Exploring Language Style in Chatbots to Increase Perceived Product Value and User Engagement

Ela Elsholz, Jon Chamberlain, Udo Kruschwitz
School of Computer Science and Electronic Engineering, University of Essex
Colchester, Essex, UK
ela.elsholz@posteo.de,[jchamb|udo]@essex.ac.uk

ABSTRACT

Chatbots that can automatically answer customer requests have become a common feature on e-commerce Web sites. There are many factors that might affect overall customer satisfaction with such services. We explore how adding language style to e-commerce chatbots can be used to increase user satisfaction, perceived product value, user interest in a product, and user engagement with a chatbot service. We conducted an experimental pilot study, where two chatbots were used to sell theatre tickets: one communicating in modern English and one in a Shakespearean-style dialect. 169 participants interacted with a randomly-assigned version of the chatbot. The results indicate that the bot talking in modern English showed a significantly higher user satisfaction, whereas the Shakespearean-styled chatbot showed higher user engagement and perceived product value. It was also found that the modern chatbot version was more often referred to as being 'easy to use', whereas the Shakespearean chatbot version was more often referred to as being 'fun to use'.

CCS CONCEPTS

• **Information systems** → **Chat**; **Electronic commerce**; • **Human-centered computing** → *Usability testing*; *Natural language interfaces*;

KEYWORDS

Chatbots, Dialogue system, Styled conversational interface, Conversational agent, Language style transfer, Shakespeare

ACM Reference Format:
Ela Elsholz, Jon Chamberlain, Udo Kruschwitz. 2019. Exploring Language Style in Chatbots to Increase Perceived Product Value and User Engagement. In *2019 Conference on Human Information Interaction and Retrieval (CHIIR '19), March 10–14, 2019, Glasgow, Scotland Uk.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3295750.3298956

1 INTRODUCTION

Conversational approaches to information interaction are not at all new, e.g. [3, 5], but they have attracted renewed interest in recent years, e.g. [8, 11]. As a result, chatbots are becoming ever more popular [17]. With the increasing integration of chatbots into daily life, their personalisation is starting to play an important role. Google has recently published a chatbot that can hardly be distinguished from a human.[1] In fact, several studies indicate that users prefer personalised chatbots [1, 9]; however, while the style of the conversation in dialogue systems has been the centre of research attention [18], to the best of our knowledge no studies to date have explored the impact of applying different language styles to chatbot responses on user satisfaction and engagement. As language style is a key component to the way chatbots communicate with users, it may influence a user's experience, and possibly even influence a user's attitude towards a product or service. This study is a first step to addresses this lack of research.

2 RELATED WORK

A chatbot service can be defined as 'a computer software program which interacts with a user through natural language dialogue and provides some form of service' [14]. For the purpose of this study, we simply refer to the term 'chatbot' for all forms of dialogue systems. Such systems can be traced back more than 50 years to *ELIZA* [19], which simulates a psychotherapy session. Although the algorithm is very simple, and the bot responds with a limited, predefined set of answers, it was reported that users felt emotionally touched when interacting with the system. Other example systems followed, e.g. *ALICE* [13]. Currently, *Mitsuku*[2] appears to be one of the most successful chatbots, having won the *Loebner prize* multiple times (in 2013, 2016, 2017).[3]

A large share of *task-oriented* chatbots can be found in the field of e-commerce, where chatbots offer customer services. Ikea's chatbot 'Anna' informs customers about furniture sizes and delivery options. Royal Mail customers can use a dialogue interface to ask for updates of their current parcel status. In fact, rapid developments in AI over the last few years have manifested in a new dimension of possible features in chatbots [7]. In particular, the personalisation of chatbots is gaining momentum. Studies indicate that users prefer personalised chatbots with human-like character traits [1, 9]. Personalisation begins by simply giving a chatbot a name or adding a picture to the dialogue and goes as far as to reacting to user emotions. Kuligowska [9] studied the impact that different aspects of personalisation in e-commerce chatbots have on user experience and satisfaction. It turned out that adding a picture which represents the chatbot as a person – be it a human or a comic figure image – has a positive effect on users' experience and satisfaction. They found that personalisation of chatbots also affects the length of the

CHIIR '19, March 10–14, 2019, Glasgow, Scotland Uk

ACM ISBN 978-1-4503-6025-8/19/03…$15.00
https://doi.org/10.1145/3295750.3298956

[1] https://www.wired.com/story/google-duplex-phone-calls-ai-future/
[2] https://www.pandorabots.com/mitsuku/
[3] http://www.aisb.org.uk/events/loebner-prize

conversation and the user's choice of words. Generally, Kuligowska reported that personalised chatbots were rated higher than neutral ones. This offers a good starting point for further explorations.

