SHOULD MY CHATBOT BE REGISTER-SPECIFIC? DESIGNING APPROPRIATE UTTERANCES FOR TOURISM INTERACTIONS

By Ana Paula Chaves Steinmacher

A Dissertation

Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

in Informatics & Computing

Northern Arizona University

November 2020

Approved:

Marco Aurelio Gerosa, Ph.D., Chair

Morgan Vigil-Hayes, Ph.D.

Eck Doerry, Ph.D.

Jesse Egbert, Ph.D.

ABSTRACT

SHOULD MY CHATBOT BE REGISTER-SPECIFIC? DESIGNING APPROPRIATE UTTERANCES FOR TOURISM INTERACTIONS

ANA PAULA CHAVES STEINMACHER

Chatbots are often designed to mimic social roles attributed to humans. However, little is known about the impact on user's perceptions of using language that fails to conform to the associated social role. This research draws on sociolinguistic theory to investigate how a chatbot's language choices can adhere to the expected social role the agent performs within a given context; i.e., we seek to understand whether a chatbot's design should account for linguistic register, which is the language variety associated with a particular situation of use. This research focuses on investigating whether register-specific language influences users' perceptions and experiences with chatbots. We produced parallel corpora of conversations in the tourism domain with similar content and varying register characteristics and evaluated users' preferences of a chatbot's linguistic choices in terms of appropriateness, credibility, and user experience as well as the chatbot's social presence and anthropomorphism. Results revealed significant associations between certain linguistic features present in utterances within the parallel corpora and user perceptions of appropriateness, credibility, and overall user experience. Moreover, the results showed that the linguistic features are a stronger predictor of this association than the variables representing individual biases. We also confirmed that using appropriate language positively influences the chatbot's social presence, which, in turn, positively influences anthropomorphism. We discuss how high levels of anthropomorphism may have a negative impact on the overall user experience. Findings strongly suggests that attention to appropriate register is an important factor for the perceived quality of chatbot conver-

sations and, therefore, critical to future chatbots' success. Although this study focused on the tourism domain, we expect these outcomes to be applicable to other interactions that share similar situational parameters. More generally, this study demonstrates that the theoretical and empirical foundation of register analysis can be an effective tool for characterizing the conversational register used in other target domains, and can systematically expose the specific linguistic features within conversational utterances that most strongly impact user perceptions of the interaction.

ACKNOWLEDGEMENTS

This dissertation is the result of a long, exciting journey. I would like to take a moment to show my gratitude to those who are part of this achievement.

First of all, I would like to express my sincere appreciation to the Federal University of Technology-Parana (UTFPR) and my colleagues at the Academic Department of Computing for allowing me the opportunity to be away from my duties to pursue my Ph.D.

I am extremely grateful to my advisor, Dr. Marco Aurelio Gerosa, for all the guidance, support, and encouragement over these years. Thank you for trusting my work and for always being there for me, inside and outside of NAU. I also extend my gratitude to the committee members whose valuable contributions helped improve and shape this work.

The completion of my dissertation would not be possible without the support of the Flagstaff Convention & Visitors Bureau–represented by Mr. Trace Ward, the Flagstaff Visitor Center, and the tourist assistant professionals who worked with us in the data collection. Their contributions with time, advice, resources, and experience in the tourism business helped make this work come true.

Pursuing a Ph.D. is an everyday learning and discovering process. Several people were part of my educational growth, including teachers and SICCS colleagues. I cannot leave NAU without mentioning Dr. Jesse Egbert, who patiently walked me through sociolinguistics when I knew nothing about it. Thank you for the many meetings, author recommendations, text reviews, lectures, and all the knowledge about linguistics you passed on to me. Also, Dr. Toby Hocking straightened the path to machine learning, devoting time and effort to assist me with model selection, data visualization, and interpretation of statistical results. Finally, I thank the Statistical Consulting Lab team, represented by Dr. Roy St. Laurent, for the helpful advice and suggestions on data collection and quantitative

analysis.

I gratefully acknowledge the financial assistance I received through both Graduate Teaching (2017-2019) and Research (2019-2020) assistantships. As a GTA, I am profoundly grateful to SICCS for the opportunity to work as a Capstone mentor, supervised by Dr. Eck Doerry, which was an inestimable professional experience. I extend my gratitude to Professor Steven Jacobs and Dr. Frédéric Loulergue, my GTA supervisors, for guidance and assistance.

Particularly helpful to me during this time were Caitlin Abuel and Tyler Conger, NAU CS undergraduate students, whose hard work greatly facilitated the recruitment and qualitative data collection during the lab sessions, the text modification, the chatbots and user study's page development, as well as data collection during online studies. I am incredibly thankful for all the valuable work you have done as members of our research team.

I am deeply indebted to all our anonymous research participants (from both online and lab sessions) for their dedication to the performed tasks. I also wish to thank the support I received from Prolific representatives whenever requested.

Finally, I do not have words to express my gratitude and love to the most important people in my life: Igor, Alice, and Dante. Thank you for crossing the seas to live my dream; for suffering my pains; for comforting, supporting, and caring even when you were the ones who needed care. Thank you for understanding my absences and tiredness, and for all the morning smiles that give me purpose. I love you and I always will.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

To Adriana Mendes Polato–a linguist, a friend, a warrior.

Chapter 1

INTRODUCTION

Chatbots are "computer programs that interact with users using natural language" (Shawar and Atwell, 2007) [1] . The origin of the concept dates back to 1950 when Alan Turing proposed the Imitation Game ("Can machines think?") (Turing, 1950). Since then, making computers that interact with humans through conversation has been a challenge for researchers (Vinciarelli *et al.*, 2015; Zue and Glass, 2000). ELIZA (Weizenbaum, 1966) was the first software to play the Imitation Game, followed by other early technologies such as TinyMud (Mauldin, 1994), SHRDLU (Winograd, 1971), and A.L.I.C.E. (Wallace, 2009) [2] . The main goal of these chatbots was to mimic human conversations. Over the years, this technology has evolved. Advances in artificial intelligence (Goertzel, 2014), natural language processing (Hirschberg and Manning, 2015), and cognitive systems (Noor, 2015) have boosted the adoption of chatbots.

