An Analysis of Older Users' Interactions with Spoken Dialogue Systems

Jamie Bost, Johanna D. Moore
University of Edinburgh, School of Informatics
10 Crichton Street, Edinburgh, EH8 9AB
E-mail: s1145415@ed-alumni.net, j.moore@ed.ac.uk

Abstract

This study explores communication differences between older and younger users with a task-oriented spoken dialogue system. Previous analyses of the MATCH corpus show that older users have significantly longer dialogues than younger users and that they are less satisfied with the system. Open questions remain regarding the relationship between information recall and cognitive abilities. This study documents a length annotation scheme designed to explore causes of additional length in the dialogues and the relationships between length, cognitive abilities, user satisfaction, and information recall. Results show that primary causes of older users' additional length include using polite vocabulary, providing additional information relevant to the task, and using full sentences to respond to the system. Regression models were built to predict length from cognitive abilities and user satisfaction from length. Overall, users with higher cognitive ability scores had shorter dialogues than users with lower cognitive ability scores, and users with shorter dialogues were more satisfied with the system than users with longer dialogues. Dialogue length and cognitive abilities were significantly correlated with information recall. Overall, older users tended to use a human-to-human communication style with the system, whereas younger users tended to adopt a factual interaction style.

Keywords: Spoken dialogue corpora, Spoken dialogue systems, Cognitive ageing

1. Introduction

Health care-related services such as in-home monitoring, delivery services, and communication between health professionals and patients increasingly rely on the use of modern technology (Czaja & Lee, 2007). Older populations (65+ years) could benefit from these systems that can improve their quality of life, but studies have shown that fewer members of this population use certain kinds of modern technology (Czaja & Lee, 2007). Thus, it is important to understand how older populations use technology and where difficulties arise, so systems can be designed to accommodate users with varying technological experiences and abilities.

A type of technology with which older users interact is a Spoken Dialogue System (SDS), which accepts speech input and produces speech output (Wolters et al., 2009). In addition to perhaps having less experience and familiarity with technology, many older users' cognitive and perceptual (i.e. hearing) abilities tend to decline, so they may employ different dialogue strategies when using an SDS (Georgila et al., 2010; Möller et al., 2008). While studies have been conducted to describe existing home or tele-care systems (Black et al., 2005; Pollack, 2005; Zajicek, 2004), more research is needed to address the needs of older users and examine how older users interact with an SDS compared to younger users. The JASMIN-CGN (Cucchiarini et al., 2006) corpus began to address this gap, as it is a collection of older users' utterances used to aid SDS design. The MeMo project (Möller et al., 2008) is directly relevant to this study, as it compares older and younger users' interactions with a command-and-control SDS and examines task success. The current study examines older users' interactions with an SDS and the differences in the communication styles of older and younger users contained in the MATCH corpus (Georgila et al., 2010). Whereas previous studies on the MATCH corpus showed quantitatively that older users' dialogues were significantly longer than those of younger users, the present study examines these differences between groups of users qualitatively to determine the reasons for the additional length in older users' dialogues.

2. The MATCH Corpus

The MATCH corpus contains 446 dialogues of older and younger users with a system that schedules appointments with health care professionals. In addition to providing a corpus that includes older users' dialogues, a primary purpose of the original study was to examine the effects of cognitive ageing on older users' interactions with an SDS. The study systematically varied (1) the number of appointment options that users were presented with (one option, two options, four options), and (2) the confirmation strategy employed (explicit, implicit, or no confirmation). This 3 X 3 design yielded 9 different dialogue strategies. The 9 dialogue strategies were simulated using a Wizard of Oz (WoZ) method, in which a human "wizard" interprets the user's speech input and performs dialogue management to determine the system's next dialogue move. The system controlled the number of options presented and confirmation strategy, and ensured that the dialogues progressed in stages of determining which health care professional to see, which half-day in a week, and at what time to schedule the appointment.

The corpus is richly annotated with Information State Update information, which includes information about the preceding dialogue relevant to making dialogue management decisions, e.g., which time slots have been confirmed. The corpus has also been annotated with dialogue acts, each of which is a < speech act, task > pair where the speech act is task independent (e.g., accept-info) and the task corresponds to one of the stages of the appointment scheduling dialogue (e.g., time-slot). For full details about the corpus and annotation, see Georgila et al. (2010).