A key way to personalise a chatbot is by altering its language style. This is the area that we are interested in. Language style transfer can be achieved through different methods ranging from rule-based, e.g. [20], to the application of deep neural networks, e.g. [2, 4, 6, 15, 16]. Io and Lee demonstrate research gaps for chatbots applied in commercial settings [7]. We use these findings as a motivation to start exploring how applying language style to e-commerce chatbots affects customer experience and their attitude towards a presented product.

The research question we address is *How does adding language style to an e-commerce chatbot affect user satisfaction, user interest in a product, perceived product value and user engagement*?

Perceived value in our context is defined as the price a customer is willing to pay for a product. It is assumed that these are indicators that can be correlated with profit.

3 METHOD

A ticket selling scenario was created where theatre tickets for a Shakespeare play were sold by two differently styled chatbots. One chatbot communicates in modern English (our baseline), whereas the second chatbot in a Shakespearean manner. The functionality of both chatbots was identical, however the responses were differently styled renditions of the same response sentence.[4]

Subjects were assigned one of the chatbots randomly and asked to use the chatbot service to book a ticket for a play at a fictional Shakespeare theatre. The booking of the ticket took place in an imaginary setup, so no real booking happened. Before and after the booking, the participants were asked to fill in a questionnaire. Figure 1 shows a screenshot of the implemented chatbot with a booking dialogue example shown from an interaction with the Shakespearean dialect version. Here is a specific example of the same response in two different styles:

> *Modern English*: 'Type the requested number of tickets again. Only the number, please. I am handsome, but a simple bot. Thanks for your understanding.'
> *Shakespearean-styled*: 'Could thou say again the ticket number only? Handsome I am, though but a simple servant bot. Pray take my gratitude for thy understanding.'

Both systems are based on a rule-based approach that uses regular expression matching and slot filling. A rule-based system was deemed easier to control than a trained system avoiding issues such as unnecessary language generation errors when trying to explore the link between independent variable (the language style) and dependent variables (user satisfaction, user engagement, user interest in a presented product, and perceived product value). The system nevertheless does allow some flexibility in the interaction. For example, if a user tries to involve the chatbot in casual conversation, the chatbot will go into *'chatmode'* and will defer the booking process until input is detected that is interpreted as a valid filler for one of the slots. Chatmode is initiated in two ways. Firstly, if the user enters a specific keyword or phrase unrelated to the booking process, for example "love", or "what is your name", the chatbot will reply with something related to the keyword. Secondly, if a user enters an input that cannot be matched by the slot filling three times in a row, it will respond by a Shakespeare quote. Chatmode stops when a slot-relevant input is entered.

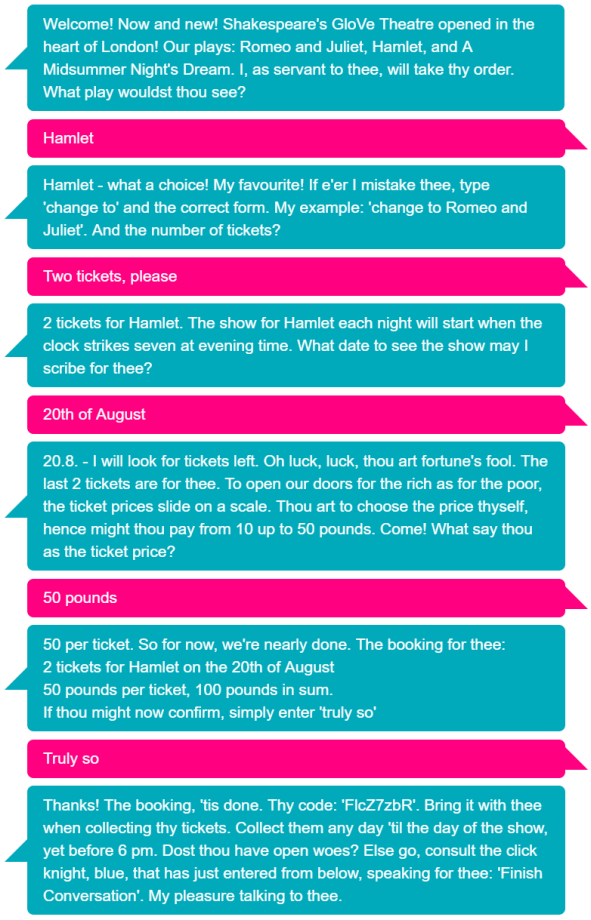

Figure 1: Example booking dialogue with Shakespearean-styled dialect chatbot version

