Being Old Doesn’t Mean Acting Old: How Older Users Interact with Spoken Dialog Systems

MARIA WOLTERS, KALLIRROI GEORGILA,
JOHANNA D. MOORE, and SARAH E. MACPHERSON
University of Edinburgh

Most studies on adapting voice interfaces to older users work top-down by comparing the interaction behavior of older and younger users. In contrast, we present a bottom-up approach. A statistical cluster analysis of 447 appointment scheduling dialogs between 50 older and younger users and 9 simulated spoken dialog systems revealed two main user groups, a “social” group and a “factual” group. “Factual” users adapted quickly to the systems and interacted efficiently with them. “Social” users, on the other hand, were more likely to treat the system like a human, and did not adapt their interaction style. While almost all “social” users were older, over a third of all older users belonged in the “factual” group. Cognitive abilities and gender did not predict group membership. We conclude that spoken dialog systems should adapt to users based on observed behavior, not on age.

Categories and Subject Descriptors: H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Ergonomics, interaction styles, natural language, user-centered design*

General Terms: Design, Experimentation, Human Factors

Additional Key Words and Phrases: Aging, spoken dialog systems, clustering, cognitive aging, voice interfaces

ACM Reference Format:
Wolters, M., Georgila, K., Moore, J. D., and MacPherson, S. E. 2009. Being old doesn’t mean acting old: How older users interact with spoken dialog systems. ACM Trans. Access. Comput. 2, 1, Article 2 (May 2009), 39 pages. DOI = 10.1145/1525840.1525842.
http://doi.acm.org/10.1145/1525840.1525842.

This research was supported by the MATCH project (SFC-HR04016) and a Wellcome Trust Value-In-People award to K. Georgila.

Authors’ addresses: M. Wolters, K. Georgila, and J. D. Moore, School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, United Kingdom; email: mwolters@inf.ed.ac.uk; S. E. MacPherson, Centre for Cognitive Aging and Cognitive Epidemiology, Human Cognitive Neurosciences, University of Edinburgh, United Kingdom.

1. INTRODUCTION

In the future, more and more older people will find themselves interacting with Spoken Dialog Systems (SDS). In contrast to traditional touch-tone Interactive Voice Response (IVR) systems, full SDS take speech input and produce speech output. SDS have been successfully deployed in a range of commercial applications, from flight information to cinema ticket booking. Increasingly, SDS are being developed for telecare applications, such as diabetes symptom management [Black et al. 2005] and behavioral interventions designed to decrease hypertension [Giorgino et al. 2005]. Although these systems were not specifically adapted to older people, they address conditions which become more prevalent with age. Therefore, they need to accommodate older users' needs, abilities, and preferences. SDS have also been used in task and medication reminder systems such as the NURSEBOT robot [Roy et al. 2000; Pollack et al. 2003], which was tested in retirement homes with care facilities. Finally, SDS have been developed for controlling smart homes [Möller et al. 2006]. Such SDS can be extended to support environmental control systems, which can help older people remain in their own home for longer and improve their quality of life [Tang and Venables 2000].

Much recent work in SDS and related fields has focused on adapting systems to the needs and preferences of the user [Buchanan et al. 1995; Walker et al. 2005; Carenini and Moore 2006; Demberg and Moore 2006]. It is well known that older users are a particularly difficult group to design for because of their diverse needs and abilities [Gregor et al. 2002]. When adapting to older users, we have several choices:

(1) design a single strategy that works well for all types of users,
(2) design for prototypical older and younger users,
(3) design a system that works with a range of user profiles, including those of "extreme" older users [Pullin and Newell 1997].

The first approach works well if a given strategy indeed benefits both older and younger users, regardless of age-related changes in ability. An example of such a result are the guidelines for adapting speech synthesis to older users that were proposed by Wolters et al. [2007]. Although these guidelines have the potential to make synthetic speech easier and more pleasant to listen to for all age groups, the beneficial effect will be higher for older users. The second approach, designing for prototypical older and younger users, requires designers to identify user age reliably and to develop specific dialog strategies tailored to the typical auditory, vocal, and cognitive abilities of older users. For example, Müller et al. [2003] classify users into age groups according to their voice. This information is then passed on to the underlying SDS, which selects a dialog strategy that can accommodate the effects of cognitive aging. The third approach motivates detailed studies of fully implemented systems or system prototypes with a small number of older users, such as Zajicek et al. [2004], that focus on the needs and problems of individual users.

But what if the relevant user groups are not delineated by age, but by characteristic patterns of behavior? This is the question we pursue in this study.

Using data from a large corpus of 447 interactions between older and younger users and simulated SDS, we investigate whether users can be grouped according to the way in which they interact with the system. The statistical analysis methodology that we use allows us to both identify "extreme" users and "typical" users. Age does not enter into the initial analysis at all.

We quantify the *interaction style* of each user based on a detailed linguistic analysis of their dialogs. For the purpose of this article, interaction style encompasses the linguistic choices that users make when they interact with an SDS. Examples of relevant linguistic choices are choosing between different expressions of assent (e.g., "yes" versus "that's fine"), or choosing to use politeness markers such as "please".

The interactions were collected during an experimental laboratory study that compared different dialog strategies for accommodating cognitive aging [Wolters et al. 2009]. In this laboratory study, we observed how users interact with a range of spoken dialog systems, and measured three central facets of usability [ISO 1998]: effectiveness, efficiency, and user satisfaction.

In the present article, we address three research questions, next given.

(1) Can users be categorized into distinct groups depending on how they speak to the system? If yes, how can these groups be characterized?
(2) Can the interaction style of a user be predicted?
(3) Does interaction style affect usability?

The article is structured as follows. In Section 2, we introduce important theoretical concepts related to the study of SDS, review relevant work on adapting SDS to older users, and summarize the literature on cognitive aging and language production that underpins our design and analysis. In Section 3, we describe the dataset that was analyzed for this study. The statistical methodology used in this study, cluster analysis, is described in Section 4. In Section 5, we report the clusters found in our data, discuss how these clusters relate to age and cognitive abilities, and examine the effect of interaction style on usability. The implications of our results for the design of spoken dialog systems for older people are discussed in Section 6. We conclude with a description of future work in Section 7.

2. LITERATURE REVIEW

2.1 What Are Spoken Dialog Systems?

Spoken dialog systems (SDS) enable users to interact with computers naturally and efficiently using one of the most natural communication modalities of all, speech. Due to their potential for commercial exploitation as well as the technological challenges they impose, SDS have attracted increased attention in both industry and research.

SDS have been developed for many different domains, including information provision [Seneff et al. 1998; Levin et al. 2000; Raux et al. 2006], command-and-control [Paek and Chickering 2007], tutoring [Zinn et al. 2002; Litman and Silliman 2004], simulation-based training [Traum et al. 2008], controlling smart homes [Möller et al. 2006], delivering reminders [Roy et al. 2000; Pollack et al. 2003], telecare [Giorgino et al. 2005; Black et al. 2005], and companionship [Catizone et al. 2008].

SDS typically consist of five main components. Automatic Speech Recognition (ASR) converts audio signals of human speech into text strings, Natural Language Understanding (NLU) determines the meanings and intentions of the recognized utterances, Dialog Management (DM) controls the interaction, Natural Language Generation (NLG) generates the text of system responses, and Text-To-Speech synthesis (TTS) converts the system utterances into actual speech output.

Based on the dominant dialog management strategy, SDS can be divided into three main categories: system-initiative, mixed-initiative, and user-initiative. Most currently deployed commercial SDS are system-initiative, that is, the user is not allowed to deviate from the dialog script imposed by the system, which may lead to long and tedious interactions and generally unnatural dialogs. At the other extreme, in user-initiative SDS the user has full control and is free to change the dialog structure as desired. Given the limitations of current ASR and NLU systems, user-initiative often leads to many errors and misunderstandings throughout the interaction. In task-oriented dialogs, task completion rates are low and users are frequently left frustrated. Mixed-initiative interaction allows both system and user to take initiative. The user is expected to respond to system prompts but may also provide more information than the system requests. This is called "overanswering". Another dialog phenomenon widely studied in the dialog literature is "grounding". This is the process by which conversational participants develop a body of mutually agreed upon information [Traum 1994]. Different types of users exhibit different grounding behavior.

Current research in dialog focuses on building systems that can dynamically adapt to user behavior and dialog context [Chickering and Paek 2007; Dzikovska et al. 2007] or that can be tailored to specific user models [Carenini and Moore 2000; Moore et al. 2004; Winterboer and Moore 2007] in order to increase efficiency and user satisfaction. At the same time a new paradigm is emerging: automatically learning dialog strategies from data and/or user simulations using statistical optimization methods [Young 2000; Georgila et al. 2008; 2005; Lemon et al. 2006]. These statistical approaches are very attractive because of their potential for efficient development and automatic optimization of dialog systems, and easy adaptation of existing applications to new domains.

In terms of NLU, most commercial systems are based on shallow parsing techniques such as identifying predefined words and phrases [Swift and Allen 2004]. This is sufficient for applications where user input is restrained to simple answers such as "yes", "no", "Monday afternoon", "three p.m.", but it breaks down when confronted with sentences such as "I can only make afternoons on Mondays". Such a sentence is ambiguous: It can mean that the user is only free on Monday afternoons, but not on other afternoons, or it can mean that the user cannot make Monday mornings. In order to cope with such sentences, researchers are developing complex syntactic and semantic processing modules

[Klein and Manning 2003; Nivre and Nilsson 2005; Dzikovska et al. 2007]. In general, the more complex the user behavior the dialog system aims to deal with, the more advanced its NLU component should be.

2.2 Why Does Interaction Style Matter?

The way in which users interact with an SDS affects all system components. Here, we will concentrate on the effects on ASR, NLU, and DM.

Most modern ASR systems consist of two components:

(1) An *acoustic model* which uses information about the acoustic features of an utterance to generate a list of hypotheses about the words the user may have spoken. Typically, the acoustic features are derived from the frequency spectrum of the utterance.

(2) A *language model* which establishes constraints on possible word sequences. Effectively, language modeling ensures that the words included in the recognition hypotheses will be in the correct context and follow some syntactic structure.

The larger the vocabulary and richer the syntax of users' utterances, the more complex the language model needs to be. However, due to performance constraints, typically language models only take into account the previous two words and ignore larger contexts. This makes them unable to model longer-range dependencies.

In addition, NLU components typically only cover a very restricted set of inputs. Substantial research and large computational resources are required to deal with complex linguistic phenomena and perform sophisticated semantic processing.

Whilst SDS should be able to accommodate a range of individual interaction styles, the system may not have to do all the work itself. It is well known that people tend to adapt their interaction style to their conversational partner. This tendency has been explored in great detail by social or communication accommodation theory [Giles 2001]. Niederhoffer and Pennebaker [2002] demonstrate that people match their linguistic styles both in verbal human-human interactions and in computer-mediated interactions such as conversations in chat rooms. The tendency to adapt one's language to one's communication partner is a largely unconscious effect of a fundamental process: *alignment* of situation models [Pickering and Garrod 2004; Branigan et al. 2009]. Situation models are complex cognitive models of the situation under discussion in the current dialog. In order to have a successful conversation, dialog partners need to align their situation models with each other. This is often achieved covertly by adopting aspects of each other's speech, such as syntactic constructions or ways of referring to objects.

These alignment processes are also at work when people interact with machines [Branigan et al. 2009]. People tend to treat computers as social actors [Nass and Brave 2005], projecting emotion and personality onto the machine. Oviatt et al. [2004] showed that children adapted their intonation to the intonation of the computer-generated characters they were interacting with. Branigan et al. [2003] observed that users adapt their syntax to that of their interlocutor, regardless of whether they believe that they are talking to a computer or to a real person. Word choice is also influenced by users' theories about the system they interact with [Pearson et al. 2006]. Users were more likely to adapt their vocabulary to the system's vocabulary if they believed that the system's ability to understand their utterances was restricted. Since users adapt their linguistic style to incorporate words and syntax modeled by computers, a system can "shape" user input to ensure successful interactions by providing adequate templates at strategic points during the interaction [Leiser 1989; Zoltan-Ford 1991; Sheeder and Balogh 2003; Wolters et al. ms].

2.3 Adapting SDS to Older Users

Existing work on adapting SDS to older people falls into two main groups: experimental assessments of full end-to-end systems [Pollack et al. 2003; Zajicek et al. 2004; Black et al. 2005; Giorgino et al. 2005] and guidelines that are largely based on the literature on cognitive and perceptual aging [Hawthorn 2000; Petrie 2001; Gregor and Dickinson 2007].

Many of the end-to-end systems, such as the HOMEY hypertension management system [Giorgino et al. 2005], are mixed-initiative. HOMEY asked patients with hypertension questions about lifestyle and relevant symptoms. When asked about their blood pressure, patients could give systolic and diastolic pressure in one utterance or they could wait for the system to ask for each value in turn. However, in mixed-initiative systems, user utterances can be complex. Without adequate ASR and NLU models, task success may decline. To improve reliability, system developers may opt for a more system-initiative design [Black et al. 2005].