A recent report on the chatbot market (Grand View Research, 2017) attests to their increasing demand, predicting a global chatbot market of USD 1.25 billion by 2025. The BotList [3] website indexes thousands of chatbots for education, entertainment, games, health, productivity, travel, fun, and several other categories. At the 2018 F8 conference, Facebook announced having 300K chatbots active on Facebook Messenger (Boiteux, 2019). Chatbots are changing how companies engage with their customers (Brandtzaeg and

[1] In this research, we specialize the term *chatbot* to refer to *a disembodied conversational agent that held a natural language conversation via a text-based environment, such as through an instant messenger tool.*

[2] A list of all the acronyms and abbreviations can be found in Appendix A

[3] https://botlist.co

Følstad, 2018; Gnewuch *et al.*, 2017), how students participate in their learning groups (Hayashi, 2015; Tegos *et al.*, 2016a), and how patients self-monitor the progress of their treatment (Fitzpatrick *et al.*, 2017), among many other applications. Some scholars (Luger and Sellen, 2016) claim that conversation is the next natural form of human-computer interactions (HCI).

The increasing interest in chatbot technologies has brought new challenges for the HCI field (Brandtzaeg and Følstad, 2018; Følstad and Brandtzæg, 2017; Neururer *et al.*, 2018), and despite this growing popularity, users remain unsatisfied with their experiences with chatbots (Kiseleva *et al.*, 2016; Luger and Sellen, 2016), which may affect their behavior towards the technology (Komatsu *et al.*, 2012). Whereas traditional user interfaces apply visual elements such as buttons, menus, or hyperlinks to communicate with users, conversational interfaces rely almost entirely on language as the primary resource to achieve communicative goals. Therefore, developing a more comprehensive understanding of the linguistic design of chatbot conversations and its effects on user perceptions is critical to the success of chatbot technologies.

To date, language design for chatbots has focused primarily on ensuring that chatbots produce coherent and grammatically correct responses, and on improving functional performance and accuracy (see e.g. (Jiang and E Banchs, 2017; Maslowski *et al.*, 2017; Massaro *et al.*, 1999; Zdravkova, 2000)). Although current chatbots may, at some functional level, provide users with the answers they seek, the utterances portray arbitrary patterns of language that often fail to take into account the interactional situation in choosing a proper conversational tone for the interaction. For example, in Duijst (2017), users complained that the financial advisor chatbot used emojis combined with a formal language in a situation of urgency (a stolen bank card). On the other hand, using appropriate linguis-

tic choices potentially increases human-likeness (Gnewuch *et al.*, 2017; Hill *et al.*, 2015; Jenkins *et al.*, 2007) and believability (Jenkins *et al.*, 2007; Morris, 2002; Morrissey and Kirakowski, 2013; Tallyn *et al.*, 2018), as well as enhancing the overall quality of the interaction (Jakic *et al.*, 2017). Currently, the design of a particular chatbot's linguistic choices is often based on ad-hoc analyses of user characteristics or the chatbot's persona. For machine language generation, models are trained using available corpora in the target domain, but they do not consider the particular context of the corpora's conversations.

Little is known about the effect of these design decisions on user perceptions, much less about how to tailor chatbot design to the particular situation of use. At the same time, empirical studies have repeatedly demonstrated that a chatbot's linguistic choices influence user perceptions and behavior toward chatbots (Elsholz *et al.*, 2019; Araujo, 2018; Tariverdiyeva, 2019). Previous research on chatbot design suggests that when chatbots misuse language (e.g., conveying excessive (in)formality or using incoherent style), the conversation sounds strange to the user, and leads to frustration (Duijst, 2017; Kirakowski *et al.*, 2009; Mairesse and Walker, 2009). Using appropriate linguistic choices potentially increases human-likeness (Hill *et al.*, 2015; Jenkins *et al.*, 2007; Gnewuch *et al.*, 2017) and believability (Jenkins *et al.*, 2007; Morris, 2002; Morrissey and Kirakowski, 2013; Tallyn *et al.*, 2018), as well as enhancing the overall perception of the quality of the interaction (Jakic *et al.*, 2017). Developing a strong basis for designing not just what a chatbot says but also how it says it must be a priority for creating the next generation of chatbots. This research establishes a framework for analyzing the effect of linguistic choices on user perceptions, and takes a first step toward developing a prescriptive basis for tailoring chatbot linguistic choices to specific interactional situations.

Humans have developed a highly-refined sense of how to adapt their tone, idioms, and

formulations to various conversational contexts (Niederhoffer and Pennebaker, 2002). We refer to sociolinguistics to understand how language variation happens in human-human communication and how to evaluate whether the same linguistic behavior could apply to chatbots. A sociolinguistic theory (Biber and Conrad, 2019) states that each utterance in a conversation reflects the immediate situation of interaction, which is formed through the relationship between the theme, interlocutors, and immediate context (Bakhtin, 2010). Varying contexts result in different patterns of language and ultimately affect a speaker's linguistic choices (Kilgarriff, 2005; Kamberelis, 1995) that are used to accomplish social purposes. The "language variety associated with a particular situation of use" is called **register** (Biber and Conrad, 2019)–a concept that has emerged as one of the most important predictors of linguistic variation in human-human communication (Biber, 2012). Register theory establishes, for example, that the core linguistic features (e.g., personal pronouns, verb tenses, adverbs) presented in this dissertation differ from those one would use to discuss this same topic for a lecture or while texting a friend.