3.1 Participants

169 subjects (111 female, 58 male) were recruited via the crowdsourcing platform *Figure Eight*.[5] Crowdsourcing falls under the category of *convenience sampling*, a form of *non-probability sampling* [12]. It is the most common sampling type in many research areas including IIR [10]. Convenience sampling generally does not exclude bias and therefore generalisations made on the basis of the results must be exercised with caution [10]. Users were allocated to one system at random and were unaware of the underlying research question.

[4] The conversion of style was performed manually by Ela Elsholz by consulting the online Shakespeare translation tool https://www.sparknotes.com/shakespeare/ and attempting to replicate the metre, rhythm and humour whilst preserving the meaning. Translations were proofread by an additional annotator before a final, agreed version being used for the chatbot.

[5] https://www.figure-eight.com/

Dependent variable	Mean (Var) *BotMod*	Mean (Var) *BotShake*
Overall user rating	6.07 (1.90)	5.61 (2.73)
Usability	6.35 (1.62)	6.09 (1.70)
Price willing to pay (in £)	35 (190)	41 (862)
Interest in visiting the theatre	5.82 (1.71)	5.46 (2.30)
Interest in visiting a Shakespeare play	5.62 (1.84)	5.30 (2.16)
Change of interest (post-task - pre-task)	0.35 (0.60)	0.18 (0.66)
Time with bot (min)	2.4 (1.53)	3.6 (8.08)
Non-task related inputs (per user)	0.16 (0.87)	1.01 (19.5)

Table 1: Mean and variance for all dependent variables (Interest in visiting the theatre and visiting a Shakespeare play are judgements obtained in the post-task questionnaire)

3.2 Questionnaires

There were two questionnaires, a pre-task questionnaire before the interaction with the chatbot system and a post-task questionnaire thereafter (see Appendix A for screenshots). The pre-task questionnaire captured demographic data about the user as well as their theatre habits. Additionally, a question was added about how high subjects estimate the ticket price for a professional theatre production in London to be. This question was added as a baseline to test that there were no significant differences between the two user groups, of which one group used only the modern chatbot and the other only the Shakespearean, before the chatbot experience.

For most of the questions a seven-point Likert-type scale – ranging from 1: 'not at all' to 7: 'very much so' – was used. One open question in form of a comment was included in the questionnaire to enhance a qualitative analysis of the user experience.

To test for differences between the two systems (the system using modern language, *BotMod*, and the system that uses Shakespearean style, *BotShake*) we applied significance tests to the results obtained for each dependent variable over the two user samples.

4 RESULTS

4.1 Quantitative Analysis

Table 1 illustrates the main results which we will now discuss in more detail (median scores left out for brevity).

User satisfaction. With a median score being greater or equal to 6 (7-point Likert-like scale), both bots show a high user rating of general bot experience and its usability[6]. We found that the modern chatbot had a higher user rating and reached a higher usability score (both at $p<0.05$, unpaired two-sample two-tailed Mann-Whitney U-test).

Perceived product value. Users stated how much they would be willing to pay for a ticket at the showcase theatre. The mean of this ticket price was £35 for the modern and £41 for the Shakespearean version although the difference is not statistically significant.

User interest in the product. Marginal differences were found regarding users' interest in visiting the showcase theatre between the users of the different bot versions, the interest in visiting a Shakespeare performance as well as the change in interest of visiting a Shakespeare performance from before to after using the bot.

User engagement. It was found that users spent more time with the Shakespearean chatbot version ($p<0.05$, unpaired two-sample two-tailed Mann-Whitney U-test). Also, more non-task related user inputs were recognized in conversations with this chatbot ($p<0.05$, unpaired two-sample two-tailed Mann-Whitney U-test).

Testing user groups for bias. The two user groups where investigated for bias in their theatre habits and price estimation of theatre tickets. This information was collected in a pre-task questionnaire, such that no experiment-related difference had occurred between the user groups so far. No significant difference can be found between the two user groups regarding their theatre habits or pre-task price estimation before using the chatbot.