Sharit et al. [2003] demonstrated that cognitive aging affects the usability of touch-tone Interactive Voice Response (IVR) systems. They found that older users performed less well on both easy and highly complex tasks than younger and middle-aged users. Older users were also less efficient and took longer to call the system. When older users were provided with a graphical aid that explained the menu structure of the IVR systems they were operating, performance improved. Sharit et al. [2003] expect similar results for systems that replace touch-tone by speech input, even though systems that accept speech input do not require users to remember mappings between tasks and keys on the telephone keypad.

Gödde et al. [2008] presented 15 older and 16 younger users with two versions of a smart home system. In the first version, context-sensitive help was given early on in the dialog, in the second version, context-sensitive help was only given when needed. Older users had lower task success and were less likely to speak to the system in a way that was easy for the system to understand. However, when given help early on in the dialog, older people were able to adapt their interaction style, in particular their vocabulary [Wolters et al. ms].

The dataset used for the present study differs from that of Gödde et al. [2008] in several important aspects. We used a different domain (appointment scheduling), we investigated more dialog strategies (nine), and our corpus was annotated in more linguistic detail (see Georgila et al. [2008] for a detailed overview of the annotations of the present corpus). Finally, while Gödde et al. [2008] only assessed the short-term memory span of their participants, our users took part in a comprehensive battery of cognitive tests (refer to Section 3.1.1).

2.4 Relevant Properties of Older Users

2.4.1 *General Cognitive Aging.* Cognitive abilities such as processing speed, working memory, and fluid intelligence affect the extent to which people can successfully use computers [Czaja and Lee 2007]. In healthy adult aging, these abilities decline [Salthouse 2004], with the decline starting as early as middle age [Garden et al. 2001]. However, there is considerable interindividual variation in terms of the speed and extent to which abilities change with increasing age [Rabbitt and Anderson 2006]. While some of this variation is predicted by an individual's cognitive abilities in childhood, it can also be attributed to the events experienced during his/her lifetime [Deary et al. 2004]. The wide range of individual differences among older adults makes it very difficult to define prototypical older users for the design process.

Aging may also affect our ability to process social information. As people age, they may find it more difficult to take another person's perspective and to attribute and reason about the mental states of others [Maylor et al. 2002; Bailey and Henry 2008]. This ability, often referred to as *theory of mind*, is crucial when interlocutors adapt their own cognitive model of the situation to resemble the other's model. This may affect language production. Horton and Spieler [2007] have shown that older people are less likely to adapt textual descriptions to their audience.

2.4.2 *Changes in Language Production.* Not all aspects of language production decline with age. Most notably, vocabulary, which is linked to acquired knowledge, is preserved [Schaie 1994; Verhaeghen 2003]. However, older people may experience word finding problems [Burke and Shafto 2004]. Patterns of word use also change during the life span. Pennebaker and Stone [2003] observe that older people use fewer words that denote negative emotions and fewer self-referential words. In addition, aspects of language production that are linked to general physiological and cognitive aging are affected by age. Anatomical and physiological changes to the vocal tract affect a person's voice [Ramig et al. 2001]. Since working memory is needed for sentence construction, sentence length and sentence complexity decrease as the working memory span shrinks with age [Nef and Hupet 1992; Kemper et al. 2003]. The speech of older people tends to be less coherent, with more ambiguous references [Nef and Hupet 1992] and more off-topic speech [James et al. 1998]. At the same time, the overall structural complexity of discourse increases. Older people are more likely to talk about multiple episodes and embed stories into the ongoing episode [Nef and Hupet 1992].

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?

Fig. 1. Presentation of options.

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?
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?

Fig. 2. Confirmation Strategies.

3. THE CORPUS

3.1 The Wizard-of-Oz Experiment

The interactions we analyze here were collected in the context of a laboratory experiment [Wolters et al. 2009]. In this experiment, we tested two approaches to accommodate users with low working memory span.

—*Reduce Number of Options.* If users are presented with a large number of options at each step in the appointment scheduling dialog, they are less likely to use the dialog system successfully. Users were presented with one, two, or four options at a time. For examples, see Figure 1.
—*Provide Confirmations.* If the system confirms each aspect of the appointment, users will find it easier to remember the appointment, since relevant information is repeated. Confirmation was given explicitly, implicitly, or not at all. For examples, see Figure 2.

The combination of these 3×3 design choices yielded nine different dialog systems. We expected that users with lower Working Memory Span (WMS) would benefit more from reduced numbers of options and repeated confirmations than users with higher WMS.

Each of the nine systems was simulated using a Wizard-of-Oz (WoZ) design [Dahlbaeck et al. 1993]. In a WoZ setup, users interact with a human wizard but they think they are interacting with an automated dialog system. WoZ experiments are an invaluable tool for investigating different design options for spoken dialog systems without the cost of actually implementing these systems. Furthermore, they allow experimenters to isolate the effects of high-level information presentation and dialog management from the problems introduced by the limitations of current ASR and NLU systems.

3.1.1 *Participants.* 50 participants were recruited: 26 older participants and 24 younger participants. The older users' age ranged from 52 to 84 years (M=66, SD=9.1). The younger users were aged between 18 to 29 years (M=22, SD=2.7). 62% of the older users and 71% of the younger users were female. Older users had spent an average of 15 years in formal education, younger users an average of 17 years.

All participants took part in a battery of four cognitive tests. These tests were carefully chosen to cover the two main dimensions of intelligence: *fluid intelligence*, which is linked to abstract reasoning, and *crystallized intelligence*, which is linked to acquired knowledge, as well as *working memory* and *information processing speed*. All tests were presented visually, to avoid problems due to age-related hearing loss [Rabbitt 1990]. The full battery took 60 to 90 minutes to administer.

Fluid intelligence was assessed using Ravens' Progressive Matrices [Raven et al. 1998, Ravens]. Participants were not timed. Crystallized intelligence was measured using the Mill Hill Vocabulary test [Raven et al. 1998, MillHill]. Working memory span was assessed using a sentence reading span task [Unsworth and Engle 2005, SentSpan]. Reading span was chosen because this measure requires participants to process the stimuli instead of simply remembering them. Thus, it taps into the key function of working memory as a short-term store for information processing. In this article, we report the absolute score, which aggregates participants' scores for all test items. Information processing speed was assessed using the Digit Symbol Substitution subtest of the Wechsler Adult Intelligence Scale-Revised [Wechsler 1981, DSST].

3.1.2 *The Wizard-of-Oz System.* All dialogs 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 time slot on that half-day was agreed. All dialogs were strictly system-initiative: The WoZ system not only controlled the choice of options presented to the user at each stage of the dialog, 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 dialog.

3.1.3 *Procedure.* Appointments could be made with four different health care professionals: a community nurse, a diabetes nurse, an occupational therapist, and a physiotherapist. Participants were asked to book nine appointments in total. Each appointment was booked with a different prototype dialog system. Four randomized lists with nine tasks each were created in which each health professional appeared at least twice. Each user was randomly assigned to one of the lists. After an appointment had been booked, participants were asked to fill in a detailed user satisfaction questionnaire with 39 items which are listed in the relevant result tables (Table XV through XVI). The questionnaire was administered on paper. Completion took on average five minutes. Once the questionnaire had been completed, participants were asked to recall the appointment. In order to mitigate effects of tiredness, participants were able to take breaks at any time in between dialogs.

System:	Good morning, this is your appointment scheduling system. [*opening_closing, greeting*]
System:	Who would you like to see, the occupational therapist, the community nurse, the physiotherapist or the diabetes nurse? [*suggest_4, hp*]
User:	Physiotherapist [*accept_info, hp*] please. [*social_polite*]
System:	You wish to see the physiotherapist. Is that correct? [*confirm_yes, hp*]
User:	Yes. [*confirm_pos, hp*]
System:	When would you like to come, Monday afternoon, Tuesday afternoon, Wednesday afternoon or Thursday afternoon? [*suggest_4, halfday*]
User:	Monday afternoon [*accept_info, halfday*] please. [*social_polite*]
System:	You would like to make an appointment on Monday afternoon. Is that correct? [*confirm_yes, halfday*]
User:	Yes. [*confirm_pos, halfday*]
System:	At what time would you like to book an appointment, 2:00 p.m., 2:30 p.m., 3:00 p.m. or 4:00 p.m.? [*suggest_4, slot*]
User:	Two p.m. [*accept_info, slot*] please. [*social_polite*]
System:	You would like to book an appointment at 2:00 p.m. Is that correct? [*confirm_yes, slot*]
User:	Yes. [*confirm_pos, slot*]
System:	I have now booked for you an appointment with the physiotherapist on Monday, at two in the afternoon at the Community Treatment Centre. Is that correct? [*confirm, final_booking*]
User:	Yes. [*confirm_pos, final_booking*]
System:	Thank you very much for using the appointment scheduling system. Goodbye. [*opening_closing, goodbye*]

Fig. 3. Dialog with a younger user (4 options at a time, explicit confirmation), labeled with [*speech act, task*] pairs. For *social* speech acts, the task is not defined.

3.1.4 *Scoring*. We measured two aspects of task success: *completion* and *recall*. Completion measures whether users successfully arranged an appointment with the correct health professional that fits their schedule, while *recall* assesses whether users were able to remember the appointment they scheduled. Users received points for remembering location, weekday, morning/afternoon, hour, half-hour, class of health professional (nurse versus therapist), and exact name of health professional (community nurse, diabetes nurse, physiotherapist, occupational therapist). Both completion and recall are equally important: If a user manages to arrange an appointment but fails to attend because crucial details have been forgotten, the appointment has essentially not been scheduled successfully. We also assessed the effect of the nine different dialog strategies on efficiency and user satisfaction. Contrary to expectations, we found no differences in task success. Users were more efficient when more options were presented at a time and explicit confirmations were omitted. Older users tended to be less satisfied with the dialog systems [Wolters et al. 2009].

3.2 Corpus Construction

Each dialog consists of a sequence of user and system turns. Each system turn corresponds to a system message. User turns are coherent sequences of one or

System:	Good morning, this is your appointment scheduling system. [*opening_closing, greeting*]
System:	Would you like to see the occupational therapist, the community nurse, the physiotherapist or the diabetes nurse? [*suggest_4, hp*]
User:	Good morning. [*social_greeting*] Occupational therapist, [*accept_info, hp*] please. [*social_polite*]
System:	You wish to see the occupational therapist. Is that correct? [*confirm_hp_yes, hp*]
User:	That is correct. [*confirm_pos, hp*] Occupational therapist, [*repeat_info, hp*] thank you. [*social_polite*]
System:	When would you like to come, Monday afternoon, Tuesday morning, Friday morning or Friday afternoon? [*suggest_4, halfday*]
User:	Friday morning would be grand, [*accept_info, halfday*] but not at nine o'clock. [*provideblock_info, slot*] Um I'll be free after nine thirty. [*provide_info, slot*]
System:	You would like to make an appointment on Friday morning. Is that correct? [*confirm_yes, halfday*]
User:	That is correct, [*confirm_pos, halfday*] with the occupational therapist, [*reprovide_info_overall, hp*] Friday morning [*repeat_info, halfday*] but has to be after nine thirty. [*reprovide_info_overall_notfilled, slot*]

Fig. 4. Dialog with an older user (4 options at a time, explicit confirmation), labeled with [*speech act, task*] pairs. For *social* speech acts, the task is not defined.

more utterances produced by the user. Each user utterance may consist of one or more speech acts. User and system turns may overlap.

For 47 users, the full set of 9 dialogs was recorded, and for the remaining 3 users, we have 8 dialogs each. All 447 dialogs 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, [Barras et al. 2000]). The transcriber followed the guidelines developed by the AMI project (http://www.amiproject.org) for the creation of the AMI meeting corpus [Carletta 2007]. Figure 3 shows a typical interaction between a younger user and one of our dialog systems, with the system that always gives explicit confirmations and presents four options at a time. Figure 4 presents an excerpt from an interaction between the same system and an older user.

All transcriptions and annotations are stored in NXT format [Carletta et al. 2003]. Orthographic transcriptions are linked to the corresponding audio files. Information about users' scores on the cognitive tests, the agreed appointment, the recalled appointment, and user satisfaction ratings are also stored in the NXT representation of each interaction.

In addition to orthographic transcriptions, the corpus was annotated with dialog acts, because it is intended as a resource for research into dialog management, in particular for learning dialog policies [Georgila et al. 2008]. An initial automatic annotation was subsequently hand-corrected [Georgila et al. 2008].

Speech acts were defined on the basis of the annotation scheme presented in Georgila et al. [2008]. Speech acts specify the action that is performed by a sequence of words [Searle 1969]. For example, in Figure 4, the utterance "That is correct." confirms that the user intends to see the occupational therapist (speech act `confirm_pos`), while the user's first utterance, "Good morning", is a greeting and therefore belongs to the general class of `social` speech acts. For an overview of all user speech acts, see Table I.