Participants include 26 older users (aged 50-85) and 24 younger users (aged 20-30). Results from statistical analyses showed that older users produced significantly longer dialogues than younger users, and they used a wider variety of speech acts and vocabulary (Georgila et al., 2008). Older users tended to use more social interaction or give additional details about availability during certain time slots, whereas younger users simply tended to accept, reject, or confirm slots.

Usability surveys were also conducted, which allowed users to rate their satisfaction on a scale of 1-5 (where 1 was "very poor" and 5 was "very good") and showed older users to be less satisfied with the system than younger users.

In order to assess the effect of users' cognitive abilities on their interaction with each of the nine systems, all participants underwent a comprehensive battery of cognitive assessments. These tests measure: crystallised intelligence, the ability to use acquired skills and experiences (MillHill, Raven and Court, 1998); fluid intelligence, the ability to reason and solve novel problems (Ravens, Raven and Court, 1998); information processing speed, the speed at which sensory input is processed (DSST, Wechsler, 1981); and working memory capacity, the limitations of cognitive performance (SentSpan, Unsworth and Engle, 2005). Although it was predicted that presenting fewer options and using explicit confirmations would aid older users or users with lower working memory span, there was a ceiling effect for task performance, as 92% of tasks were completed successfully, with users scheduling appointments with the correct health professional at possible times. Thus, neither dialogue strategy nor cognitive abilities affected task success in this study. Information recall was also measured with a cumulative score (on a scale from 0-2 for each detail) of how well a user remembered the details of the booking (i.e., health professional, day, time, and location of the appointment). Working memory span was not correlated with appointment recall, but users with lower information processing speed had less success recalling appointment details (Wolters et al., 2009). The current study aims to provide a deeper understanding of what contributes to dialogue length and explores the relationships between measures of dialogue length with cognitive abilities, information recall, and user satisfaction.

3. Research Questions

As described above, open questions remain with respect to how older users' interactions with an SDS differ qualitatively from those of younger users in the MATCH corpus. We designed an original annotation scheme to explore the causes of additional length in the MATCH corpus and the differences in communication styles between age groups. The following research questions directed the analysis of the data created by the new length annotations:

1. Why are older users' dialogue lengths significantly longer than those of younger users?
2. How does length relate to measures of interaction success, i.e. information recall and user satisfaction?
3. How does length relate to cognitive abilities?

4. Annotation Scheme

An annotation system was specifically designed to identify causes of additional length and the differences between users' communication styles. Any words or utterances in a dialogue unnecessary for completing the task were annotated. For example, if the system asks, "Would you like to see the diabetes nurse?" the only required response from the user would be "Yes" or "No." A response containing unnecessary utterances (with annotations) would be, "No, *[long_provide* I want to see the physiotherapist] [*future_details* on Thursday afternoon] [*polite_vocabulary* please]."

The basic structure of the annotation scheme was developed by taking a small sample of dialogues (36), identifying instances of additional length, and attributing categories to these instances. The scheme was expanded by testing it against samples of the remaining dialogues.

Once the annotation scheme was complete, a length annotation tool was built using NITE XML TOOLKIT (NXT, Carletta et al., 2003).

The length annotation scheme is summarised below in Table 1. Two coders applied the scheme to a sample of the dialogues. Inter-coder reliability for all categories of the length annotation scheme was high, with Cohen's (1960) $K > 0.75$ for all annotation categories. Thus, the results of the length annotation coding were carried forward for analysis.

5. Analysis

5.1 Length Annotation Statistics

Table 2 shows the percentage of annotated text of each length annotation category for older (n=26) vs. younger (n=24) users, as well as the mean length of the length annotation for each subcategory. The categories that account for most of the annotated text were *Verbose Answers, Polite Exchanges, and Overanswering* for older users. In eight out of nine categories, the sum of words in each category was statistically significantly different between groups. In addition, older users' average category length tended to be an order of magnitude longer than that of younger users.