Given the importance of register in human-human communication (Biber, 2012) and the human tendency to respond to computers as social actors (Nass and Moon, 2000; Nass *et al.*, 1994), it is reasonable to assume that when people talk to a chatbot, they unconsciously expect coherence between a chatbot's utterance and an utterance that a human whom the agent represents might form in that interactional situation (Gnewuch *et al.*, 2017). Hence, we expect that chatbots that use register-appropriate language are more likely to meet the user's expectations, and be recognized by human-interlocutors for the role they stand in. However, the potential applicability of register has not yet been explored in the context of chatbot language design.

Argamon (2019) suggests that the sociolinguistic concept of register could be formal-

ized to provide a theoretical basis for machine language generation. To achieve that, chatbots would need to be enriched with computational models that can evaluate the conversational situation and adapt the chatbot's linguistic choices to conform with the expected register, in an effort to mimic the subconscious humans' language production process. However, before investing effort into developing algorithms to adapt a chatbot's language to a particular register, it is crucial to understand how register variation influences user experiences. To fill this gap, this research focuses on investigating the extent to which the register variation plays a role in shaping the user perception of the human-chatbot interaction. Since register is defined in terms of situational contexts and communicative purposes, we selected a domain in which a chatbot has a social role: we choose tourism as the application domain due to the adoption of chatbots to support travel decisions (Garrido *et al.*, 2017; Sano *et al.*, 2018; Ivanov and Webster, 2017; Niculescu *et al.*, 2014) and the relevance of the tourism industry to the economy of the city where this research was conducted, namely Flagstaff, AZ (Thomas Combrink, 2018). The chatbot in this research performs the role of a tourist assistant, helping tourists to search for information about a destination during a trip.

This research aims to investigate whether and to what extent chatbot utterances should conform to registers, such that they cohere with their expected social role. Since register is the main predictor of language variation in human-human communication (Biber, 2012), we hypothesize that *chatbot utterances that are register-appropriate will have a positive effect on user experience.* In this research, user experience is defined in terms of attitudinal metrics, such as credibility (Zumstein and Hundertmark, 2017), anthropomorphism, and social presence (Katkute *et al.*, 2017) as well as perceived quality and appropriateness (Jakic *et al.*, 2017). To evaluate this hypothesis, we defined the following research question:

RQ: *How does a chatbot's use of register-appropriate language affect the user perceptions and experiences with chatbots?*

To answer this question, this research brings together techniques, content, and concepts spanning three disciplines, namely: chatbot design, sociolinguistics, and tourism. The research method includes analyzing the register characteristics of two corpora of conversations (FLG and $DailyDialog$) to characterize their register differences and producing a parallel corpus (FLG_{mod}) in the tourism domain, which has similar informational content as FLG, but varies in language use patterns (register) used to convey that content. Then, we performed user studies to evaluate user's preferences in terms of appropriateness, credibility, and user experience. The results show that there is an association between linguistic register and user perceptions of the interaction, and that register is a stronger predictor of this association than other variables of individual biases (participants, their social agent orientation, and answers' authors). These outcomes have important implications for the design of chatbots, e.g., the need to design chatbots to generate register-specific language for hard-coded and dynamically generated utterances to improve acceptance and user perceptions of chatbot interactions.

Ultimately, the main contributions of this research are the assessment of the influence of register on user experiences and the formalization and adaptation of register analysis as a technique to purposefully design chatbot utterances not only in the tourism domain, but also for other task-oriented domains where chatbots would benefit from using context-

specific language.

1.1 Research Design

This research mixes quantitative and qualitative methods. As outlined earlier, this research aims to explore the extent to which user experience (in terms of perceived appropriateness, credibility, and overall user experience) is related to the conversational register used by a chatbot. For this purpose, we compared conversations expressed in different registers, presenting them to users for evaluation. To isolate the effect of register on perceived user experience, we compared conversations that are equivalent in content but vary in language patterns.

Finding such parallel data—natural language texts with the same semantic content, but expressed in different forms (Nevill and Bell, 1992)—is difficult. Previous studies requiring such parallel data have typically used written texts with multiple versions, e.g., versions of the Bible or Shakespearean texts in the original and modern language forms (see (Tikhonov and Yamshchikov, 2018)). Although perhaps useful for the analysis in an abstract context of NLP research, these corpora portray archaic language centered around topics not likely to be relevant to most modern chatbot users.

The approach presented here, therefore, was based on the production of a parallel corpus. This corpus was based on actual conversations, which were carefully manipulated based on register theory to produce conversations of equivalent content, but in differing registers. Unlike previous studies that focus on style (Elsholz *et al.*, 2019; Hoegen *et al.*, 2019; Tariverdiyeva, 2019) (i.e., preferences associated with authors or historical period), we relied on register theory to identify and reproduce language variations that would be plausible for a tourist assistant chatbot to use. Moreover, developing a concrete basis for

explicitly manipulating conversational register in the design of chatbot language requires an explicit characterization of the register. Therefore, we identified a set of linguistic features that, together, characterize the register and show how varying these features affect user perceptions of conversational quality. Figure 1.1 depicts the proposed method; each of the steps is summarized below and detailed in the following chapters [4] .