3.3 Interaction Style Measures

We characterized the interaction style of each user using three groups of measures:

—*Dialog Level Measures.* These include dialog length in terms of turns, speech acts, and word forms, the number of different speech acts and word forms, and the frequency of actions that are of theoretical interest, such as taking the initiative, grounding information that has been received, and confirming system suggestions. The full set of features is defined in Table II.

—*Speech Act Measures.* These include frequencies of groups of speech acts that are defined in Table III. Speech act groups consist of a number of related, distinct speech acts. For example, the group `provide` consists of the two speech acts in Table I that begin with `provide_`, while the group `accept` consists of six speech acts that begin with `accept_`. The frequency of a speech act group is defined as the sum of the number of occurrences of each speech act in all dialogs produced by the user. For example, the older user `O1` uses 46 `provide` speech acts in 9 dialogs, while the older user `O11` only produces 6 instances of such speech acts.

—*Word Group Measures.* These include frequencies of words that fall into particular semantic/pragmatic groups, such as social words ("hello", "thanks"), or words used to talk about the interaction ("voice", "understand"). Word groups are defined by a list of associated word forms. For example, `goodbye` is associated with the word forms "cheerio", "bye", and "goodbye". Each occurrence is counted separately. Thus, the utterance "That's fine, thanks, cheerio, goodbye." contains two instances of `goodbye`. `O1` produces 7 words that are associated with the group `sorry` in 9 dialogs, while `O11` produces 0 instances of such word forms. A full list of all word groups can be found in Table IV.

Word groups were defined with a view to ASR and language modeling. In particular, we wanted to know how often people used words that were not directly relevant to the task, words that are politeness markers, synonyms for straightforward yes/no answers, or words that indicated meta-communication about the dialog. Such material can be difficult for a speech recognizer to process. Compare the two dialogs in Figure 5. The first dialog is no problem for a standard language model. The user answers the simple yes/no question with one of the two expected keywords, "yes". Such single-word utterances, especially if they consist of frequent responses such as digits or yes/no, can be detected extremely reliably. Moreover, the utterance contains no extra material.

Table I. List of User Speech Acts

Speech Act	Description
Accepting / Rejecting System Suggestions	
accept_info	user 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 accepts option that s/he previously provided
accept_info_yes_prevprovided	user accepts option s/he previously provided by saying "yes"
accept_info_null_prevprovided	user implicitly accepts option s/he previously provided
reject_info	user rejects option suggested by the system
reject_info_no	user rejects option suggested by the system by saying "no"
reject_info_null	user rejects option suggested by the system
confirm_pos	user confirms an option when asked for confirmation
confirmimplicit_pos	user continues w/ dial. after implicit confirmation by 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 system information
correct_info_no	user corrects system information using a negative
correctblock_info	user corrects previously provided info 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 reprovides information in the same utterance or turn
reprovide_info_overall	user reprovides information for slot that has already been filled
reprovide_info_overall_notfilled	user reprovides info for slot that has not been filled yet
reprovideblock_info	user reprovides info about options that are not possible
reprovideblock_info_overall	user reprovides info for slot already marked unavailable
repeat_info	user repeats info given by system in explicit/implicit confirmation
repeatblock_info	user repeats info 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 system
social	social interaction with the system, e.g. "goodbye", "thank you"
stall_wizard	user asks the wizard to wait

In the second dialog, the user uses a four-word phrase to signal an affirmative answer and proceeds to engage in meta-communication and provide two pieces of additional information, one of which is not relevant to the task at hand (appointment scheduling). If the system only expects a yes/no answer, it will

Table II. Dialog-Level Measures Used for Analysis

Name	Description
# Turns	number of user turns in the dialog
# Word Forms	total number of distinct word forms
# Words	total number of words
Avg. Word Form Freq.	average number of occurrences of each word form
# Distinct Speech Acts	total number of distinct speech acts
# Total Speech Acts	total number of speech acts
# Confirmations	number of confirmations
# Grounding	number of times information has been grounded
# Init	number of times user has taken initiative

Table III. Speech Act Groups Used for Analysis

Name	Description
accept	user accepts system suggestion
provide	user provides additional information
reprovide	user provides information again
grounding	user provides information again, grounding a filled slot
block	user gives information about options that do not suit
repeat	user repeats information
request	user requests clarification, repetition, or help
confirm	user reacts to system suggestion
social	greetings, apologies, and other social interaction
garbage	uninterpretable

Table IV. Word Groups Used for Analysis

Name	# word forms	Description
hes	7	hesitations, filled pauses
no	2	forms of “no”
yes	4	forms of “yes”
pos	11	forms of affirmative answers other than “yes”
neg	4	forms of negative answers other than “no”
hello	2	forms of “hello”
bye	4	word forms associated with “goodbye”
please	1	forms of “please”
thanks	3	word forms associated with thanking
sorry	2	apologies
modal	7	forms of modal verbs such as “can”
qualifiers	16	qualifiers of information such as “most”
request	14	forms of requests
meta	67	communication about the dialog

be impossible to understand what the user is saying. If the system has access to a richer grammar of affirmative and negative responses, it might be able to map “that’s great” to a “yes” answer. But in order to answer the user fully, the ASR component needs to be able to recognize at least key words such as “Tuesdays”, “Wednesdays”, and “understand”, the NLU engine would need to detect the two subsequent speech acts and extract the relevant information, and the DM component would have to select a reaction that indicates whether the system can understand the user.

Easy to parse:
System: Would you like to see the occupational therapist?
User: Yes.

Difficult to parse:
System: Would you like to see the occupational therapist?
User: Oh that would be great, I have a really bad back. I can only make Tuesdays or Wednesdays though. Just wondering, can you understand what I'm saying?

Fig. 5. Two types of positive answers.

4. METHOD

In this study, we examine whether our users can be divided into groups based on their interaction style. Each user is represented by a data point that is specified by the frequency of a number of features as defined in Section 3. Underlying structures in datasets can be found using clustering methods. Clustering is used to partition a given dataset into groups such that each item in the original dataset belongs to exactly one group, and items in each group are more similar to each other than to items in other groups. There are many techniques for finding groups in data that differ both in the way the clusters are created and in the way in which the quality of the resulting clusters is judged [Webb 2002]. In this study, we used seven classic methods that have complementary strengths and weaknesses.

4.1 Cluster Analysis

The groups that are found through cluster analysis can be affected by many factors: presence of outliers, normalization of data, distance metric used, presence of irrelevant features, the clustering algorithm itself, and the metric for determining cluster quality [Webb 2002; Tan et al. 2005]. Quality metric, relevant features, and outliers were determined once at the beginning of the analysis. The remaining aspects, data normalization, distance metrics, and clustering algorithm, were systematically varied using the R package `ClusterSim` [Walesiak 2008]. In total, we examined 560 combinations of parameters per feature set.

4.1.1 *Quality Metric.* As our quality metric, we chose Rousseeuw's Silhouette measure [Rousseeuw 1987]. This metric favors compact clusters that are well separated. Let $a(i)$ be the average dissimilarity between i and all other items in A, and $b(i)$ the minimum dissimilarity between i and the nearest member of the cluster B that is closest to A. Then the silhouette of a data item i from cluster A is defined as

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}. \tag{1}$$

The silhouette of a cluster A is the average silhouette of all data points $i \in A$, and the silhouette of a given set of clusters C is the maximum of the silhouettes of each cluster $A \in C$. As a rough guide, overall silhouettes of 0.71 and higher indicate that a very reliable structure has been found, while values between 0.50 and 0.70 indicate that a reasonable clustering has been detected [Rousseeuw 1987].

4.1.2 *Outliers.* Outliers in the dataset can skew the clustering such that we end up with several tiny clusters that consist of the outliers, and one big cluster that contains the rest of the dataset. Such an analysis completely obscures any more detailed structure in the main dataset. Therefore, as a first step, outliers need to be detected and removed before proceeding with analysis.

4.1.3 *Feature Sets.* Each user is represented by one feature vector. All feature values except for the average frequency of word forms and distinct speech acts are counts. These counts represent the total number of occurrences of a feature in all dialogs recorded for a given user.

We tested four feature sets.

—*Dialog.* Measures of conversation length and frequencies of characteristic actions as given in Table II.

—*Speech Acts.* Number of occurrences of speech act groups as given in Table III.

—*Words.* Number of occurrences of word groups as given in Table IV.

—*All.* Combination of all three feature sets.

These feature sets were derived from the features reported in Georgila et al. [2008] through a combination of a priori considerations and inspection of the feature value distribution. Word and speech act groups that only occurred in very few of our 447 dialogs were excluded. Features that only occurred in dialogs from a small number of users were included if they were of theoretical interest.

The frequency distribution of many features is highly skewed. Take, for example, the number of times when users provided information about times and days that they cannot make (*blocking*). We only find blocking in dialogs from 7 users, all of whom are older. Five of these users block less than 3 times in all 9 dialogs, while one user provides 52% of all instances of blocking.

4.1.4 *Normalization.* Five normalization procedures were tested, standardization (Eq. (2)), Weber standardization (Eq. (3)), unitization (Eq. (4)), unitization with zero minimum (Eq. (5)), and normalization with a range of [-1,1] (Eq. (6)).

$$x' = (x - \text{mean}(x))/\text{stddev}(x) \tag{2}$$

$$x' = (x - \text{median}(x))/\text{median absolute dev.}(x) \tag{3}$$

$$x' = (x - \text{mean}(x))/\text{range}(x) \tag{4}$$

$$x' = (x - \min(x))/\text{range}(x) \tag{5}$$

$$x' = (x - \text{mean}(x))/\max(|x - \text{mean}(x)|) \tag{6}$$

4.1.5 *Distance Metric.* The distance metric defines the extent to which two items i and j are similar. Many of our features do not occur in at least 20% of our users, which means that they have an absolute frequency of 0 for that user. This can greatly skew distance measures. Therefore, we transformed the $n \times m$ matrix of feature values, where n is the number of features and m the number of users, into the corresponding covariance matrix. In our experiments, we compared five distance metrics: Euclidean distance, squared Euclidean distance, Manhattan distance, Chebyschev distance, and the General Distance Measure [Walesiak 1999; Jajuga et al. 2003]. The first four distances can be found in standard textbooks [Webb 2002, Appendix A]. The General Distance Measure was designed to accommodate ratio-scale, interval-scale, and ordinal data. Here, we used GDM1, the version for ratio and interval-scale data. Let x_{ij} be the value of feature j for object i. Then the GDM distance between two objects i,k is given as

$$\mathrm{GDM1}(i,k) = \frac{1}{2} - \frac{\sum_{j=1}^{m}(x_{ij}-x_{kj})(x_{kj}-x_{ij}) + \sum_{j=1}^{m}\sum_{l=1,l\neq i,k}^{n}(x_{ij}-x_{lj})(x_{kj}-x_{lj}))}{2\left[\sum_{j=1}^{m}\sum_{l=1}^{n}(x_{ij}-x_{lj})^2 \sum_{j=1}^{m}\sum_{l=1}^{n}(x_{kj}-x_{lj})^2\right]^{1/2}}. \tag{7}$$

4.1.6 *Clustering Method.* There are many different clustering methods. In this section, we will only discuss the methods used in our study as implemented in the `ClusterSim` package; for a more detailed overview, see, for example, Kaufman and Rousseeuw [1990] and Webb [2002].

We tested seven methods: six hierarchical methods and one partitioning-based method. The hierarchical methods were single linking, complete linking, average linking, median linking, centroid linking, and the Ward method as implemented in the R method `hclust` [R Development Core Team 2008]; the partitioning-based method was Kaufman and Rousseeouw's 1990 Partitioning Around Medoids (PAM) as implemented in the R method `pam` [R Development Core Team 2008]. All of these methods are well known and widely used.

All six hierarchical methods are agglomerative. They start with a partitioning of the dataset where each data point is assigned its own cluster. In each step, the two nearest clusters are merged until the data is represented by one single cluster. The clusters can be represented as a binary tree where each node is associated with a cluster. Each internal node n has exactly two daughters d_1, d_2. When we join the clusters which are associated with d_1, d_2, we get the cluster associated with n. In order to partition a dataset into x clusters, we simply cut the tree at the appropriate height. Figure 6 shows the dendrogram for a sample clustering of our dataset.