5.2 Regression Analysis

SPSS, version 19 (IBM, 2012), was used to perform statistical analysis, such as independent samples *t*-tests to compare group differences (see Table 2), and multiple regressions. For these regressions, we hypothesised:

H1: Cognitive abilities will predict dialogue length, and they will be negatively correlated.

H2: Length and cognitive abilities will predict information recall, and length will be negatively correlated with information recall, whereas cognitive abilities will be positively correlated with information recall.

H3: Length and cognitive abilities will predict user satisfaction, and they will be negatively correlated with

Category	**Subcategory**	**Description**
Task Communication	wrong_task	The user communicates about a future but not current task
	delayed_response	The user delays response so the system repeats the prompt
Repeat Information	repeat_booking_details	The user repeats booking details already stated
	repeat_confirmation	The user restates a confirmation
Overanswering	long_provide	The user provides information about the current task in a long statement
	future_details	The user provides booking details that have not yet been addressed
	multiple_options	The user provides multiple options as an answer
Verbose Answers	long_confirmation	The user's confirm statement is long
	long_acceptance	The user's accept statement is long
	long_rejection	The user's reject statement is long
Polite Exchanges	polite_vocabulary	The user uses polite vocabulary
User Produces Incorrect or Extra Statement	user_changes_mind	The user changes his/her mind about booking details
	provides_incorrect_information	The user provides incorrect information
	comment_extra	The user makes an additional comment
	understand_prompt_neg	The user does not understand the prompt
User Requests	request _possible	The user makes a request that the system can accommodate
	request_ impossible	The user makes a request that the system cannot accommodate
System Re-requests	system_stalls	The system stalls the dialogue
	system_requests	The system asks the user to repeat him/herself or make a selection
	system_error	The system makes an error
	user_responds	The user responds to a system request or error
Disfluencies	disfluency	The user's speech contains a disfluency
Other	other	The dialogue's length is caused by an undescribed factor

Table 1: Length Annotation Scheme Summary

Category/Subcategory	**Older**	**Younger**	**Sig.**
% of words with length annotations	72.5%	20.2%	
Task Communication	*32 (0.6%)*	*23 (4.8%*	*n.s.*
wrong_task	7.0±0.8	2.3±2.9	n.s.
delayed_response	2.0±0.0	1.0±0.0	n.s.
Repeat Information	*786 (14.3%)*	*35 (7.3%)*	***
repeat_booking	2.7±0.3	1.1±0.3	**
repeat_confirmation	1.75±1.1	1.1±0.3	**
Overanswering	*934 (17.0%)*	*70 (14.6%)*	***
long_provide	5.9±2.5	4.8±1.7	n.s.
future_details	4.5±2.4	3.1±1.1	n.s.
multiple_options	4.4±2.4	NA	NA
Verbose Answers	*1141 (20.8%)*	*71 (14.8%)*	***
long_confirmation	3.5±1.1	3.0±0.0	n.s.
long_acceptance	6.1±2.5	5.0±1.8	n.s.
long_rejection	4.7±1.7	2.0±1.0	*
Polite Exchanges	*1079 (19.6%)*	*143 (29.7%)*	***
polite_vocabulary	1.5±0.8	1.1±0.3	**
Extra Statement	*533 (9.7%)*	*13 (2.7%)*	***
user_changes_mind	2.7±2.2	1.3±0.8	*
incorrect_information	2.4±2.6	NA	NA
comment_extra	4.5±5.6	1.0±0.0	n.s.
understand_neg	3.9±2.9	3.0±0.0	n.s.
User Requests	*350 (6.4%)*	*19 (4.0%)*	***
request _possible	5.3±1.8	5.3±1.2	n.s.
request_ impossible	5.8±2.9	3.0±0.0	n.s.
User Responds	*127 (2.3%)*	*31 (6.4%)*	*
user_responds	3.3±4.4	1.6±0.7	n.s.
Disfluencies	*510 (9.3%)*	*73 (15.2%)*	***
disfluency	1.4±0.9	1.2±1.1	n.s.

Table 2: Length in number of words of annotation categories: sum of words used in category, percentage of annotated words in a category within a user group, and mean length of an annotation ± standard deviation. Significance was measured using independent samples, one-tailed t-tests. *: $p < 0.05$, **: $p < 0.01$, ***: $p < .001$

user satisfaction.