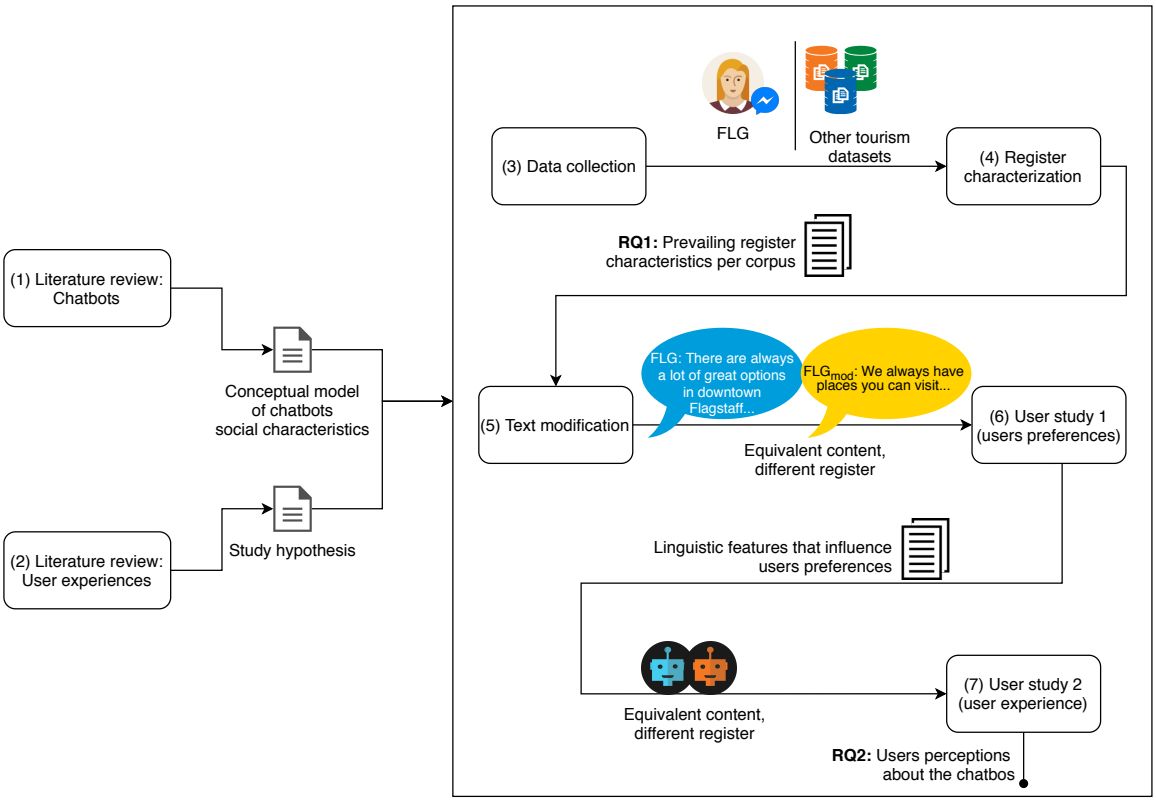


Figure 1.1: Overview of the research method. After literature review and hypothesis definition, the main method consists of seven steps, and the outcomes of one step is seeded into the next one.

Step 1–Literature review: we performed a qualitative analysis of the literature on the social aspects that influence human-chatbots interactions. As the main outcome, we developed a conceptual model of chatbot social characteristics. Additionally, this step pro-

[4] A summary of this research was presented at ACM CHI Doctoral Consortium (Chaves, 2020b) in May 2020.

vided evidence of language use as a source of frustration when interacting with chatbots. The outcomes of this step are summarized in Chapter 2 [5] .

Step 2–User experiences: we referred to the literature on chatbots and human-human communication to identify what constructs could potentially be impacted by the patterns of language in human-chatbots interactions. Based on that, we derived a set of hypotheses that we evaluated in step 7 (user study 2). The outcomes of step 2 are presented in Chapter 6.

Step 3–Data collection: to provide a foundation for the analysis, we collected conversations of human domain experts (tourist assistants) interacting with tourists in a text-based tourist information search scenario; we refer to this as the FLG corpus. Because the conversational register is characterized by comparing linguistic expression in varying interactional situations, we also selected another corpus of conversations in the tourism domain that is available online and is commonly used in natural language research, namely $DailyDialog$ (Li *et al.*, 2017). Conversations in this corpus span a random variety of daily-life topics from ordinary life to politics, health, tourism, and other topics. Details about the corpora collection are provided in Chapter 3.

Step 4–Register characterization: the next aim was to characterize the conversational registers present in the two corpora. Based on a broad set of linguistic features that have been previously identified as relevant for characterizing conversational register (Biber, 1988), we performed register analysis of the FLG and $DailyDialog$ corpora individually, and then statistically compared the patterns of language between them, similar to the analysis performed by Chaves *et al.* (2019a). The register characterization step

[5]The literature review has been accepted for publication in the International Journal of Human-Computer Interaction (IJHCI) (to appear) Chaves and Gerosa (2020).

is detailed in Chapter 4.

Step 5–Text modification: having identified discrete register variations present in the two corpora, the focus shifted to using these register characterizations to produce a parallel corpus in which conversations had equivalent information content, but used a different linguistic format. Specifically, for every answer provided by a tourist assistant in the FLG corpus, we performed linguistic modifications to produce a new corresponding answer that portrays a language pattern that mimics the register characteristics from the $DailyDialog$ corpus; we call this produced parallel corpus FLG_{mod}. To assess whether the modified answers in FLG_{mod} preserved the informational content of the original, we performed a study to validate the text modification. We invited participants to compare the parallel answers in terms of naturalness and content preservation. Chapter 5 details the text modification and validation. After performing these foundational steps, we ended up with two parallel corpora (FLG and FLG_{mod}).

Step 6–User study 1 (user preferences): after developing the parallel corpora that differ solely in the portrayed register, we performed a study to reveal whether users perceived register variations and, if so, what linguistic variations within a register characterization appear to have the greatest impact on aspects of user experience. Overall, we expected to find a preference for the original answers from the FLG corpus, since these are register-specific language produced by humans. To perform the analysis, we selected a subset of tourist questions and their corresponding answers from both FLG and FLG_{mod}. Participants were presented with these individual question-answer exchanges and, for each, were asked to choose which answer they preferred based on three distinct measures of quality: appropriateness, credibility, and overall experience. Then, we fitted a statistical learning model to identify the linguistic features that best predict user choices

[6] . This step is detailed in Chapter 6.