The six hierarchical methods differ with respect to the rule used to determine the distance between clusters. For *single linking*, the distance between two clusters A, B is taken to be the distance between the two nearest data points.

$$d_{AB} = \min_{i\in A,\, j\in B} d_{ij} \tag{8}$$

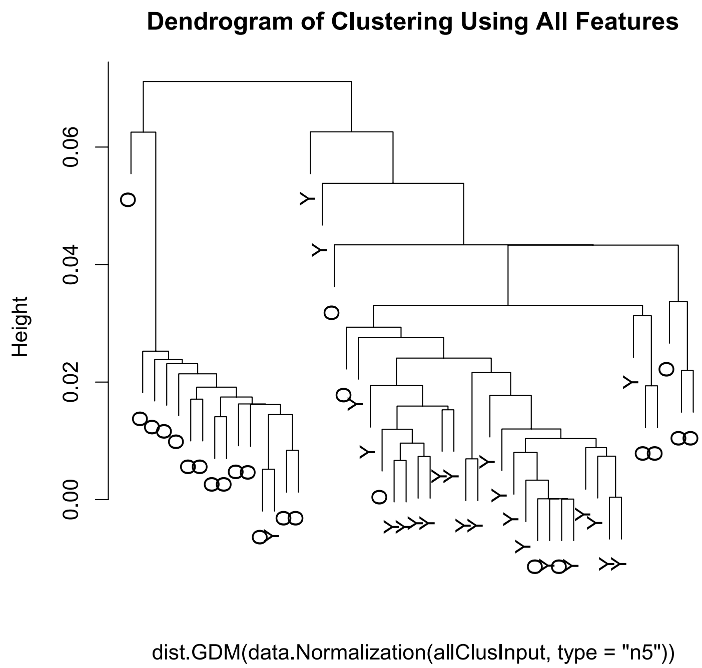

Fig. 6. Dendrogram of clustering for all users excluding outliers, feature set `All`. Left daughter of root corresponds to cluster 1, right daughter to cluster 2. O: older users, Y: younger user.

Single linking favors elongated clusters; it is less well suited to detecting compact, globular structures [Tan et al. 2005]. In *complete linking*, the distance between the two farthest data points is used instead.

$$d_{AB} = \max_{i \in A, j \in B} d_{ij} \tag{9}$$

Average linking uses the average distance between data items, while *centroid* linking is based on the distance between the cluster means, and *median* linking determines the distance between clusters by calculating the distance between cluster medians. Complete, average, centroid, and median linking favor compact, globular clusters. Median linking is useful if some of the clusters have a small diameter compared to others [Webb 2002]. *Ward* clustering uses a slightly different approach. The core measure is the sum of squares of dissimilarities between data points. The lower the sum of squares, the more similar the cluster items. In each step, the algorithm merges those two clusters that yield the new cluster with the smallest sum of squares. Thus, the Ward algorithm also favors very compact clusters.

Partitioning around medoids (PAM) [Kaufman and Rousseeuw 1990] follows a different strategy. Whereas the hierarchical approaches discussed earlier yield a sequence of possible clusters, in PAM, the desired number of clusters k is one of the inputs into the algorithm. Briefly, for each cluster, the algorithm systematically searches for one data item that represents the center of the cluster, the "medoid". Data points are assigned to the cluster that is associated

Table V. Best Combinations for Each Dataset

Feature Set	Method	Distance Measure	Normalisation	Silhouette
All	Single Linking	GDM 1	Normalisation to [-1,1]	0.713
Word	Single Linking	GDM 1	Normalisation to [-1,1]	0.681
Speech Acts	Complete Linking	GDM 1	Normalisation to [-1,1]	0.606
Dialogue	Complete Linking	GDM 1	Weber standardization	0.798

with the nearest medoid. PAM is a variant of classical k-means approaches to clustering. From a design perspective, it is useful for defining "typical" users.

5. RESULTS

5.1 Interaction Style Groups

Before running the clustering algorithms, the dataset was examined for outliers using Principal Component Analysis (PCA, R method `prcomp` with scaling and centering of feature values [R Development Core Team 2008]). All 34 features present in the full dataset (`All`) were used in the analysis. The first two principal components covered 61.9% of the overall variance. When plotting the dataset against the first two principal components, two outliers were identified and removed from the dataset for all subsequent cluster analyzes. Both of these outliers were older users. This leaves us with a dataset consisting of 48 users, leaving us with 24 older and 24 younger users.

In the next step, we determined optimal clustering methods for each of the four feature sets, `All`, `Dialog`, `SpeechActs`, and `Words`, using `ClusterSim`. The best combination for each dataset is shown in Table V. For an explanation of the relevant parameters, see Section 4. The preferred normalizations are Weber standardization, which relies on medians, and normalization to [-1, 1]. The preferred distance measure was always the General Distance Measure. The silhouette values indicate that the clustering algorithms found a reasonable structure. It is interesting that the best methods alternate between single and complete linking, since the first favors long, straggly clusters and the second compact, globular clusters. The explanation lies in our dataset: As Figure 7 shows, one of the two clusters is very compact, an ideal candidate for complete linking, while the other is long, extended, and fuzzy, the type of cluster that single linking detects very well.

In order to determine the number of clusters, we examined the ten best solutions for each of the four feature sets. 75% of the best solutions are for two clusters. The remaining solutions are five-cluster solutions when clustering on dialog-level data only. This strongly suggests that there are indeed two groups of users in the data. Therefore, we chose two-cluster solutions for all four feature sets.

In the next step, we identified the users that belonged to each of the two clusters. First, the dataset was clustered four times, once per feature set, using the best algorithm and settings for this particular feature set. We then compared the clusters that users were assigned to. There was considerable overlap: For 85% (41 out of 48), the four cluster analyzes agreed. Of these unambiguous

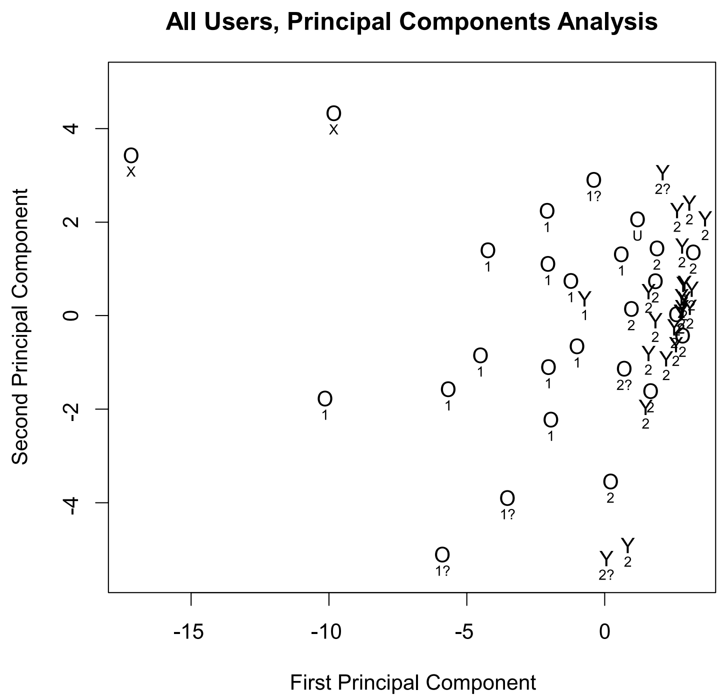

Fig. 7. Principal Component Analysis of all users. O: older, Y: younger, 1,2: Cluster, ?: assigned by majority vote, U: undecided, X: outlier.

cases, 19 were older users and 22 younger users. The first cluster contains 12 users, while the second group consists of 29 users. The high level of agreement indicates that our cluster analysis is uncovering a real structure in the data, not imposing a spurious order. Of the users that could not be assigned reliably, taking a majority vote assigns 3 to cluster 1, and 3 to cluster 2. Only one user cannot be assigned to a cluster by majority vote.

Figure 7 shows the distribution of all 50 users along the two first principal components identified by the PCA. The 6 users that were assigned to clusters by a majority vote are marked with a "?". Four of these users are at the boundary between clusters, while the two older users who are potentially in cluster 1, are at the outer boundary of this cluster. The single user for whom no decision can be made sits at the boundary between the two clusters. The two outliers, who are not marked with a cluster label, are older users who are firmly on the side of cluster 1. They may well be "extreme" users in the sense of Pullin and Newell [1997].

In order to establish the characteristics of these two clusters, let us now consider the 41 unambiguous cases. Table VI shows the distribution of age groups across clusters. While almost all of the users in cluster 1 are older (92%), a significant proportion of the older users are in cluster 2, which is dominated by younger users. The clusters cannot be completely segregated by age.

Table VI. Distribution of Age Groups Across Clusters

Age Group	Cluster 1	Cluster 2	Total
Older	11	8	19
Younger	1	21	22
Total	12	29	41

This is a key result: Older users differ greatly in how they interact with the simulated systems. Although cluster 1 is dominated by older people, with a lone younger straggler, 8 older people have been assigned to cluster 2 because they appear to behave very much like younger users. This number corresponds to two-fifths of the older people in our unambiguous sample of 41 cases and a third of our original sample of 26 users.

Figure 7 shows all users and their associated clusters. We see that cluster 2 is very compact and well defined, while cluster 1, which is mostly populated by older users, is highly diffuse, just as we would expect from the literature on aging (refer to Section 2.4.1).

In the following subsections, we analyze the behavior of the users in each cluster in detail. Once typical patterns of behavior have been identified (Section 5.1.1), we turn our attention to older users. In particular, we would like to know in what respect the older users in cluster 2 differ from the older people in cluster 1, which consists exclusively of older users (Section 5.1.2). We also examine whether there are any age-related differences in interaction style within cluster 2, which contains a mix of older and younger users (Section 5.1.3).

In order to get a clear picture of the characteristics of each cluster, the analyzes presented in the rest of this section are restricted to those 41 users that could be assigned unambiguously to one of the two clusters. Since cluster analysis is an exploratory technique, it is wise to be conservative when using its results for further analysis. However, as Figure 7 shows, most of the remaining users are very close to one of the two clusters, and the older users who have been marked as outliers are "extreme" examples of cluster 1. Indeed, when repeating our analyzes with all users, classifying the two "extreme" users and the undecided user into cluster 1, we obtained very similar results to the ones reported here.

5.1.1 *Cluster Characteristics.* The differences between the two clusters are illustrated in Table VII. Users in cluster 1 need more turns to complete the whole set of nine dialogs. They also produce roughly three times as many individual words and twice as many individual speech acts as users in cluster 2. During their interaction, they use a richer vocabulary, and a more varied repertoire of speech acts. Despite this, the token/type ratio is similar between clusters. Users in cluster 1 are also more likely to use confirmations, ground information, and take the initiative in providing additional information.

The additional speech acts seen in cluster 1 users can be classified into four groups, next described.

—*Managing Information.* Users from cluster 1 are significantly more likely to provide additional information about aspects of the appointment. They do this around 15 times during the whole set of 9 dialogs, while users from

Table VII. Frequency of Dialog-Level, Speech-Act-Level, and Word-Level Features by Cluster.

Group	Measure	Cluster 1	Cluster 2	Sig.
Dialog	Turns	78.75	61.14	p < 0.000 ***
	Words	349.67	98.66	p < 0.000 ***
	Word Forms	99.17	29.69	p < 0.000 ***
	Avg. Word Form Freq.	3.47	3.38	p < 0.931 n.s.
	Distinct Speech Acts	134.17	73.07	p < 0.000 ***
	Total Speech Acts	16.08	8.76	p < 0.000 ***
	Confirmations	31.75	29.52	p < 0.011 .
	Grounding	33.00	29.52	p < 0.002 **
	Initiative	18.50	1.69	p < 0.000 ***
Speech Act Groups	Provide	15.25	1.62	p < 0.000 ***
	Reprovide	3.17	0.07	p < 0.000 ***
	Block	1.33	0.00	p < 0.001 **
	Request	1.00	0.21	p < 0.125 n.s.
	Repeat	4.25	0.14	p < 0.000 ***
	Social	33.50	4.21	p < 0.000 ***
	Garbage	5.92	0.38	p < 0.000 ***
Word Groups	Yes	16.33	25.66	p < 0.001 **
	No	2.25	6.38	p < 0.001 **
	Positive	9.25	1.69	p < 0.000 ***
	Negative	1.50	0.90	p < 0.169 n.s.
	Thanks	9.42	0.48	p < 0.000 ***
	Bye	4.67	0.24	p < 0.000 ***
	Please	14.58	3.21	p < 0.000 ***
	Hello	0.25	0.10	p < 0.13 n.s.
	Sorry	0.67	0.17	p < 0.159 n.s.
	Meta	5.25	0.24	p < 0.000 ***
	Extra	1.33	0.03	p < 0.034 .
	Modal	7.58	0.14	p < 0.000 ***
	Qual	3.58	0.00	p < 0.000 ***

.: p<0.05, *: p<0.01, **: p<0.001, ***: p<0.0001 or better.

cluster 2 only give unprompted information once or twice during all 9 interactions. Cluster 1 users are also more likely to ground slots that have already been filled by the system.

—*Being Sociable.* Users from cluster 1 are more likely to interact socially with the system. They produce an average of around 33 `social` speech acts during all 9 dialogs (more than 3 per dialog), whereas cluster 2 users only produce 5 speech acts of this type in 9 dialogs (less than 1 per dialog).

—*Repeating Information.* Users from cluster 1 are more likely to to repeat information that comes from the system (`repeat`) and from themselves (`reprovide`). While cluster 1 users do this on average seven times during their interactions with the nine SDS, cluster 2 users almost never repeat information.

—*Garbage.* Users from cluster 1 are more likely to produce uninterpretable utterances. On average, six utterances from a typical cluster 1 user are classified as `garbage`, compared to less than one utterance from a typical cluster 2 user.