Dialogue length was measured by the ratio of annotated words to the total number of words (per dialogue). Cognitive abilities were measured by the test scores for each participant as described above. Information recall and user satisfaction were measured by the score and satisfaction ratings, respectively, as described in Wolters et al. (2009). We divided the annotation categories into two groups: one was a "verbose" group of categories including *Verbose Answers, Overanswering,* etc., and the other was a "system navigation" group including *System Re-requests, Task Communication,* etc.

6. Results

6.1 Predicting Length

We found that age, SentSpan, Raven and MillHill tests were significant predictors of length, yielding a model with an adjusted $R^2 = .365$, where $F_{4,424} = 62.5, p < 0.001$, using the stepwise method. For Age, the correlation is positive, because older users tend to have longer dialogues. SentSpan and DSST are negatively correlated with dialogue length, showing that users with lower working memory capacity and lower information processing speed have longer dialogues.

6.2 Predicting Information Recall

Although we were able to build a statistically significant regression model to predict information recall, it only accounted for 5% of the variance in information recall score. We believe that this result is caused by the observed ceiling effect of task success and information recall discussed above.

In lieu of a reliable regression model, it is worth reporting the correlations (Spearman's ρ) of the relevant variables (see Table 3). The variables used in the regression models, e.g., Ravens, DSST, and length measures, are significantly correlated with information recall score. Based on the signs of the correlation coefficients', these results suggest that users with shorter dialogues have higher information recall scores than users with longer dialogues, as do users with higher information processing speed and fluid intelligence.

Variable	**Correlation Coefficient**	***p***
No. annotated words	-.107	*
Verbose group freq.	-*.131*	**
Verbose group lengths	-.114	**
Ravens	.155	**
DSST	*.182*	***

Table 3: Correlations (Spearman's ρ) with information recall score. *: $p < 0.05$, **: $p < 0.01$, ***: $p < .001$

6.3 Predicting User Satisfaction

We hypothesised that length and cognitive abilities would predict user satisfaction, and they would be negatively correlated with user satisfaction (H3). Users with shorter dialogues and higher cognitive test scores would be more satisfied with the system than users with longer dialogues and lower cognitive test scores.

A model using age, cognitive test scores, and "verbose" and "system navigation" categories' lengths to predict user satisfaction had an adjusted $R^2 = .179$, where $F_{5,440} = 20.5, p < 0.001$, using the stepwise method.

As expected, all correlations with satisfaction are negative. Users with longer dialogues or who had more difficulty communicating with the system have lower satisfaction ratings, as did users with lower information processing speed.

7. Future Work

While this study provides valuable qualitative information about user groups' communication styles and explores the different relationships between dialogue length, cognitive abilities, user satisfaction, and information recall, further work is needed to address task success and information recall, in particular. Another experiment should be run with a more challenging, yet still relevant, task, which would generate more normally distributed task success. Many questions could be re-examined, from the original design to test dialogue strategy and working memory capacity, to the current issue of whether dialogue length and cognitive abilities can reliably predict score. From the existing correlations, it seems likely that there is a significant relationship, which could be explored with a more normally distributed set of information recall scores.

8. Conclusion

Overall, we were able to provide answers to our research questions. Using the data provided from our length annotation scheme, we found that older users' dialogues were significantly longer than those of younger users. Older users' additional dialogue length was primarily comprised of *Verbose Answers, Polite Exchanges*, and *Overanswering*, as they tended to respond in full sentences, repeat the details of their appointments, or provide additional details about their availability. Their average category length was also an order of magnitude longer than that of younger users. We were able to build significant regression models that predict dialogue length from users' cognitive ability scores and predict user satisfaction from dialogue length. In general, users with lower cognitive ability scores had longer dialogues than users with higher cognitive ability scores, and users with longer dialogues were less satisfied with the system than users with shorter dialogues. We also found significant correlations between dialogue length and cognitive ability scores with information recall, suggesting that users with higher cognitive ability scores and shorter dialogues have higher information recall scores than other users. Similar to previous studies (Georgila et al., 2010; Möller et al., 2008), we found that older users tend to communicate with an SDS as if they were communicating with a human, whereas younger users adopt a factual style of interaction to communicate with the system. Finally, these findings contribute qualitative information that can be used to improve SDS design, making this important and increasingly common technology accessible to users who can benefit from it**.**

9. References

Castor, A., Pollux, L.E. (1992). The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1), pp. 37--53.

