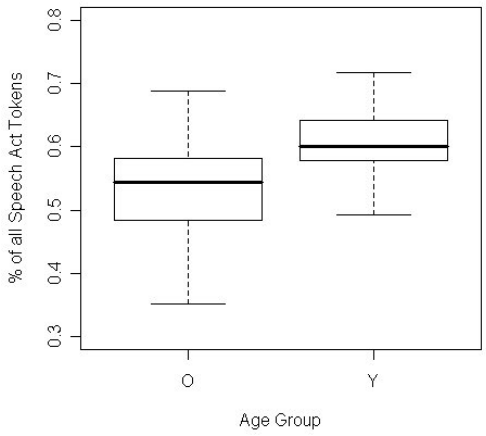

Figure 4: Relative frequency of the three most frequent speech acts.

6. Future Work

In the future we intend to annotate the corpus with part-of-speech tags and syntactic information and use these annotations to study further the differences between older and younger users. We will also train user simulations and dialogue strategies for these two types of users. Furthermore,

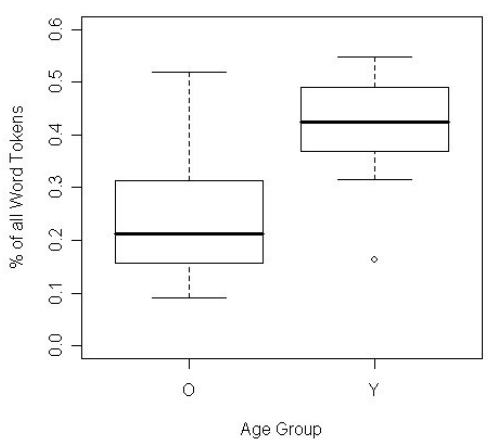

Figure 5: Relative frequency of the three most frequent words.

we will use the wave files and the transcriptions of the user utterances in order to train acoustic models for older people or adapt existing general models to the peculiarities of older people's speech. As previously mentioned, speech recognition for older people is known to be challenging compared to younger populations (Anderson et al., 1999; Müller et al., 2003).

7. Conclusions

We have presented a richly annotated corpus of older and younger users' interactions with simulated spoken dialogue systems that contains information about task success, task completion, users' cognitive abilities, and users' subjective ratings of each system. All of this information has been stored using the open standard NITE XML (Carletta et al., 2003). We hope that this corpus will prove a rich resource for learning dialogue management strategies, creating realistic user simulations, investigating how older users interact with dialogue systems, assessing the impact of cognitive ageing on spoken human-machine interaction, and last, but not least, adapting speech recognisers to older voices.

8. Acknowledgments

This research was supported by the MATCH project (SHEFC-HR04016, `http://www.match-project.org.uk`) and the Wellcome Trust VIP Award. We would like to thank Mark Core for our interesting discussions regarding the set of dialogue acts. We also thank the anonymous reviewers.

9. References

S. Anderson, N. Liberman, E. Bernstein, S. Foster, E. Cate, B. Levin, and R. Hudson. 1999. Recognition of Elderly Speech and Voice-Driven Document Retrieval. In *Proc. ICASSP*.

M. Aylett, C. Pidcock, and M.E. Fraser. 2006. The Cerevoice Blizzard Entry 2006: A prototype database unit selection engine. In *Proc. BLIZZARD Challenge*.

L. Baeckman, B. J. Small, and A. Wahlin. 2001. Aging and memory: Cognitive and biological perspectives. In J. E. Birren and K. W. Schaie, editors, *Handbook of the Psychology of Aging*, pages 349–377. Academic Press, San Diego, CA etc.

L. A. Black, C. McMeel, M. McTear, N. Black, R. Harper, and M. Lemon. 2005. Implementing autonomy in a diabetes management system. *J Telemed Telecare*, 11 Suppl 1:6–8.

J. Carletta, S. Evert, U. Heid, J. Kilgour, J. Robertson, and H. Voormann. 2003. The toolkit: Flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments, and Computers*, 35:353–363.

Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation*, 41(2):181–190.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

P. Commarford. 2006. *Working Memory, Search, and Signal Detection: Implications for Interactive Voice Response System Menu Design*. Ph.D. thesis, University of Central Florida.

Catia Cucchiarini, Hugo Van Hamme, Olga van Herwijnen, and Felix Smits. 2006. Jasmin-cgn: Extension of the spoken dutch corpus with speech of elderly people, children and non-natives in the human-machine interaction modality. In *Proc. LREC*, pages 135–138.

N. Dahlbaeck, A. Joensson, and L. Ahrenberg. 1993. Wizard of oz studies - why and how. *Knowledge-Based Systems*, 6:258–266.

Kallirroi Georgila, Oliver Lemon, and James Henderson. 2005a. Automatic annotation of COMMUNICATOR dialogue data for learning dialogue strategies and user simulations. In *Proc. DIALOR*.

Kallirroi Georgila, Oliver Lemon, and James Henderson. 2005b. Learning user simulations for information state update dialogue systems. In *Proc. Interspeech*.