This suggests that, overall, cluster 1 users are more likely to actively manage the interaction as they would try to manage the interaction with a human receptionist. They provide information that they expect the system to process, even though all dialogs are system-initative and system questions are specifically designed to elicit brief, factual responses. Cluster 1 users are also more likely to repeat information and to produce problematic utterances.

As we can see in Table VII, the additional vocabulary that characterizes cluster 1 falls into four distinct groups:

(1) affirmative answers couched in terms other than forms of "yes" (category `positive`);
(2) social expressions such as "thanks", "bye", and "please" (categories `thanks, bye, please`);
(3) communication about the discourse, about the user's internal reasoning processes, and about reasons for the users' decisions (category `meta`); and
(4) modal expressions and qualifiers (categories `modal, qualifier`).

Cluster 1 users are also far more likely to use nonwords such as backchannels or filled pauses.

Most of the modals and qualifiers are part of polite idiomatic expressions signaling levels of preference, such as "I would like to" or "I can make" or "I would rather". This is in addition to variants of "no" and "yes". Cluster 1 users are also more likely to follow social conventions in bidding "good-bye" to the system, adding "please", or thanking the system. However, they are as unlikely as cluster 2 users to apologize or greet the system with "hello".

To summarize, users who belong to cluster 1 have an interaction style that can be characterized as sociable, chatty, and communicative. In contrast, users from cluster 2 are more terse and dispense with many of the social niceties that would be expected in conversation with a human. In the following discussion, we will characterize these two tendencies with the keywords "social" (for cluster 1) and "factual" (for cluster 2). These keywords are solely intended as descriptors of user behavior; they should not be taken to allude to stereotypical views of older users.

5.1.2 *Social Older Users versus Factual Older Users.* Table VIII presents the differences between older users in the Social and older users in the Factual cluster. Overall, the differences between the two groups of older users mirror the differences between the two clusters. However, there are a few commonalities across clusters. Older users from both clusters are less likely to use "yes", and more likely to use alternatives to "yes" (category `pos`) for affirmative answers. They are also equally likely to use grounding speech acts, and to repeat information. In addition, they share a tendency to use words associated with talking about the dialog (category `meta`).

5.1.3 *Age-Group Differences Between Factual Users.* Table IX shows that the behavior of older "factual" users barely differs from that of younger "factual" users. The only significant difference is the number of affirmative expressions other than "yes" (`pos`). Weakly significant differences include the amount of metacommunication, where "factual" older users behave like their more

Table VIII. Frequency of Features by Cluster (Older Users Only)

Group	Measure	Social	Factual	Sig.
Dialog	Turns	80.36	66.75	p < 0.013 .
	Words	358.64	115.00	p < 0.000 ***
	Word Forms	101.73	35.88	p < 0.000 ***
	Avg. Word Form Freq.	3.47	3.27	p < 0.869 n.s.
	Distinct Speech Acts	136.82	79.25	p < 0.000 ***
	Total Speech Acts	16.45	9.88	p < 0.001 **
	Confirmations	31.91	30.12	p < 0.17 n.s.
	Grounding	33.09	30.12	p < 0.089 n.s.
	Initiative	19.00	1.38	p < 0.000 ***
Speech Act Groups	Provide	15.45	1.25	p < 0.000 ***
	Reprovide	3.27	0.12	p < 0.001 **
	Block	1.45	0.00	p < 0.064 n.s.
	Request	4.64	0.00	p < 0.016 .
	Repeat	1.09	0.25	p < 0.481 n.s.
	Social	33.91	5.50	p < 0.000 ***
	Garbage	6.09	0.88	p < 0.002 **
Word Groups	Yes	16.36	23.25	p < 0.117 n.s.
	No	2.27	6.00	p < 0.013 .
	Positive	9.45	5.62	p < 0.135 n.s.
	Negative	1.64	0.75	p < 0.128 n.s.
	Thanks	9.64	1.25	p < 0.001 **
	Bye	4.64	0.62	p < 0.004 **
	Please	14.27	3.00	p < 0.001 **
	Hello	0.27	0.00	p < 0.117 n.s.
	Sorry	0.73	0.25	p < 0.422 n.s.
	Meta	5.64	0.75	p < 0.066 n.s.
	Extra	1.45	0.12	p < 0.416 n.s.
	Modal	7.73	0.38	p < 0.001 **
	Qual	3.82	0.00	p < 0.003 **

.: p<0.05, *: p<0.01, **: p<0.001, ***: p<0.0001 or better.

"social" counterparts, the number of turns, which is higher for older users, and the number of distinct speech acts.

5.2 Users Adapting to the System

Another interesting aspect of interaction style is the extent to which users adapt to the system. As we have seen in Section 2, people tend to change how they speak when they interact with another person, and this accommodation may even extend to artificial interlocutors. It is difficult to assess whether our participants changed their behavior in response to the system, since they were presented with nine system variants and only had one conversation with each of these. However, the systems were very similar in several key aspects.

—*Discouraging User Initiative and Overanswering.* The dialogs were tightly scripted. Users were required to react to options suggested by the system. By default, options were generated randomly. Information that the system had not explicitly asked for was not taken into account. This is typical of system-initiative SDS and a very common way of making systems more robust. Moreover, the "wizard" was unable to overtly acknowledge or repeat

Table IX. Frequency of Features by Age Group (Factual Users Only)

Group	Measure	Older	Younger	Sig.
Dialog	Turns	66.75	59.00	p < 0.04 .
	Words	115.00	92.43	p < 0.087 n.s.
	Word Forms	35.88	27.33	p < 0.059 n.s.
	Avg. Word Form Freq.	3.27	3.42	p < 0.494 n.s.
	Distinct Speech Acts	79.25	70.71	p < 0.142 n.s.
	Total Speech Acts	9.88	8.33	p < 0.041 .
	Confirmations	30.12	29.29	p < 0.489 n.s.
	Grounding	30.12	29.29	p < 0.489 n.s.
	Initiative	1.38	1.81	p < 0.781 n.s.
Speech Act Groups	Provide	1.25	1.76	p < 0.761 n.s.
	Reprovide	0.12	0.05	p < 0.47 n.s.
	Block	0.00	0.00	n/a
	Request	0.00	0.19	p < 0.537 n.s.
	Repeat	0.25	0.19	p < 0.328 n.s.
	Social	5.50	3.71	p < 0.302 n.s.
	Garbage	0.88	0.19	p < 0.161 n.s.
Word Groups	Yes	23.25	26.57	p < 0.221 n.s.
	No	6.00	6.52	p < 0.695 n.s.
	Positive	5.62	0.19	p < 0.003 **
	Negative	0.75	0.95	p < 0.875 n.s.
	Thanks	1.25	0.19	p < 0.129 n.s.
	Bye	0.62	0.10	p < 0.058 n.s.
	Please	3.00	3.29	p < 0.539 n.s.
	Hello	0.00	0.14	p < 0.374 n.s.
	Sorry	0.25	0.14	p < 0.502 n.s.
	Meta	0.75	0.05	p < 0.02 .
	Extra	0.12	0.00	p < 0.105 n.s.
	Modal	0.38	0.05	p < 0.106 n.s.

.: p<0.05, *: p<0.01, **: p<0.001, ***: p<0.0001 or better.

back additional information offered by the user, because there were no message templates for such utterances. Thus, users should provide less additional information in the later dialogs than in the earlier ones.

—*Factual System Prompts.* All prompts were designed to be neutral, clear, and concise. In addition, the "wizard" was unable to react to social chitchat. Thus, users should be less likely to use social dialog acts and words associated with interpersonal interaction in the latter dialogs.

We identified six markers that are relevant to these two points and occur reasonably frequently in the data: the frequency of `provide_*` and `reprovide_*` speech acts (taking initiative and overanswering), the frequency of social speech acts, and the respective frequencies of the three word groups "thanks", "bye", and "please". The number of occurrences of each of these markers in the first three dialogs was compared to their frequency in the last three dialogs. Since the sequence of dialog systems was randomized for each user, the distribution of conditions in the first three dialogs and in the last three dialogs is the same (Wilcoxon test, $p < 0.9$). Hence, it is feasible to statistically compare aggregate data from these two groups of dialogs.

Figure 8 shows a sample dialog where the user both takes the initiative and includes many social elements. This particular exchange took place right after

System: Would you like to see the community nurse?
User: Yes *[acknowledgement]*
good morning, *[social]*
I'd like to see the community nurse *[accept_info]*
if at all possible, *[social]*
on # Wednesday morning *[provide_info]*
at nine thirty, *[provide_info]*
failing that on Friday *[provide_info]*
at ten o'clock. *[provide_info]*
Can you confirm, th- thank you. *[social]*

Fig. 8. Overanswering with social speech acts.

System: When would you like to come, Monday afternoon, Tuesday morning, Friday morning or Friday afternoon?
User: Friday morning would be grand, *[accept_info]*
but not at nine o'clock. *[provideblock_info]*
Um I'll be free after nine thirty. *[provide_info]*

Fig. 9. Overanswering with blocking.

Table X. Frequency of Key Features in First Three versus Last Three Dialogs by Cluster

Measure	Social Cluster			Factual Cluster		
	First 3	Last 3	Sig.	First 3	Last 3	Sig.
Social	12.58	11.50	p < 0.488 n.s.	3.41	0.83	p < 0.008 *
Provide	6.42	5.92	p < 0.5 n.s.	1.10	0.24	p < 0.001 **
Reprovide	1.42	0.83	p < 0.137 n.s.	0.07	0.00	p < 0.167 n.s.
Thanks	3.42	3.58	p < 0.637 n.s.	0.52	0.00	p < 0.011 .
Bye	2.00	1.08	p < 0.032 .	0.14	0.07	p < 0.095 n.s.
Please	4.58	4.67	p < 0.547 n.s.	2.38	0.72	p < 0.014 .

.: p<0.5, *: p<0.01, **: p<0.001, ***: p<0.0001 or better.

the start of the dialog, after the system greeting. We see three social speech acts: The user greets the system with “good morning”, qualifies a request with “if at all possible”, and asks for confirmation with “thanks”.

The user also provides four pieces of information (hence the four `provide_info` speech acts), namely day and time for two alternative appointments. Figure 9 comes from the next stage of the dialog, where the half-day is agreed. The user explicitly accepts one of the options, Friday morning, but then goes on to specify a time slot that is impossible (`provideblock_info`) and a time slot that is suitable (`provide_info`).

The results are summarized in Tables X through XII. The tables show mean differences in frequency between the first and the last dialogs. Table X shows changes in behavior by cluster. Users in the Social cluster only become less likely to sign off with “good-bye”. Users in the Factual cluster, on the other hand, change their behavior in the expected direction: less overanswering, fewer social speech acts, fewer social responses. When we unpick this tendency further, we notice a clear difference between older users in the two clusters (refer to Table XI): Older users in the Factual cluster adapt more aspects of

Table XI. Frequency of Key Features by Cluster (Older Users Only)

Measure	Older (Social Cluster)			Older (Factual Cluster)		
	First 3	Last 3	Sig.	First 3	Last 3	Sig.
Social	13.18	11.45	$p < 0.346$ n.s.	4.75	0.75	$p < 0.084$ n.s.
Provide	6.36	6.09	$p < 0.579$ n.s.	0.88	0.00	$p < 0.006$ *
Reprovide	1.55	0.82	$p < 0.089$ n.s.	0.00	0.00	$p < 1$ n.s.
Thanks	3.73	3.45	$p < 0.408$ n.s.	1.50	0.00	$p < 0.038$.
Bye	1.91	1.18	$p < 0.072$ n.s.	0.25	0.25	$p < 0.35$ n.s.
Please	4.73	4.45	$p < 0.382$ n.s.	2.38	0.38	$p < 0.046$.

.: $p<0.05$, *: $p<0.01$, **: $p<0.001$, ***: $p<0.0001$ or better.

Table XII. Frequency of Key Features by Age Group (Factual Users Only)

Measure	Older (Factual Cluster)			Younger (Factual Cluster)		
	First 3	Last 3	Sig.	First 3	Last 3	Sig.
Social	4.75	0.75	$p < 0.084$ n.s.	2.90	0.86	$p < 0.025$.
Provide	0.88	0.00	$p < 0.006$ *	1.19	0.33	$p < 0.014$.
Reprovide	0.00	0.00	$p < 1$ n.s.	0.10	0.00	$p < 0.17$ n.s.
Thanks	1.50	0.00	$p < 0.038$.	0.14	0.00	$p < 0.081$ n.s.
Bye	0.25	0.25	$p < 0.35$ n.s.	0.10	0.00	$p < 0.081$ n.s.
Please	2.38	0.38	$p < 0.046$.	2.38	0.86	$p < 0.063$ n.s.

.: $p<0.05$, *: $p<0.01$, **: $p<0.001$, ***: $p<0.0001$ or better.

their interaction style to the system than older users in the Social cluster. Within the Factual cluster, older users again behave much like younger users (Table XII).