Chercheur, J.L. (1994). *Case-Based Reasoning.* San Mateo, CA: Morgan Kaufman Publishers.

Grandchercheur, L.B. (1983). Vers une modélisation cognitive de l'être et du néant. In S.G Paris, G.M. Olson, & H.W. Stevenson (Eds.), *Fondement des Sciences Cognitives*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 6--38.

Martin, L.E. (1990). Knowledge Extraction. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 252--262.

Superman, S;-Batman, B ; Catwoman, C. and Spiderman, S. (2000) *Superheroes experiences with books*. Gotham City: The Phantom Editors Associates.

Zavatta, A. (1992). Un Générateur d'Insultes s'intégrant dans un Système de Dialogue Humain-Machine. Thèse de Doctorat en Informatique. Université Paris-sud, Centre d'Orsay.

Black, L.-A., McMeel, C., McTear, M., Black, N., Harper, R., and Lemon, M. (2005). Implementing autonomy in a diabetes management system. *Journal of Telemedicine and Telecare*, 11(1), pp. 6--8.

Carletta, J., Evert, S., Heid, U., Kilgour, J., Robertson, J., and Voormann, H. (2003). The NITE XML Toolkit : Flexible annotation for multimodal language data. *Behavior Research Methods, Instruments, & Computers*, 35(3), pp. 353--363.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), pp. 37--46.

Cucchiarini, C., Van Hamme, H., Van Herwijnen, O., and Smits, F. (2006). Jasmin-CGN: Extension of the spoken dutch corpus with speech of elderly people, children and non-natives in the human-machine interaction modality. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*(1), pp. 135--138.

Czaja, S. J., and Lee, C. C. (2007). The Impact of Aging on Access to Technology. *Universal Access in the Information Society*, 5(4), pp. 341--349.

Georgila, K., Wolters, M., Karaiskos, V., Kronenthal, M., Logie, R., Mayo, N., et al. (2008). A Fully Annotated Corpus for Studying the Effect of Cognitive Ageing on Users' Interactions with Spoken Dialogue Systems. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*(1), pp. 938--944.

Georgila, K., Wolters, M., Moore, J. D., and Logie, R. H. (2010). The MATCH corpus : a corpus of older and younger users' interactions with spoken dialogue systems. *Language Resources and Evaluation*, 44(3), pp. 221--261.

Georgila, K., Wolters, M. K., and Moore, J. D. (2010). Learning Dialogue Strategies from Older and Younger Simulated Users. In *Proceedings of 11th Annual Meeting of the ACL Special Interest Group on Discourse and Dialogue (SIGDIAL)*(1), pp. 103--106.

IBM (2014). *SPSS*. Available from http://www-01.ibm.com/software/uk/analytics/spss/

Möller, S., Gödde, F., and Wolters, M. (2008). Corpus Analysis of Spoken Smart-Home Interactions with Older Users. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*(1), pp. 735--740.

Pollack, M. (2005). Intelligent technology for an aging population: The use of AI to assist elders with cognitive impairment. *AI Magazine*, 26(2), pp. 9--24.

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

Unsworth, N., and Engle, R. (2005). Individual differences in working memory capacity and learning: evidence from the serial reaction time task. *Memory and Cognition*, 33(1), pp. 213--220.

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

Wolters, M., Georgila, K., Moore, J. D., Logie, R. H., Macpherson, S. E., and Watson, M. (2009). Reducing working memory load in spoken dialogue systems. *Interacting with Computers*, 21(4), pp. 276--287.

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 Dialogue Systems. *ACM Transactions on Accessible Computing*, 2(1), pp. 1--31.

Zajicek, M. (2004). Successful and available: interface design exemplars for older users. *Interacting with Computers*, 16(1), pp. 411–430.