Step 7–User study 2 (user experiences): given the user preferences identified in Step 6, my next goal was to identify whether the previous inferences still stand when the user experiences the interaction with a chatbot that portrays the expected linguistic register, instead of examining conversation excerpts side-by-side. Additionally, we wanted to assess the appropriateness of language as a predictor of chatbot's credibility, social presence, anthropomorphism and overall user satisfaction. We developed two chatbots, one of which uses the register that was consistently preferred by the participants of user study 1 and one of which uses the paired responses. In this experiment, we expected participants to consistently rate the benchmark chatbot as the one that provides the best user experiences. Chapter 7 details this study.

1.2 Publications

The research presented in this dissertation resulted in the following publications:

- CHAVES, Ana Paula; EGBERT, Jesse; GEROSA, Marco Aurelio. 2019. Chatting like a robot: the relationship between linguistic choices and users' experiences. In: *ACM CHI Workshop on Conversational Agents: Acting on the Wave of Research and Development*, Glasgow, UK (Chaves *et al.*, 2019b).

- CHAVES, Ana Paula; DOERRY, Eck; EGBERT, Jesse; GEROSA, Marco Aurelio. 2019. It's How You Say It: Identifying Appropriate Register for Chatbot Language Design. In: *Proceedings of the 7th International Conference on Human-Agent Interaction (HAI '19)*. Association for Computing Machinery, New York, NY, USA, 102–109 (Chaves

[6] A paper that reports these results is under review for publication in the ACM Transactions on Computer-Human Interaction (TOCHI) Chaves *et al.* (2021).

et al., 2019a).

- CHAVES, Ana Paula. 2020. Should my Chatbot be Register-Specific? Designing Appropriate Utterances for Tourism. In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)–Doctoral Consortium.* Association for Computing Machinery, New York, NY, USA, 1–11 (Chaves, 2020b).

- CHAVES, Ana Paula; GEROSA, Marco Aurelio. 2020. How should my chatbot interact? A survey on human-chatbot interaction design. In: *International Journal of Human-Computer Interaction.* DOI: 10.1080/10447318.2020.1841438 (Chaves and Gerosa, 2020).

- CHAVES, Ana Paula; HOCKING, Toby; EGBERT, Jesse; DOERRY, Eck; GEROSA, Marco Aurelio. 2020. Chatbots language design: the influence of language use on user experience. *Under review.*

I have also collaborated in a research paper with researchers in chatbots for Software Engineering (Wessel *et al.*, 2018) and served the program committee of the CONVERSATIONS workshop for two years (2019 and 2020).

Chapter 2

LITERATURE REVIEW

The motivation behind this research is the claim that people react to computers as social actors (Nass and Moon, 2000), and the more human-like a computer representation is, the more people tend to respond to them socially (Gong, 2008). Chatbots are, by definition, designed to have at least one human-like trait: the ability to interact in (human) natural language. Hence, making a chatbot acceptable to users is primarily a social, not only technical, problem to solve (Neururer *et al.*, 2018). The first step of this research was to understand a chatbot's social characteristics and its impact on user interaction. To accomplish that, we performed a survey of the literature that explores desired social characteristics for chatbots and, as a result, we derived a conceptual model of social characteristics, which is presented in Section 2.1.

The results of the survey highlight challenges to designing chatbot's language and the potential implications of language choices on the user experience, which motivated the next steps. In the subsequent sections, we summarize the literature on chatbot language design (Section 2.3) and the rationale for applying sociolinguistic register analysis as the theoretical foundation for this research (Section 2.4). Finally, in Section 2.5, we discuss the choice of the tourism domain, specifically information search interactions in tourism

contexts, as a testbed for this research.

2.1 Social characteristics of chatbots

The literature survey on chatbot social characteristics includes 56 studies in a variety of domains. Appendix B presents an overview of the study selection process and a summary of the surveyed literature. A detailed report of the results can be found in Chaves and Gerosa (2020). This section describes the identified social characteristics grouped into categories, as depicted in Table 2.1. We also grouped the social characteristics based on the domain in which they were investigated (see Table 2.2).

Table 2.1: Conceptual model of chatbots social characteristics

	Social Characteristics	Benefits	Challenges	Strategies
Conversational Intelligence	*Proactivity*	**[B1]** to provide additional information **[B2]** to inspire users and to keep the conversation alive **[B3]** to recover from a failure **[B4]** to improve conversation productivity **[B5]** to guide and engage users	**[C1]** timing and relevance **[C2]** privacy **[C3]** user perception of being controlled	**[S1]** to leverage conversational context **[S2]** to select a topic randomly
	Conscientiousness	**[B1]** to keep the conversation on track **[B2]** to demonstrate understanding **[B3]** to hold a continuous conversation	**[C1]** to handle task complexity **[C2]** to harden the conversation **[C3]** to keep the user aware of the chatbot's context	[S1] conversational flow **[S2]** visual elements **[S3]** confirmation messages
	Communicability	**[B1]** to unveil functionalities **[B2]** to manage the user's expectations	**[C1]** to provide business integration **[C2]** to keep visual elements consistent with textual inputs	**[S1]** to clarify the purpose of the chatbot **[S2]** to advertise the functionality and suggest the next step **[S3]** to provide a help functionality
Social Intelligence	*Damage control*	**[B1]** to appropriately respond to harassment **[B2]** to deal with testing **[B3]** to deal with lack of knowledge	**[C1]** to deal with unfriendly users **[C2]** to identify abusive utterances **[C3]** to balance emotional reactions	**[S1]** emotional reactions **[S2]** authoritative reactions **[S3]** to ignore the user's utterance and change the topic **[S4]** *conscientiousness* and *communicability* **[S5]** to predict the user's satisfaction
	Thoroughness	**[B1]** to adapt the language dynamically **[B2]** to exhibit believable behavior	**[C1]** to decide on how much to talk **[C2]** to be consistent	Not identified
	Manners	**[B1]** to increase human-likeness	**[C1]** to deal with face-threatening acts **[C2]** to end a conversation gracefully	**[S1]** to engage in small talk **[S2]** to adhere turn-taking protocols
	Moral agency	**[B1]** to avoid stereotyping **[B2]** to enrich interpersonal relationships	**[C1]** to avoid alienation **[C2]** to build unbiased training data and algorithms	Not identified
	Emotional intelligence	**[B1]** to enrich interpersonal relationships **[B2]** to increase engagement **[B3]** to increase believability	**[C1]** to regulate affective reactions	**[S1]** to use social-emotional utterances **[S2]** to manifest *conscientiousness* **[S3]** reciprocity and self-disclosure
	Personalization	**[B1]** to enrich interpersonal relationships **[B2]** to provide unique services **[B3]** to reduce interactional breakdowns	**[C1]** privacy	**[S1]** to learn from and about the user **[S2]** to provide customizable agents **[S3]** visual elements
Personification	*Identity*	**[B1]** to increase engagement **[B2]** to increase human-likeness	**[C1]** to avoid negative stereotypes **[C2]** to balance the *identity* and the technical capabilities	**[S1]** to design and elaborate on a persona
	Personality	**[B1]** to exhibit believable behavior **[B2]** to enrich interpersonal relationships	**[C1]** to adapt humor to the user's culture **[C2]** to balance the *personality* traits	**[S1]** to use appropriate language **[S2]** to have a sense of humor