5.3 Predicting Older Users' Interaction Style

Now that we have identified two very distinct patterns of interaction style, it would be useful to find a way of determining which group a given user will fall into. The obvious candidate is age. If we assign users to clusters by majority vote, then almost all younger users (96%) belong to the Factual cluster. So, for younger users, age group predicts interaction style fairly well. Older users, on the other hand, are split. 35% of the 26 older users are in the Factual cluster, either unambiguously (8 users) or by majority vote (1 user).

In order to attempt to predict the interaction style of an older user, we now look for significant differences in cognitive abilities and demographic variables. There is a slight difference in chronological age between older users in the two clusters (Social cluster: M=67, SD=9.5 years, Factual cluster: M=62, SD=5.8 years), but this is not statistically significant (Kruskal-Wallis test, $p<0.21$). There is also no significant difference in gender distribution ($p<0.37$) or years of education ($p<0.45$) between older users in the Factual cluster and those in the Social cluster. This leaves us with our last potential explanatory variable, cognition. As Table XIV demonstrates, none of the four aspects of cognition we tested can account for differences in interaction style, even though younger and older users in the Factual cluster differ significantly on all cognitive measures

Table XIII. Differences in Cognitive Abilities Between Age Groups (Factual Cluster)

Measure	Older		Younger		Sig.
	Mean	Stddev	Mean	Stddev	
Ravens	49.00	6.84	54.64	3.55	$p < 0.006$ *
MillHill	52.58	7.81	42.23	6.91	$p < 0.004$ **
DSST	52.11	11.68	75.59	8.57	$p < 0.000$ ***
SentSpan	31.22	16.57	37.09	15.59	$p < 0.981$ n.s.

.: p<0.05, *: p<0.01, **: p<0.001, ***: p<0.0001 or better.

Table XIV. Differences in Cognitive Abilities Between Clusters (Older Users Only)

Measure	Social		Factual		Sig.
	Mean	Stddev	Mean	Stddev	
Ravens	48.36	8.35	49.88	4.39	$p < 0.934$ n.s.
MillHill	53.36	9.67	51.50	4.57	$p < 0.214$ n.s.
DSST	51.45	14.22	53.00	7.75	$p < 0.869$ n.s.
SentSpan	28.40	16.86	34.75	16.61	$p < 0.286$ n.s.

except SentSpan (Table XIII).[1] We will discuss this finding in the context of the literature on cognitive aging and social cognition in Section 6.

5.4 Interaction Style and Usability

Now that differences in interaction style have been firmly established, the final question is: Do they affect usability? Since the two interaction styles require, at the very least, different language models, usability of a full SDS will definitely suffer if the social users cannot be accommodated, especially given that these users are the least likely to adapt their behavior to the system. Even in our WoZ experiments, where speech recognition and language understanding are near perfect, we can detect important differences in usability between the two groups.

First, we examine *task success*. Overall, users in the two clusters do not differ significantly in terms of task success (Kruskal-Wallis test, $p<0.441$ not significant). Looking at each cluster in turn, we found no significant effect of dialog strategy on performance. This fits well with our previous analyzes of the complete dataset [Wolters et al. 2009].

When it comes to *efficiency*, on the other hand, we find very clear differences between the clusters, as we have discussed earlier. "Social" users take significantly longer to reach their goal than "factual" users (refer to Table VII).

The clusters also differ in terms of *user satisfaction*. Table XV shows the main significant differences in user ratings between "social" and "factual" users. We examined differences in ratings for all 38 interval-scaled and ordinal items in the questionnaire, excluding the binary perceived task completion. "Social" users are invariably more negative than "factual" users. They notice

[1]Incidentally, despite the large age range, there is no correlation between chronological age and working memory span, Ravens scores, or the MillHill vocabulary test results in our population. We do find a significant correlation between information processing speed and age (Pearson's ρ=-0.502, $p < 0.009$).

Table XV. Differences in User Satisfaction Between Clusters

Feature	Social	Factual	Sig
Overall impression	3.46	3.69	$p < 0.076$ n.s.
Did not do what I wanted	2.67	2.24	$p < 0.000$ ***
System info clear	3.76	3.97	$p < 0.001$ **
System info incomplete	2.39	2.10	$p < 0.001$ **
System efficient	3.70	3.71	$p < 0.682$ n.s.
System unreliable	2.35	2.05	$p < 0.000$ ***
System understood me	3.27	3.78	$p < 0.000$ ***
Knew what to say to system	3.40	3.67	$p < 0.002$ **
Had to concentrate to hear	2.94	2.68	$p < 0.007$ *
System natural	2.55	2.93	$p < 0.002$ **
System slow	3.08	2.89	$p < 0.081$ n.s.
System friendly	2.96	3.36	$p < 0.003$ **
System reactions not as expected	2.89	2.48	$p < 0.000$ ***
System's expectations not clear	2.71	2.48	$p < 0.009$ *
Too many errors	2.16	1.93	$p < 0.000$ ***
Easy error recovery	3.30	3.26	$p < 0.198$ n.s.
System like human	2.54	2.83	$p < 0.015$.
System cooperative	3.31	3.74	$p < 0.000$ ***
Easy to lose way in dialog	2.51	2.17	$p < 0.000$ ***
Conversation unnatural	3.22	2.82	$p < 0.002$ **
Could Direct Conversation	2.90	2.90	$p < 0.819$ n.s.
Conversation too long	2.57	2.72	$p < 0.706$ n.s.
Reached aim quickly	3.41	3.49	$p < 0.141$ n.s.
Conversation balanced	3.04	3.10	$p < 0.852$ n.s.
System pleasant	3.00	3.21	$p < 0.224$ n.s.
Felt relaxed	3.11	3.43	$p < 0.005$ *
Had to concentrate mentally	3.11	2.82	$p < 0.005$ *
Fun to use	2.33	2.92	$p < 0.000$ ***
Satisfied with system	3.13	3.69	$p < 0.000$ ***
System difficult to use	2.34	2.17	$p < 0.002$ **
System easy to learn	3.75	4.02	$p < 0.000$ ***
System comfortable to use	3.08	3.47	$p < 0.001$ **
System inflexible	2.89	2.77	$p < 0.284$ n.s.
System not helpful	2.64	2.19	$p < 0.000$ ***
Would prefer different way	3.61	3.32	$p < 0.007$ *
Will use system again	3.33	3.56	$p < 0.009$ *
System like receptionist	2.83	3.07	$p < 0.042$.
Using system is worthwhile	3.34	3.65	$p < 0.000$ ***

.: $p<0.05$, *:$p<0.01$, **:$p<0.001$, ***:$p<0.0001$ or better.

that the systems do not react to their initiative, and rate the systems as less friendly and natural. They also report greater effort in using the system, and are significantly less satisfied overall and less likely to use the system again.

These judgements are tied to interaction style, not to age. When we examine differences between older and younger users in the Factual cluster (Table XVI), we see that, overall, older users are *more positive* than younger users. They are more likely to rate the system as being like a human and natural, while younger users show some frustration with the length of dialogs.

In summary, these analyzes show that it is not primarily age that affects key aspects of usability such as efficiency or user satisfaction. Instead, it appears to matter whether the system can accommodate the user's interaction style.

Table XVI. Differences in User Satisfaction Between Age Groups (Factual Cluster)

Feature	Older	Younger	Sig
Overall impression	3.74	3.67	$p < 0.751$ n.s.
Did not do what I wanted	2.27	2.22	$p < 0.372$ n.s.
System info clear	3.88	4.00	$p < 0.36$ n.s.
System info incomplete	2.13	2.09	$p < 0.445$ n.s.
System efficient	3.81	3.66	$p < 0.14$ n.s.
System unreliable	2.07	2.04	$p < 0.46$ n.s.
System understood me	3.79	3.78	$p < 0.808$ n.s.
Knew what to say to system	3.72	3.65	$p < 0.543$ n.s.
Had to concentrate to hear	2.57	2.73	$p < 0.243$ n.s.
System natural	3.33	2.76	$p < 0.000$ ***
System slow	2.58	3.03	$p < 0.003$ **
System friendly	3.34	3.37	$p < 0.431$ n.s.
System reactions not as expected	2.61	2.43	$p < 0.241$ n.s.
System's expectations not clear	2.34	2.54	$p < 0.228$ n.s.
Too many errors	1.97	1.91	$p < 0.229$ n.s.
Easy error recovery	3.23	3.27	$p < 0.717$ n.s.
System like human	3.21	2.67	$p < 0.000$ ***
System cooperative	3.71	3.75	$p < 0.608$ n.s.
Easy to lose way in dialog	2.12	2.19	$p < 0.394$ n.s.
Conversation unnatural	2.76	2.85	$p < 0.681$ n.s.
Could Direct Conversation	2.87	2.92	$p < 0.842$ n.s.
Conversation too long	2.37	2.86	$p < 0.002$ **
Reached aim quickly	3.56	3.45	$p < 0.759$ n.s.
Conversation balanced	3.45	2.96	$p < 0.000$ ***
System pleasant	3.21	3.21	$p < 0.604$ n.s.
Felt relaxed	3.41	3.44	$p < 0.798$ n.s.
Had to concentrate mentally	2.77	2.85	$p < 0.641$ n.s.
Fun to use	2.83	2.95	$p < 0.208$ n.s.
Satisfied with system	3.53	3.76	$p < 0.057$ n.s.
System difficult to use	2.23	2.15	$p < 0.537$ n.s.
System easy to learn	3.89	4.07	$p < 0.121$ n.s.
System comfortable to use	3.28	3.56	$p < 0.044$.
System inflexible	2.60	2.84	$p < 0.124$ n.s.
System not helpful	2.34	2.13	$p < 0.066$ n.s.
Would prefer different way	3.43	3.27	$p < 0.237$ n.s.
Will use system again	3.38	3.63	$p < 0.079$ n.s.
System like receptionist	3.07	3.07	$p < 0.955$ n.s.
Using system is worthwhile	3.57	3.68	$p < 0.419$ n.s.

.: $p<0.05$, *:$p<0.01$, **:$p<0.001$, ***:$p<0.0001$ or better.

6. DISCUSSION

Our analysis yields the following answers to our research questions (Section 1).

(1) Can users be categorized into distinct groups? Yes. The interaction style of older and younger users falls into two main groups, a "social" group which is characterized by more interpersonal communication, higher verbosity, and great variability between users, and a "factual" group that is characterized by the ability to adapt to the system, a concise communication style, and fairly uniform behavior.

(2) Can the interaction style of older users be predicted? Not from the data available to us. Whether a user falls into the "social" or the "factual" group cannot be predicted by age, gender, years in education, or cognitive abilities. Although older users dominate the "social" group, and younger users the "factual" group, a considerable minority of older users behave like younger users.
(3) Does interaction style affect usability? Yes. Interaction style affects efficiency and user satisfaction: "Social" users are less efficient and less satisfied with the system, which is tailored to the "factual" interaction style.

Our "social" older users conformed in several ways to the predictions made by research into aging. In terms of language production, they were more likely to supply additional information, which led to a more complex discourse. They also used a richer vocabulary. In terms of social cognition, they did not adapt to the very formulaic interaction offered by the system; instead, they tended to interact with the computer as they would with a human. This includes politeness markers and social speech acts that are not only unnecessary, but also likely to severely confuse an actual end-to-end SDS. Our "factual" older users, on the other hand, did not conform to these predictions at all; in fact, they were virtually indistinguishable from younger users.

6.1 Implications for Research

These results illustrate the need to include a wide variety of users of all ages in datasets for research into speech and language interfaces. Since older users may well behave just like younger users, it is important to recruit a substantial range and number of people. All five components of SDS will benefit greatly from corpora that contain a substantial number of older users. For illustration purposes, we discuss one aspect in depth, building statistical models for simulating user behavior [Schatzmann et al. 2006].

User simulations are a cornerstone of statistical approaches to spoken dialog systems. Data collection with real users requires substantial time and effort. In addition, every time a dialog strategy is modified, all experiments with human users must be restarted from scratch. Simulated users, on the other hand, allow different dialog policies to be tested efficiently and cost effectively. Using statistical techniques for learning dialog policies with simulated users enables exploration of strategies that are not present in existing corpora of human-machine dialogs. The systems can then learn new and potentially better dialog strategies.

What constitutes a good user simulation model is still an open question [Schatzmann et al. 2006; Georgila et al. 2008]. It is generally accepted in the field that the more diverse the available data, the more likely it is to train simulated users that capture realistic user behavior [Schatzmann et al. 2006]. Therefore it is important that simulated users are trained on large corpora representative of a wide range of user behaviors. Current corpora used for training simulated users, such as the DARPA Communicator corpus of flight-booking dialogs [Walker et al. 2002], do not appear to have been designed to ensure older people are sufficiently represented. It would be interesting to see whether a group similar to our Social users can be found in the DARPA Communicator corpus.