4.2 Qualitative Analysis

The qualitative analysis was based on the user comments regarding the bot experience in the post-task questionnaire. Analysing all user comments, six general comment topics were categorised: 'works well', 'easy', 'difficult', 'fun', 'don't like bot', 'concept'. Ela Elsholz labelled each comment manually (without knowledge of which system produced this comment) to determine whether it can be assigned to one or more of these categories. It was then investigated whether there was a significant difference in frequency for each comment category between the two bots using z-tests. Table 2 gives an overview of the test results.

category	total	*BotMod*	*BotShake*	p-value
works well	19	12	7	0.25
easy	47	33	14	0.001 **
difficult	8	0	8	not applicable
fun	41	13	28	0.004 **
don't like bot	7	1	6	not applicable
concept	34	18	16	0.79

Table 2: Results for frequency of certain concepts in user comments (in some cases no z-test could be applied due to data sparsity).

It was found that users' comments regarding the modern-language bot version were more often related to the 'easy' category ($p<0.05$, unpaired two-sample two-tailed z-test), while comments on the Shakespearean-styled version were more frequently related to the 'fun' category ($p<0.05$, unpaired two-sample two-tailed z-test).

5 DISCUSSION

The purpose of this research was to investigate the effects of adding language style to e-commerce chatbots. The main focus has been to answer the research question: How does adding language style to an e-commerce chatbot affect *user satisfaction*, *user interest in a product*, *perceived product value* and *user engagement*?

By using the Shakespearean chatbot experiment, a difference in user satisfaction and user engagement was found between the

[6]Usability in the context of this research is the subjective rating given by the user of how easy the system was to use to perform the task.

test user groups, of which one used the modern and the other the Shakespearean chatbot. User satisfaction was found to be higher for the modern bot, whereas user engagement turned out to be higher for the Shakespeare bot.

The results of the qualitative analysis showed that the modern bot was described by more users to be *'easy to use'*, while the Shakespearean chatbot was more often described as *'fun to use'*. Also, the usability was rated higher for the modern chatbot implying that the test users preferred a fast and easy booking over an entertaining one; however, test users were recruited through the platform *Figure Eight* and participated in the experiment to earn money. Moreover, they were paid per experiment, not per time. Therefore *Figure Eight* users might show a stronger tendency to prefer a fast and easy process than an average customer in a daily life setting. Further experiments would be needed to see whether this preference can be generalised to a larger population.

Although the Shakespearean chatbot achieved a lower user rating, the ticket price users stated they would be willing to pay (as a proxy for user interest in the product) was on average higher. Although the difference was marginal, this can be interpreted as an indication that a more elaborate presentation might give the impression a product is worth more and hence users might be willing to pay more.

The user engagement was evaluated by measuring the time spent with the bot and counting non-task related user inputs in a conversation. That Shakespearean bot users spent more time in the booking dialogue can also be traced back to the fact that the chatbot responses were harder to understand as the qualitative analysis showed that the Shakespearean bot was more often described as difficult to use. However, more users of the Shakespearean bot group engaged in non-task related inputs than in the modern bot group. This can be interpreted as evidence that the Shakespearean styling of the bot increased users' interest in engaging the chatbot in conversation. In other words, users were more engaged with the bot in the conversation. Surprisingly few conversations contained non-task related user inputs (20 out of 169). While designing the bot, it was expected that more users would try to challenge the bot or experiment with how it reacts to different inputs. That this was not the case could possibly be traced back to the fact that *Figure Eight* users took part in the experiment to earn money. Also, the design of the bot could play a role as the questions provided by the bot to fill the slots were very straightforward and might not have invited users to engage in a more open dialogue. Finally, the test users were informed before the experiment that their conversation was to be recorded. They might, therefore, have been eager to complete the experiment properly and did not want to risk making a mistake by entering non-task relevant inputs. Again, to find out whether the results in user engagement can be generalised over a larger population further experiments with other test user groups would be necessary.

In addition to this, the results were found using an artificial booking setup. It can therefore only be speculated how far those results could be reproduced in a real world booking scenario. The setup also contained elements that evidently confused the test users, like the self-assessment of ticket prices and lack of information about seating. Also, some users had to repeat the date several times, as shows were booked out for their inputs or, in some cases, because their input date format was not recognised.

In the study, only one particular language style was investigated. It might be more a question of whether users like a Shakespearean dialect or find it annoying, artificial or hard to understand. The experimental results might be very different when adding another style. In addition, the findings from using the simple user interface of the experiment may not necessarily be generalisable to other dialogue systems, where the complexity of the task and range of potential responses mask the benefit of a stylised responder.