Table 2.2: Social characteristics grouped by studies domain

Domain	Social Characteristics	Studies
Open domain	*Proactivity, Conscientiousness, Damage control, Thoroughness, Manners, Moral agency, Emotional intelligence, Personalization, Identity, Personality*	(Thies *et al.*, 2017) (Portela and Granell-Canut, 2017) (Shum *et al.*, 2018) (Morrissey and Kirakowski, 2013) (Curry and Rieser, 2018) (De Angeli *et al.*, 2001) (Hill *et al.*, 2015) (Kirakowski *et al.*, 2009) (Mairesse and Walker, 2009) (De Angeli and Brahnam, 2006) (Banks, 2018) (Brahnam and De Angeli, 2012) (Ho *et al.*, 2018) (De Angeli, 2005) (Corti and Gillespie, 2016) (Ptaszynski *et al.*, 2010)
Ethnography	*Proactivity, Conscientiousness, Thoroughness, Personalization*	(Tallyn *et al.*, 2018)
Task management	*Proactivity, Damage control, Manners, Personalization, Identity*	(Liao *et al.*, 2016)(Toxtli *et al.*, 2018)
Tourism	*Proactivity, Thoroughness, Manners*	(Chaves and Gerosa, 2018)
Business	*Proactivity, Personalization*	(Duijvelshoff, 2017)
Information search	*Proactivity, Damage control, Manners, Emotional intelligence*	(Avula *et al.*, 2018)(Wallis and Norling, 2005)
Decision-making coach	*Proactivity, Damage control, Manners*	(Mäurer and Weihe, 2015)
Health-care	*Proactivity, Emotional intelligence*	(Fitzpatrick *et al.*, 2017)(Miner *et al.*, 2016)
Credibility assessment interviews	*Proactivity, Conscientiousness*	(Schuetzler *et al.*, 2018)
Education	*Proactivity, Conscientiousness, Damage control, Thoroughness, Manners, Emotional intelligence, Identity, Personality*	(Ayedoun *et al.*, 2017)(Coniam, 2008)(Dyke *et al.*, 2013)(Hayashi, 2015)(Kumar *et al.*, 2010)(Silvervarg and Jönsson, 2013)(Sjödén *et al.*, 2011)(Tegos *et al.*, 2016b)
Financial services	*Conscientiousness, Communicability, Damage control, Thoroughness, Personalization, Identity*	(Candello *et al.*, 2017)(Duijst, 2017)
Customer service	*Conscientiousness, Communicability, Damage control, Thoroughness, Manners, Emotional intelligence, Personalization, Identity*	(Araujo, 2018)(Brandtzaeg and Følstad, 2018)(Gnewuch *et al.*, 2017)(Jenkins *et al.*, 2007)(Lasek and Jessa, 2013)
E-commerce	*Conscientiousness, Manners*	(Jain *et al.*, 2018a)
News reading	*Communicability*	(Valério *et al.*, 2017)
Human resources	*Communicability, Damage control, Manners, Identity*	(Liao *et al.*, 2018)
Virtual assistant	*Thoroughness, Emotional intelligence, Personalization, Identity*	(Ciechanowski *et al.*, 2018)(Zamora, 2017)
Gaming	*Thoroughness, Emotional intelligence, Personality*	(Dohsaka *et al.*, 2014)(Morris, 2002)
Race-talk	*Moral agency, Identity*	(Marino, 2014)(Schlesinger *et al.*, 2018)
Humorous talk	*Personality*	(Meany and Clark, 2010)
Not defined	*Proactivity, Conscientiousness, Communicability, Damage control, Personalization, Identity, Personality*	(Brandtzaeg and Følstad, 2017) (Jain *et al.*, 2018b) (Neururer *et al.*, 2018)

2.1.1 Conversational intelligence

Conversational intelligence includes characteristics that help the chatbot perform a proactive, attentive, and informative role in the interaction by enabling the chatbot to actively participate in the conversation and demonstrate awareness of the topic discussed, the evolving conversational context, and the dialogue flow. The highlighted benefits relate to how a chatbot manages the conversation to make it productive, interesting, and neat.

To achieve that, designers and researchers must pay close attention to the timing and relevance of provided information, privacy, interactional flexibility, and consistency. The social characteristics in this category are as follows:

Proactivity. Reported in 18 studies, *proactivity* is the capability of a chatbot to share initiative with the user, contributing to the conversation in a more natural way (Morrissey and Kirakowski, 2013). Chatbots may manifest *proactivity* when they initiate exchanges, suggests new topics, provide additional information, or formulate follow-up questions. *Proactivity* was mostly investigated in open domain and education chatbots. In open domain chatbots, *proactivity* helps to keep the conversation alive (i.e., engagement) (Shum *et al.*, 2018). Educational chatbots rely on *proactivity* to motivate human students to think, share, and collaborate (e.g., (Dyke *et al.*, 2013)). *Proactivity* was also observed in other eight task-oriented domains, including task management and information search.