Our data is also restricted by the fact that our simulated systems were strictly system-initiative. We chose this approach because the original cognitive psychology experiment demanded that all users be exposed to the chosen dialog strategy several times during the dialog. We achieved this by imposing a clear, inflexible structure on all dialogs. However, as we have seen, our "social" users did not take kindly to this fixed, system-initiative structure. Thus, we hypothesize that a mixed-initiative approach would lead to higher efficiency and greater user satisfaction for this user group. Although well-designed mixed-initiative dialog systems can outperform comparable system-initiative dialog systems in terms of both user satisfaction and task efficiency [Chu-Carroll and Nickerson 2000], in practice, any mixed-initiative system that is tested with older users should also include system-initiative dialogs because these are a useful fall-back strategy for error recovery [Chu et al. 2007].

We recognize that our results are limited by the characteristics of our participant pool. Overall, both younger and older users were highly educated. Since the majority of older users were recruited from the participant pool of the Department of Psychology, University of Edinburgh, they may well be unusually open to experience and interested in technology. We would expect to find similar clusters in a more diverse population.

Another open question is whether the interaction style of older users can be predicted from user characteristics, and if yes, from which. The only predictor we found, age, was unreliable, and cognitive abilities did not explain any of the remaining variation. There are three additional candidates for predictors which we have not considered yet: attitude to technology, personality, and social cognition. Measures such as computer anxiety and computer self-efficacy affect whether older people will use technology [Ellis and Allaire 1999; Czaja et al. 2006]. Computer anxiety has been shown to affect the efficiency with which younger and older people perform computer-related tasks [Mahar et al. 1997; Laguna and Babcock 1997]. Personality is not only linked to computer anxiety, it also affects a person's interaction style, in particular in terms of vocabulary [Mairesse et al. 2007; Pennebaker et al. 2003]. Social cognition, in particular the capacity for theory of mind, is potentially even more important. Our "social" older users not only treated the computer system more like a human than like a computer, they were also less likely to adapt to the system's interaction style.

6.2 Implications for Design

Our "social" users pose real challenges for system design. They are more likely to use synonyms for simple answers such as "yes" or "no", more likely to ask for help and use meta-communication, more likely to be sociable, and more likely to supply the system with information that it cannot process. In related work on the present corpus, we have shown that the ASR component will need more sophisticated language models to accommodate the additional vocabulary [Vipperla et al. 2009]. The NLU module will require adequate strategies for identifying and discarding irrelevant material. However, due to the added complexity, such ASR and NLU systems are also more likely to fail. It remains to be seen whether the benefit of adapting to the user's preferred interaction style outweighs the added cost of error recovery dialogs and task failures. Other useful measures include using task-specific help prompts early on in the dialog [Zajicek et al. 2004; Wolters et al. ms], improving problem detection [Walker et al. 2002], and better error recovery [Skantze 2005; McTear et al. 2005]. It also remains to be seen what happens when our "social" users are faced with an end-to-end dialog system that is unable to process social speech acts and unexpected provision of information. It is not clear whether these users would be able to adopt a more effective interaction style, or whether they would give up in frustration. We would expect to see both outcomes.

In terms of dialog management and language generation, our usability results suggest that "social" users expect the system to adapt to their interaction style. This may mean that system prompts and messages need to be more polite and that the system may well benefit from adding appropriate social dialog acts. Note that such changes could be counter-productive for "factual" users, who value efficiency highly.

Based on our results, we would also caution against using a uniform "older people strategy" for all older users. Our "factual" older users actually slightly preferred the pared-down highly efficient systems they were presented with (Table XVI). As for the "social" cluster, the behavior of users in this cluster is highly variable (see Figure 7). This supports the user-sensitive inclusive design strategy of Newell and Gregor [2000]. The most extreme users are the most verbose ones with the largest vocabulary. A SDS with which these users can interact successfully should be able to handle most of the users in the "social" cluster relatively easily.

If designers prefer to work with "representative" users, these can be found with the `pam` cluster method. For all four datasets, `pam` led to reasonable partitions of the dataset. A solution based on this clustering technique was among the top four algorithms for each feature set.

7. CONCLUSION

In this article, we have shown that being old does not necessarily mean acting old. Even though older users were more likely to have a "social" interaction style than younger users, a sizeable proportion of older users preferred the same "factual" interaction style as younger users.

We plan to annotate the corpus further to allow more detailed analyzes of both language and usability. This should yield additional insights into the relation between interaction style and usability. Linguistic analyses will include part-of-speech tagging and basic syntactic analysis, while usability annotations will focus on problems and misunderstandings. They will be annotated using the schema outlined by Möller et al. [2007]. We will also run more detailed acoustic analyses to investigate the function of disfluencies and look for evidence of language production difficulties such as word finding problems.

It remains to be seen whether the clusters observed in our dataset can be replicated in other domains with more challenging tasks, and to what extent including full ASR and NLU will affect users' behavior. To this end, we plan to collect and analyze a corpus of interactions between older and younger users and a flight-booking system [Winterboer and Moore 2007; Moore et al. 2004]. For this experiment, the test battery will be expanded to cover some of the potential influences on interaction style discussed in Section 6.1 such as computer anxiety and theory of mind.

Finally, our findings need to be verified against deployed systems in the field. Ai et al. [2007] found that both word use and dialog act frequency differed between participants in laboratory experiments and real users in the field. For example, real users were more likely to request help than participants in experiments. It remains to be seen whether our "factual" older users would be just as factual in the real world, or whether they are likely to at least partially revert to a more "social" interaction style.

ACKNOWLEDGMENTS

We would like to thank R. Logie for collaborating with us on the design of the original experiment, N. Mayo and J. Eddy for coding the Wizard-of-Oz interface, N. Mayo for technical help with the experiment, V. Karaiskos for administering the spoken dialog experiment, M. Kronenthal for transcribing all 447 dialogs, and M. Watson for scheduling participants, administering the cognitive test battery, and data entry.

REFERENCES

AI, H., RAUX, A., BOHUS, D., ESKENAZI, M., AND LITMAN, D. 2007. Comparing spoken dialog corpora collected with recruited subjects versus real users. In *Proceedings of SIGDIAL*.

BAILEY, P. E. AND HENRY, J. D. 2008. Growing less empathic with age: Disinhibition of the self-perspective. *J. Gerontol. Series B—Psychol. Sci. Soc. Sci. 63*, 4, 219–226.

BARRAS, C., GEOFFROIS, E., WU, Z., AND LIBERMAN, M. 2000. Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Commun. 33*.

BLACK, L.-A., MCMEEL, C., MCTEAR, M., BLACK, N., HARPER, R., AND LEMON, M. 2005. Implementing autonomy in a diabetes management system. *J. Telemed. Telecare 11 Suppl 1*, 6–8.

BRANIGAN, H., PICKERING, M. J., PEARSON, J., AND MCNEIL, J. F. 2009. Linguistic alignment between humans and computers. *J. Pragmatics* (in press).

BRANIGAN, H. P., PICKERING, M. J., PEARSON, J., MCLEAN, J. F., AND NASS, C. 2003. Syntactic alignment between computers and people: The role of belief about mental states. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*.

BUCHANAN, B. G., MOORE, J., FORSYTHE, D. E., CARENINI, G., OHLSSON, S., AND BANKS, G. 1995. An interactive system for delivering individualized information to patients. *Artif. Intell. Med. 7*, 117–154.

BURKE, D. M. AND SHAFTO, M. A. 2004. Aging and language production. *Current Direct. Psychol. Sci. 13*, 1, 21–24.

CARENINI, G. AND MOORE, J. 2000. An empirical study of the influence of argument conciseness on argument effectiveness. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*.

CARENINI, G. AND MOORE, J. D. 2006. Generating and evaluating evaluative arguments. *Artif. Intell. 170*, 11, 925–952.

CARLETTA, J. 2007. Unleashing the killer corpus: Experiences in creating the multi-everything AMI Meeting Corpus. *Lang. Resources Eval. 41*, 2, 181–190.

CARLETTA, J., EVERT, S., HEID, U., KILGOUR, J., ROBERTSON, J., AND VOORMANN, H. 2003. The NITE XML toolkit: Flexible annotation for multi-modal language data. *Behav. Res. Methods, Instrum. Comput. 35*, 3, 353–363.

CATIZONE, R., DINGLI, A., PINTO, H., AND WILKS, Y. 2008. Information extraction tools and methods for understanding dialog in a companion. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

CHICKERING, D. AND PAEK, T. 2007. Personalizing influence diagrams: Applying online learning strategies to dialog management. *User Model. User-Adapted Interact. Special Issue Statist. Probabil. Methods User Model. 17*, 1-2, 71–91.

CHU, S.-W., O'NEILL, I., AND HANNA, P. 2007. Using multiple strategies to manage spoken dialog. In *Proceedings of Interspeech Conference*.

CHU-CARROLL, J. AND NICKERSON, J. S. 2000. Evaluating automatic dialog strategy adaptation for a spoken dialog system. In *Proceedings of the 1st Annual Meeting of the North American Chapter of the ACL (NAACL)*. 202–209.

CZAJA, S. AND LEE, C. 2007. The impact of aging on access to technology. *Univer. Access Inf. Soc. 5*, 341–349.

CZAJA, S. J., CHARNESS, N., FISK, A. D., HERTZOG, C., NAIR, S. N., ROGERS, W. A., AND SHARIT, J. 2006. Factors predicting the use of technology: Findings from the center for research and education on aging and technology enhancement (CREATE). *Psychol. Aging 21*, 333–352.

DAHLBAECK, N., JOENSSON, A., AND AHRENBERG, L. 1993. Wizard of Oz studies - Why and how. *Knowl.-Based Syst. 6*, 258–266.

DEARY, I. J., WHITEMAN, M. C., STARR, J. M., WHALLEY, L. J., AND FOX, H. C. 2004. The impact of childhood intelligence on later life: Following up the Scottish Mental Surveys of 1932 and 1947. *J. Personal. Soc. Psychol. 86*, 130–147.

DEMBERG, V. AND MOORE, J. D. 2006. Information presentation in spoken dialog systems. In *Proceedings of EACL Conference*.

DZIKOVSKA, M. O., CALLAWAY, C. B., FARROW, E., MARQUES-PITA, M., MATHESON, C., AND MOORE, J. D. 2007. Adaptive tutorial dialog systems using deep NLP techniques. In *Proceedings of the 2007 Meeting of NAACL and HLT Demo Session*.

ELLIS, R. D. AND ALLAIRE, J. C. 1999. Modeling computer interest in older adults: The role of age, education, computer knowledge, and computer anxiety. *Hum. Factors 41*, 345–355.

GARDEN, S., PHILLIPS, L., AND MACPHERSON, S. 2001. Mid-Life aging, open-ended planning, and laboratory measures of executive function. *Neuropsychol. 15*, 472–482.

GEORGILA, K., HENDERSON, J., AND LEMON, O. 2005. Learning user simulations for information state update dialog systems. In *Proceedings of Interspeech Conference*.

GEORGILA, K., WOLTERS, M., KARAISKOS, V., KRONENTHAL, M., LOGIE, R., MAYO, N., MOORE, J., AND WATSON, M. 2008. A fully annotated corpus for studying the effect of cognitive aging on users' interactions with spoken dialog systems. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

GEORGILA, K., WOLTERS, M., AND MOORE, J. 2008. Simulating the behavior of older versus younger users. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, Human Language Technologies (ACL/HLT)*, 49–52.

GILES, H. 2001. Speech accommodation. In *Concise Encyclopaedia of Sociolinguistics*. Elsevier, Oxford, UK, 193–197.

GIORGINO, T., AZZINI, I., ROGNONI, C., QUAGLINI, S., STEFANELLI, M., GRETTER, R., AND FALAVIGNA, D. 2005. Automated spoken dialog system for hypertensive patient home management. *Int. J. Med. Inf. 74*, 159–67.

GÖDDE, F., MÖLLER, S., ENGELBRECHT, K.-P., KÜHNEL, C., SCHLEICHER, R., NAUMANN, A., AND WOLTERS, M. 2008. Study of a speech-based smart home system with older users. In *International Workshop on Intelligent User Interfaces for Ambient Assisted Living*, 17–22.

GREGOR, P. AND DICKINSON, A. 2007. Cognitive difficulties and access to information systems: An interaction design perspective. *Univer. Access Inf. Soc. 5*, 393–400.

GREGOR, P., NEWELL, A. F., AND ZAJICEK, M. 2002. Designing for dynamic diversity - Interfaces for older people. In *Proceedings of the 5th International ACM Conference on Assistive Technologies (ASSETS'02)*, 151–156.

HAWTHORN, D. 2000. Possible implications of aging for interface designers. *Interact. Comput. 12*, 507–528.

HORTON, W. S. AND SPIELER, D. H. 2007. Age-Related differences in communication and audience design. *Psychol. Aging 22*, 2, 281–290.

ISO 1998. ISO Ergonomic requirements for office work with visual display terminals (VDTs)–Part 11: Guidance on usability.

JAJUGA, K., WALESIAK, M., AND BAK, A. 2003. On the general distance measure. In *Exploratory Data Analysis in Empirical Research*, M. Schwaiger and O. Opitz, Eds. Springer, Berlin/Heidelberg, 104–109.