It can also be said that our baseline system, the modern chatbot, can be regarded as having a language style itself. It talked in a lively fashion, showing humour and tried to personally engage with the user. We deliberately decided to choose this as a baseline rather than a neutral version (aka an NLP interface to a database) based on Kuligowska's findings that personalised e-commerce chatbots were preferred over neutral ones [9].

There are many other issues arising from this line of research. For example, Google's recently published chatbot service[7] that adopts human behaviour by adding common speech pausing into the conversation raises ethical questions about whether chatbots are becoming too human, opening up discussion over whether it must be marked clearly that one is talking to a human or using an automatic interface. Another highly debated application of chatbot that has emerged in recent years is the use of chatbots in social media like on Facebook and Twitter – also called *Twitterbots*. These chatbots deliver advertising or purposefully influence political debates.[8] [9] There is certainly plenty of scope for further research.

6 CONCLUSION

We explored how adding language style to an e-commerce chatbot might affect overall user satisfaction, perceived product value, user interest in a product, and user engagement with the service. To do this we constructed a specific scenario of a ticket selling application. We found that adding Shakespearean language to an e-commerce chatbot increased user engagement (significantly) as well as perceived product value (though in this case only marginally). At the same time user satisfaction decreased. It was also found that the Shakespearean bot was more frequently described as *'fun'* in user comments, while the modern bot was referred to as *'easy'*. Those findings imply that in some e-commerce settings, it could be useful to accept a loss in user satisfaction to increase the perceived product value and user engagement.

Independent of which language style was used, both bot versions resulted in very high user acceptance and user satisfaction. Considering that both chatbot versions contained some degree of personality and language style, it may be promising to further investigate means of automatic style transfer. Applications of automatic style transfer are broad, and include adapting to the language styles of users. With the growing number of chatbots being applied to household devices, their personalisation will likely become more important.

[7] https://www.wired.com/story/google-duplex-phone-calls-ai-future/
[8] https://www.politico.com/story/2016/10/chatbots-are-invading-politics-229598
[9] https://politibot.io/this-is-how-politibot-platform-works/

REFERENCES

[1] R. E. Banchs. 2017. On the construction of more human-like chatbots: Affect and emotion analysis of movie dialogue data. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 1364–1367. https://doi.org/10.1109/APSIPA.2017.8282245

[2] K. Carlson, A. Riddell, and D. N. Rockmore. 2017. Zero-Shot Style Transfer in Text Using Recurrent Neural Networks. *CoRR* abs/1711.04731 (2017). arXiv:1711.04731 http://arxiv.org/abs/1711.04731

[3] A. De Roeck, U. Kruschwitz, P. Neal, P. Scott, S. Steel, R. Turner, and N. Webb. 1998. YPA - an intelligent directory enquiry assistant. *BT Technology Journal* 16, 3 (1998), 145–155.

[4] Z. Fu, X. Tan, N. Peng, D. Zhao, and R. Yan. 2017. Style Transfer in Text: Exploration and Evaluation. *CoRR* abs/1711.06861 (2017). arXiv:1711.06861 http://arxiv.org/abs/1711.06861

[5] J. Glass, G. Flammia, D. Goodine, M. Phillips, J. Polifroni, S. Sakai, S. Seneff, and V. Zue. 1995. Multilingual spoken-language understanding in the MIT Voyager system. *Speech Communication* 17, 1 (1995), 1 – 18.

[6] M. Han, O. Wu, and Z. Niu. 2018. Unsupervised Automatic Text Style Transfer Using LSTM. In *Natural Language Processing and Chinese Computing*, Xuanjing Huang, Jing Jiang, Dongyan Zhao, Yansong Feng, and Yu Hong (Eds.). Springer International Publishing, Cham, 281–292.

[7] H. N. Io and C. B. Lee. 2017. Chatbots and conversational agents: A bibliometric analysis. In *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*. 215–219. https://doi.org/10.1109/IEEM.2017.8289883

[8] H. Joho, L. Cavedon, J. Arguello, M. Shokouhi, and F. Radlinski. 2017. First International Workshop on Conversational Approaches to Information Retrieval (CAIR'17). In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. ACM, New York, NY, USA, 1423–1424.

[9] K. Kuligowska. 2015. Commercial Chatbot: Performance Evaluation, Usability Metrics and Quality Standards of Embodied Converstational Agents. *Professionals Center for Business Research* 2 (2015). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2569637

[10] D. Kelly. 2009. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends in Information Retrieval* 3, 1–2 (2009), 1–224. https://doi.org/10.1561/1500000012

[11] F. Radlinski and N. Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17)*. ACM, 117–126.