Conscientiousness. Reported in 11 studies, *conscientiousness* is a chatbot's capacity to demonstrate attentiveness to the conversation at hand (Dyke *et al.*, 2013; Duijst, 2017). It enables a chatbot to follow the conversational flow, show understanding about the context, and interpret each utterance as a meaningful part of the whole conversation (Morrissey and Kirakowski, 2013). *Conscientiousness* was investigated mainly in the education domain, where the tutor chatbot is expected to be attentive and relevant as well as to control the conversation flow to keep the students focused on the topic (see, e.g., (Dyke *et al.*, 2013)). *Conscientiousness* was also investigated in other seven domains, including open domain interactions and customer and financial services.

Communicability. Reported in 6 studies, *communicability* is the capability of a chatbot to convey its features to users (Valério *et al.*, 2017). The problematic around a chatbot's *communicability* lies in the nature of the interface: instead of buttons, menus, and

links, chatbots unveil their capabilities within conversational turns, one sentence at a time (Valério *et al.*, 2017). The lack of *communicability* may lead users to give up on using the chatbot when they cannot understand the available functionalities and how to use them (Valério *et al.*, 2017). *Communicability* was investigated in the customer service domain, where chatbots guide customers through available functionalities (Valério *et al.*, 2017). The remaining task-oriented domains reporting *communicability* are financial services, news reading, and human resources. We did not identify studies on open domain chatbots that report *communicability* since in open domain interactions, users are free to talk about varying topics, and guidance is less a concern.

2.1.2 Social intelligence

Social intelligence focuses on habitual social protocols, which refers to how a chatbot reproduces adequate social behavior for the purpose of achieving desired goals (Björkqvist *et al.*, 2000). Characteristics in this category refer to chatbot's ability to respond to social cues during the conversation, accept differences, and manage conflicts (Salovey and Mayer, 1990), as well as be empathic and demonstrate caring (Björkqvist *et al.*, 2000). The benefits relate to resolving social positioning and recovering from failures, and increasing believability, authenticity, human-likeness, engagement, and rapport. To achieve that, designers and researchers must pay close attention to privacy, emotional regulation issues, language consistency, and identification of failures and inappropriate content. The social characteristics in this category are as follows:

Damage control. Reported in 12 papers, *damage control* is the ability of a chatbot to deal with either conflict or failure situations in a socially acceptable manner (Wallis and Norling, 2005; Jain *et al.*, 2018b; Silvervarg and Jönsson, 2013). *Damage control* was

mostly investigated for open domain and customer service chatbots. In open domain interactions, users are free to wander among topics, and testing or flaming tend to be more frequent (Hill *et al.*, 2015). In the customer service context, the chatbot needs to avoid disappointing the user, as frustration may negatively reflect the business that the agent represents (Araujo, 2018). *Damage control* was also identified in other six domains, such as task management and financial services.

Thoroughness. Reported in 13 papers, *thoroughness* is the ability of a chatbot to be precise regarding how it uses language to express itself (Morrissey and Kirakowski, 2013). *Thoroughness* defines that chatbots should coherently use language that portrays the expected style (Mairesse and Walker, 2009). *Thoroughness* is mainly reported for open domain (five studies) and customer service chatbots (two studies), where the interactions are expected to be natural and credible to succeed (Morrissey and Kirakowski, 2013; Gnewuch *et al.*, 2017). *Thouroughness* was also reported in other six domains of studies, such as financial services and education.

Manners. Reported in 10 papers, refer to the ability of a chatbot to manifest polite behavior and conversational habits (Morrissey and Kirakowski, 2013). A chatbot can manifest *manners* by adopting speech acts such as greetings, apologies, and closings (Jain *et al.*, 2018b); minimizing impositions (Tallyn *et al.*, 2018; Toxtli *et al.*, 2018), and making interactions more personal (Jain *et al.*, 2018b). *Manners* potentially reduces the feeling of annoyance and frustration that may lead the interaction to fail (Jain *et al.*, 2018b). We identified studies reporting *manners* in nine different domains, with only open domain showing up twice. The list includes education, information search, and task management, among others.

Moral agency. Reported in 6 papers, *moral agency* is a manifested behavior that may

be inferred by a human as morality and agency (Banks, 2018). *Moral agency* was observed in only two domains of studies: open domain (four studies) and race-talk (two studies), which shows that this characteristic is primarily relevant when the conversational topic may raise moral concerns, which ultimately requires ethical behavior from the conversational partners.

Emotional intelligence. Reported in 14 papers, *emotional intelligence* allows an individual to appraise and express feelings, regulate affective reactions, and harness the emotions to solve a problem (Salovey and Mayer, 1990). Although chatbots do not have genuine emotions (Wallis and Norling, 2005), there are considerable discussions about the role of manifesting (pretended) emotions in chatbots (Shum *et al.*, 2018; Wallis and Norling, 2005; Ho *et al.*, 2018). An emotionally intelligent chatbot can recognize and control user's feelings and demonstrate respect, empathy, and understanding, improving the relationship between them (Salovey and Mayer, 1990; Li *et al.*, 2017). *Emotional intelligence* is mainly investigated in domains where topics may involve the disclosure of feelings (e.g., in open domain interactions (Shum *et al.*, 2018)) and expressions of empathy and understanding are appropriate (Fitzpatrick *et al.*, 2017; Dohsaka *et al.*, 2014) (e.g., health care, gaming, education).