Kallirroi Georgila, James Henderson, and Oliver Lemon. 2006. User simulation for spoken dialogue systems: Learning and evaluation. In *Proc. Interspeech*.

Kallirroi Georgila, Oliver Lemon, James Henderson, and Johanna D. Moore. 2008. Automatic annotation of context and speech acts for dialogue corpora. *Natural Language Engineering, submitted*.

Florian Gödde, Sebastian Möller, Klaus-Peter Engelbrecht, Christine Khnel, Robert Schleicher, Anja Naumann, and Maria Wolters. 2008. Study of a speech-based smart home system with older users. In *International Workshop on Intelligent User Interfaces for Ambient Assisted Living*, pages 17–22.

B. W. Huguenard, F. J. Lerch, B. W. Junker, R. J. Patz, and R. E. Kass. 1997. Working memory failure in phone-based interaction. *ACM Transactions on Computer-Human Interaction*, 4:67–102.

Staffan Larsson and David Traum. 2000. Information state and dialogue management in the trindi dialogue move engine toolkit. *Natural Language Engineering*, 6(3-4).

Speech Act	Description
Accepting / Rejecting System Suggestions	
accept_info	user explicitly accepts option suggested by the system
accept_info_yes	user accepts option by saying "yes"
accept_info_null	user implicitly accepts option suggested by the system
accept_info_prevprovided	user explicitly accepts option that s/he had previously provided
accept_info_yes_prevprovided	user accepts option that s/he had previously provided by saying "yes"
accept_info_null_prevprovided	user implicitly accepts option that s/he had previously provided
reject_info	user explicitly rejects option suggested by the system
reject_info_no	user rejects option suggested by the system by saying "no"
reject_info_null	user implicitly rejects option suggested by the system
confirm_pos	user confirms an option when asked for confirmation
confirmimplicit_pos	user continues with dialogue after implicit confirmation request by the system
confirm_neg	user rejects an option when asked for confirmation
yes_answer	user answers "yes" to system question
no_answer	user answers "no" to system question
Correcting System / Indicating Misunderstandings	
correct_info	user corrects previously provided information
correct_info_no	user corrects previously provided information by saying "no"
correctblock_info	user corrects previously provided information about options that are not possible
signal_misunderstanding	user signals that system has misunderstood previous utterance
request_info	request for help, clarification, or repetition
Taking Initiative	
provide_info	user provides information about possible options
provideblock_info	user provides information about options that are not possible
reprovide_info	user provides information again in the same utterance or turn
reprovide_info_overall	user provides information again for slot that has already been filled
reprovide_info_overall_notfilled	user provides information again for slot that has not been filled yet
reprovideblock_info	user provides information again about options that are not possible
reprovideblock_info_overall	user provides information again for slot that has already been marked as unavailable
repeat_info	user repeats information given by system in an explicit/implicit confirmation
repeatblock_info	user repeats information about options that are not possible
repeat_info_misunderstanding	user repeats information as a reaction to a misunderstanding
Social Interaction with the System	
acknowledgement	user shows that s/he can understand the system
social	social interaction with the system, e.g. "goodbye", "thank you"

Table 7: List of user speech acts.

Oliver Lemon and Oliver Pietquin. 2007. Machine learning for spoken dialogue systems. In *Proc. Interspeech*.

S. E. Linville. 2000. The aging voice. In R. Kent, editor, *Voice Quality Measurement*, pages 359–376. Singular, San Diego, CA.

S. Möller, P. Smeele, H. Boland, and J. Krebber. 2007. Evaluating spoken dialogue systems according to de-facto standards: A case study. *Computer Speech and Language*, 21(1):26–53.

Sebastian Möller, Florian Gödde, and Maria Wolters. 2008. Corpus analysis of spoken smart-home interactions with older users. In *Proc. LREC*.

C. Müller, F. Wittig, and J. Baus. 2003. Exploiting Speech for Recognizing Elderly Users to Respond to their Special Needs.

R Development Core Team, 2006. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

J. Raven, J.C. Raven, and J.H. Court, 1998. *Manual for Raven's Progressive Matrices and Vocabulary Scales.* Harcourt Assessment, San Antonio, TX.

N. Unsworth and R. W. Engle. 2005. Individual differences in working memory capacity and learning: evidence from the serial reaction time task. *Mem Cognit*, 33:213–20.

D. Wechsler, 1981. *Manual for the Wechsler Adult Intelligence Scale-Revised. New York*. The Psychological Corporation, New York.

Maria Wolters, Pauline Campbell, Christine DePlacido, Amy Liddell, and David Owens. 2007. Making synthetic speech accessible to older people. In *Proc. Sixth ISCA Workshop on Speech Synthesis, Bonn, Germany*.

Maria Wolters, Kallirroi Georgila, Robert Logie, Sarah MacPherson, Johanna Moore, and Matt Watson. 2008. Accommodating cognitive ageing: Do we need fewer options? *submitted*.

M. Zajicek. 2004. Successful and available: interface design exemplars for older users. *Interacting with Computers*, 16:411–430.