JAMES, L. E., BURKE, D. M., AUSTIN, A., AND HULME, E. 1998. Production and perception of "verbosity" in younger and older adults. *Psychol. Aging 13*, 355–367.

KAUFMAN, L. AND ROUSSEEUW, P. J. 1990. *Finding Groups in Data*. Wiley, New York.

KEMPER, S., HERMAN, R., AND LIAN, C. 2003. Age differences in sentence production. *J. Gerontol. Series B: Psychol. Sci. Soc. Sci. 58*, 5, 260–8.

KLEIN, D. AND MANNING, C. D. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 423–430.

LAGUNA, K. AND BABCOCK, R. L. 1997. Computer anxiety in young and older adults: Implications for human-computer interactions in older populations. *Comput. Hum. Behav. 13*, 317–326.

LEISER, R. G. 1989. Exploiting convergence to improve natural language understanding. *Interact. Comput. 1*, 284–298.

LEMON, O., GEORGILA, K., HENDERSON, J., AND STUTTLE, M. 2006. An ISU dialog system exhibiting reinforcement learning of dialog policies: Generic slot-filling in the TALK in-car system. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

LEVIN, E., NARAYANAN, S., PIERACCINI, R., BIATOV, K., BOCCHIERI, E., FABBRIZIO, G. D., ECKERT, W., LEE, S., POKROVSKY, A., RAHIM, M., RUSCITTI, P., AND WALKER, M. 2000. The AT &T DARPA communicator mixed-initiative spoken dialog system. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'00)*.

LITMAN, D. AND SILLIMAN, S. 2004. ITSPOKE: An intelligent tutoring spoken dialog system. In *Companion Proceedings of the Human Language Technology Conference: 4th Meeting of the North American Chapter of the Association for Computational Linguistics*.

MAHAR, D., HENDERSON, R., AND DEANE, F. 1997. The effects of computer anxiety, state anxiety, and computer experience on users' performance of computer based tasks. *Personal. Individ. Differ. 22*, 683–692.

MAIRESSE, F., WALKER, M., MEHL, M., AND MOORE, R. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Intell. Res. 30*, 457–500.

MAYLOR, E. A., MOULSON, J. M., MUNCER, A. M., AND TAYLOR, L. A. 2002. Does performance on theory of mind tasks decline in old age? *British J. Psychol. 93*, 465–485.

MCTEAR, M., O'NEILL, I., HANNA, P., AND LIU, X. 2005. Handling errors and determining confirmation strategies - An object-based approach. *Speech Commun. 45*, 249–269.

MÖLLER, S., ENGELBRECHT, K.-P., AND OULASVIRTA, A. 2007. Analysis of communication failures for spoken dialog systems. In *Proceedings of Interspeech Conference*.

MÖLLER, S., KREBBER, J., AND SMEELE, P. 2006. Evaluating the speech output component of a smart-home system. *Speech Commun. 48*, 1–27.

MOORE, J. D., FOSTER, M.-E., LEMON, O., AND WHITE, M. 2004. Generating tailored, comparative descriptions in spoken dialog. In *Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference*, 917–922.

MÜLLER, C., WITTIG, F., AND BAUS, J. 2003. Exploiting speech for recognizing elderly users to respond to their special needs. In *Proceedings of Eurospeech Conference*. 1305–1308.

NASS, C. AND BRAVE, S. 2005. *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. The MIT Press.

NEF, F. AND HUPET, M. 1992. Normal aging effects on oral and written production of spontaneous language. *Année Psychol. 92*, 3, 393–419.

NEWELL, A. AND GREGOR, P. 2000. "User sensitive inclusive design" — In search of a new paradigm. In *Proceedings of the ACM Conference on Universal Usability (CUU'00)*. 39–43.

NIEDERHOFFER, K. G. AND PENNEBAKER, J. W. 2002. Linguistic style matching in social interaction. *J. Lang. Soc. Psychol. 21*, 337–360.

NIVRE, J. AND NILSSON, J. 2005. Pseudoprojective dependency parsing. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 99–106.

OVIATT, S., DARVES, C., AND COULSTON, R. 2004. Toward adaptive conversational interfaces: Modeling speech convergence with animated personas. *ACM Trans. Comput.-Hum. Interact. 11*, 3, 300–328.

PAEK, T. AND CHICKERING, D. 2007. Improving command-and-control speech recognition on mobile devices: Using predictive user models for language modeling. *User Model. User-Adapted Interact. Special Issue Statist. Probabil. Methods User Model. 17*, 1-2, 93–117.

PEARSON, J., HU, J., BRANIGAN, H. P., PICKERING, M., AND NASS, C. 2006. Adaptive language behavior in HCI: How expectations and beliefs about a system affect users' word choice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing System (CHI'06)*.

PENNEBAKER, J. W., MEHL, M. R., AND NIEDERHOFFER, K. G. 2003. Psychological aspects of natural language use: Our words, our selves. *Ann. Rev. Psychol. 54*, 547–577.

PENNEBAKER, J. W. AND STONE, L. D. 2003. Words of wisdom: Language use over the life span. *J. Personal. Soc. Psychol. 85*, 291–301.

PETRIE, H. 2001. Accessibility and usability requirements for icts for disabled and elderly people: A functional classification approach. In *Inclusive Design Guidelines for HCI*, C. Nicolle and J. Abascal, Eds. Taylor and Francis, London, UK, 29–60.

PICKERING, M. J. AND GARROD, S. 2004. Toward a mechanistic psychology of dialog. *Behav. Brain Sci. 27*, 169–226.

POLLACK, M., BROWN, L., COLBRY, D., MCCARTHY, C. E., OROSZ, C., PEINTNER, B., RAMAKRISHMAN, S., AND TSAMARDINOS, I. 2003. Autominder: An intelligent cognitive orthotic system for people with memory impairment. *Robotics Auton. Syst. 44*, 273–282.

PULLIN, G. AND NEWELL, A. 1997. Focussing on extra-ordinary users. In *Universal Access in Human Computer Interaction. Coping with Diversity*. Springer, Berlin, 253–262.

R DEVELOPMENT CORE TEAM. 2008. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

RABBITT, P. 1990. Mild hearing loss can cause apparent memory failures which increase with age and reduce with IQ. *Acta Oto-Laryngologica Suppl. 476*, 167–175.

RABBITT, P. AND ANDERSON, M. 2006. The lacunae of loss? Aging and the differentiation of cognitive abilities. In *Lifespan Cognition: Mechanisms of Change*, E. Bialystok and F. I. Craik, Eds. Oxford University Press, New York, Chapter 23.

RAMIG, L. O., GRAY, S., BAKER, K., CORBIN-LEWIS, K., BUDER, E., LUSCHEI, E., COON, H., AND SMITH, M. 2001. The aging voice: A review, treatment data and familial and genetic perspectives. *Folia Phoniatrica Logopaedica 53*, 252–265.

RAUX, A., BOHUS, D., LANGNER, B., BLACK, A., AND ESKENAZI, M. 2006. Doing research on a deployed spoken dialog system: One year of Let's Go! experience. In *Proceedings of Interspeech Conference*.

RAVEN, J., RAVEN, J., AND COURT, J. 1998. *Manual for Raven's Progressive Matrices and Vocabulary Scales.* Harcourt Assessment, San Antonio, TX.

ROUSSEEUW, P. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math. 20*, 53–65.

ROY, N., PINEAU, J., AND THRUN, S. 2000. Spoken dialog management using probabilistic reasoning. In *Proceedings of the 38th Annual Meeting of the Association for Computation Linguistics (ACL)*, 93–100.

SALTHOUSE, T. A. 2004. What and when of cognitive aging. *Current Direct. Psychol. Sci. 13*, 140–144.

SCHAIE, K. W. 1994. The course of adult intellectual development. *Amer. Psychol. 49*, 4, 304–313.

SCHATZMANN, J., WEILHAMMER, K., STUTTLE, M., AND YOUNG, S. 2006. A survey of statistical user simulation techniques for reinforcement-Learning of dialog management strategies. *Knowl. Eng. Rev. 21*, 97–126.

SEARLE, J. R. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge.

SENEFF, S., LAU, R., PAO, C., AND ZUE, V. 1998. Galaxy-II: A reference architecture for conversational system development. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'98)*. 931–934.

SHARIT, J., CZAJA, S. J., NAIR, S., AND LEE, C. C. 2003. Effects of age, speech rate, and environmental support in using telephone voice menu systems. *Hum. Factors 45*, 234–251.

SHEEDER, T. AND BALOGH, J. 2003. Say it like you mean it: Priming for structure in caller responses to a spoken dialog system. *Int. J. Speech Technol. 6*, 103–111.

SKANTZE, G. 2005. Exploring human error recovery strategies: Implications for spoken dialog systems. *Speech Commun. 45*, 325–341.

SWIFT, M. AND ALLEN, J. 2004. Skeletons in the parser: Using a shallow parser to improve deep parsing. In *Proceedings of the International Conference on Computational Linguistics (COLING'04)*.

TAN, P.-N., STEINBACH, M., AND KUMAR, V. 2005. *Introduction to Data Mining*. Addison-Wesley.

TANG, P. AND VENABLES, T. 2000. 'Smart' homes and telecare for independent living. *J. Telemed. Telecare 6*, 8–14.

TRAUM, D. 1994. A computational theory of grounding in natural language conversation. Ph.D. thesis, University of Rochester.

TRAUM, D. R., SWARTOUT, W., GRATCH, J., AND MARSELLA, S. 2008. A virtual human dialog model for non-team interaction. In *Recent Trends in Discourse and Dialogue*, L. Dybkjaer and W. Minker, Eds. Springer, New York, 45–67.

UNSWORTH, N. AND ENGLE, R. W. 2005. Individual differences in working memory capacity and learning: Evidence from the serial reaction time task. *Memory Cognit. 33*, 213–20.

VERHAEGHEN, P. 2003. Aging and vocabulary scores: A meta-analysis. *Psychol. Aging 18*, 332–339.

VIPPERLA, R., WOLTERS, M., GEORGILA, K., AND RENALS, S. 2009. Speech input from older users in smart environments: Challenges and perspectives. In *Proceedings of HCI International*.

WALESIAK, M. 1999. Distance measure for ordinal data. *Argumenta Oeconomica 2*, 8, 167–173.

WALESIAK, M. 2008. Cluster analysis with ClusterSim computer program and R environment. *Acta Universitatis Lodziniensis, Folia Oeconomica 216*, 303–311.

WALKER, M., WHITTAKER, S., STENT, A., MALOOR, P., MOORE, J., JOHNSTON, M., AND VASIREDDY, G. 2005. Generation and evaluation of user tailored responses in multimodal dialog. *Cognit. Sci. 28*, 5, 811–840.

WALKER, M. A., LANGKILDE-GEARY, I., HASTIE, H. W., WRIGHT, J., AND GORIN, A. 2002. Automatically training a problematic dialog predictor for the HMIHY spoken dialog system. *J. Artif. Intell. Res. 16*, 105–133.

WALKER, M. A., RUDNICKY, A., PRASAD, R., ABERDEEN, J., POTAMIANOS, A., PASSONNEAU, R., ROUKOS, S., SANDERS, G. A., SENEFF, S., AND STALLARD, D. 2002. DARPA communicator: Cross-System results for the 2001 evaluation. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'02)*, 269–272.

WEBB, A. R. 2002. *Statistical Pattern Recognition*, 2nd ed. Wiley, New York.

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

WINTERBOER, A. AND MOORE, J. D. 2007. Evaluating information presentation strategies for spoken recommendations. In *Proceedings of Recommender Systems Conference*.

WOLTERS, M., CAMPBELL, P., DEPLACIDO, C., LIDDELL, A., AND OWENS, D. 2007. Making synthetic speech accessible to older people. In *Proceedings of the 6th ISCA Workshop on Speech Synthesis*, 288–293.

WOLTERS, M., ENGELBRECHT, K.-P., GÖDDE, F., MÖLLER, S., NAUMANN, A., AND SCHLEICHER, R. ms. Using help prompts to shape users' speech. Manuscript.

WOLTERS, M., GEORGILA, K., LOGIE, R., MACPHERSON, S., MOORE, J., AND WATSON, M. 2009. Reducing working memory load in spoken dialogs: Do we have to limit the number of options? *Interact. Comput.* (accepted).

YOUNG, S. 2000. Probabilistic methods in spoken dialog systems. *Philosoph. Trans. Royal Soc. (Series A) 358*, 1769, 1389–1402.

ZAJICEK, M., WALES, R., AND LEE, A. 2004. Speech interaction for older adults. *Univer. Access Inf. Soc. 3*, 2, 122–130.

ZINN, C., MOORE, J. D., AND CORE, M. G. 2002. A 3-tier planning architecture for managing tutorial dialog. In *Proceedings of Intelligent Tutoring Systems, 6th International Conference*, 574–584.

ZOLTAN-FORD, E. 1991. How to get people to say and type what computers can understand. *Int. J. Man-Mach. Studies 34*, 527–547.

Received November 2008; revised March 2009; accepted March 2009