[12] M. Schonlau, R. D. Fricker, and M. N. Elliott. 2002. *Conducting Research Surveys via E-mail and the Web*. Santa Monica, CA: RAND Corporation.

[13] B.A. Shawar and E. Atwell. 2002. *A comparison between Alice and Elizabeth chatbot systems*. University of Leeds, School of Computing research report 2002.19. http://eprints.whiterose.ac.uk/81930/ Shawar, BA and Atwell, E (c) 2002, University of Leeds. Reproduced with permission from the copyright holders.

[14] B. A. Shawar and E. Atwell. 2007. Different Measurements Metrics to Evaluate a Chatbot System. In *Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies (NAACL-HLT-Dialog '07)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 89–96. http://dl.acm.org/citation.cfm?id=1556328.1556341

[15] B. A. Shawar and E. Atwell. 2016. Stylistic Transfer in Natural Language Generation Systems Using Recurrent Neural Networks. In *Proceedings of EMNLP 2016 Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods*. Association for Computational Linguistics, 43–47.

[16] T. Shen, T. Lei, R. Barzilay, and T. Jaakkola. 2017. Style Transfer from Non-Parallel Text by Cross-Alignment. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 6830–6841. http://papers.nips.cc/paper/7259-style-transfer-from-non-parallel-text-by-cross-alignment.pdf

[17] A. Shevat. 2017. *Designing Bots*. O'Reilly.

[18] P. Thomas, M. Czerwinski, D. McDuff, N. Craswell, and G. Mark. 2018. Style and Alignment in Information-Seeking Conversation. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18)*. ACM, New York, NY, USA, 42–51.

[19] J. Weizenbaum. 1983. ELIZA - A Computer Program For the Study of Natural Language Communication Between Man And Machine (Reprint). *Commun. ACM* 26 (01 1983), 23–28.

[20] W. Xu, A. Ritter, W. B. Dolan, R. Grishman, and C. Cherry. 2012. Paraphrasing for style. In *24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers*. 2899–2914.

A APPENDIX - QUESTIONNAIRES

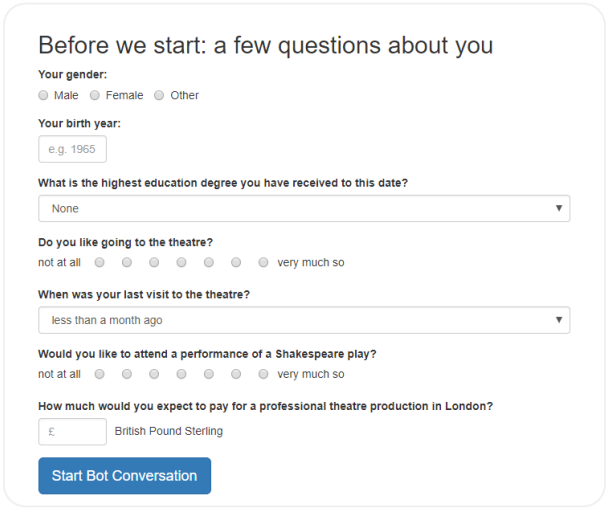

Figure 2: Pre-task questionnaire.

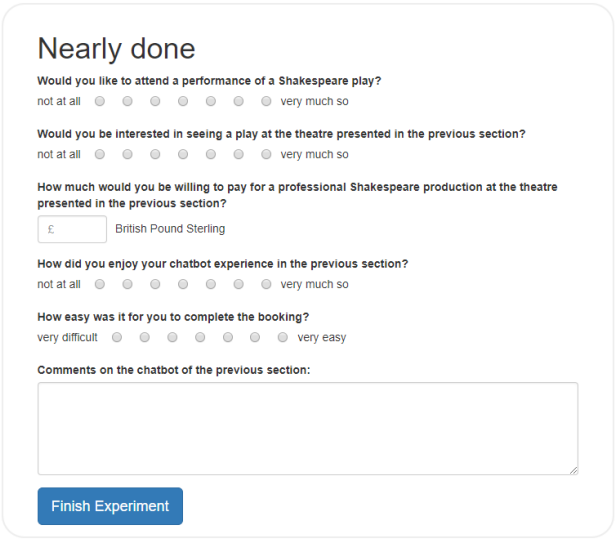

Figure 3: Post-task questionnaire.