Personalization. Reported in 11 papers, *personalization* allows a chatbot to be aware of situational context and to adapt its features dynamically to better suit individual needs (Neururer *et al.*, 2018). *Personalization* was investigated in seven different domains, including open domain and task management (two studies each). In open domain interactions, personalization is derived from remembering information from previous interactions, such as personal preferences and other user details (Thies *et al.*, 2017). In task-oriented contexts, such as task management, personalization aims to increase the rele-

vance of services to particular users (Liao *et al.*, 2016).

2.1.3 Personification

Finally, **personification** refers to the chatbot's perceived identity and personality representations, which include characteristics that help a chatbot to manifest personal and behavioral traits. The benefits relate to increasing believability, human-likeness, engagement, and interpersonal relationship, which is in line with the benefits of **social intelligence**. However, unlike the **social intelligence** category, designers and researchers should focus on attributing recognizable *identity* and *personality* traits that are consistent with a user's expectations and the chatbot's capabilities. In addition, it is important to care about adaptation to a user's culture and to reduce the effects of negative stereotypes. The social characteristics in this category are as follows:

Identity. Reported in 16 papers, *identity* refers to the ability of an individual to demonstrate belonging to a particular social group (Stets and Burke, 2000). Although chatbots do not have the agency to decide what social group they want to belong to, designers attribute *identity* to them, intentionally or not, when they define the way a chatbot talks or behaves (Cassell, 2009). Aspects that convey a chatbot's *identity* include gender, age, language style, and name. Additionally, chatbots may have anthropomorphic, zoomorphic, or robotic representations. *Identity* concerns were primarily investigated for open domain and customer service chatbots (4 studies each). In open domain interactions, *identity* is explored as a means of building common ground (De Angeli *et al.*, 2001). In the case of customer service, *identity* helps to manifest credibility and trust (Gnewuch *et al.*, 2017). Other domains include race-talk, education, and gaming.

Personality. Reported in 12 papers, *personality* refers to the set of traits that deter-

mines the agent's interaction style, describes its character, and allows the end-user to understand its general behavior (De Angeli *et al.*, 2001). Therefore, *personality* ensures that a chatbot displays behaviors that stand in agreement with user expectations in a particular context (Petta and Trappl, 1997). Chatbots with consistent *personality* are more predictable and trustable (Shum *et al.*, 2018). *Personality* was mostly investigated in open domain and education chatbots. The other two domains were gaming and humorous chatbots, which are both playful agents. Hence, *personality* is relevant when believability and attitude play a role in the interaction (Portela and Granell-Canut, 2017) (e.g., in open domain) and when a chatbot's attitude may increase the user's mental comfort when performing a task (Ayedoun *et al.*, 2017), as in educational contexts.

2.2 Chatbot Thoroughness

The survey of the literature revealed two benefits of thoroughness. Firstly, chatbots potentially increase human-likeness by adapting their language to the interactional context. When analyzing interactions with a customer representative chatbot, Jenkins *et al.* (2007) observed that the chatbot proposed synonyms to keywords, and the repetition of this vocabulary led the users to imitate it. Hill *et al.* (2015) observed a similar tendency to matching language style. The authors found that people use fewer words per message and limited vocabulary with chatbots than with a human partner. When interacting with a chatbot that uses emojis and letter reduplication (Thies *et al.*, 2017), participants reported a draining experience, since the chatbot's energy was too high to match. Since these outcomes show that adapting the language to the interlocutor is a common behavior for humans, chatbots would benefit from manifesting it. Chatbots should also adapt their language to the context in which they are implemented and adopt appropriate linguistic

registers (Morrissey and Kirakowski, 2013; Gnewuch *et al.*, 2017; Duijst, 2017).

Secondly, thoroughness increases a chatbot's believability. Morrissey and Kirakowski (2013) found that chatbot's formal grammatical and syntactical chiefly determine whether they are effective or ineffective, and that chatbots should use consistent grammar and spelling. Morris (2002) states that believable chatbots also need to display unique characters through their linguistic choices; this point is affirmed by Mairesse and Walker (2009), who demonstrated that personality could be expressed by language patterns and proposed a computational framework for producing utterances that manifest a target personality. The outcomes showed that a single utterance using appropriate linguistic form could manifest a believable personality. Participants in Jenkins *et al.* (2007) described some interactions as "robotic" when the chatbot repeated keywords in the answers. Similarly, in Tallyn *et al.* (2018), participants complained about the "inflexibility" of pre-defined responses and expressed the desire for the chatbot to talk more "as a person."

Despite the relevance of language for chatbot design and the benefits of thoroughness, the survey revealed a lack of studies focusing on chatbot linguistic choices or how they impact the user's experiences. Out of 13 papers that report this characteristic, only three focus the investigation on how patterns of language influence user perceptions and behavior toward chatbots (Duijst, 2017; Mairesse and Walker, 2009; Hill *et al.*, 2015). Gnewuch *et al.* (2017) and Morris (2002) suggest design principles that consider language choices. In the remaining studies, issues related to thoroughness emerged as exploratory findings.

Additionally, the surveyed literature does not discuss strategies to provide thoroughness. Tamayo-Moreno and Pérez-Marín (2017) state that designers "seems to create the agent according to their expertise and needs, without a unified procedure". Although the literature in computational linguistics has proposed algorithms and statistical mod-

els to manipulate language style and matching (see, e.g., (Prabhumoye *et al.*, 2018; Zhao *et al.*, 2018b; Zhang *et al.*, 2017)) as well as the relevance and informativeness of the utterances (Bi *et al.*, 2019), these strategies have not been evaluated from the perceived user experience perspective.

The surveyed literature highlights that the main challenge in effectively implementing thoroughness is to be consistent regarding language choices. For example, when analyzing an open-domain interaction, Kirakowski *et al.* (2009) found that participants consider it inappropriate when chatbots used more formal language or unusual vocabulary since general-purpose chatbots generally engage in casual interactions. These outcomes underscore that social perceptions affect communication with chatbots, and align with human-human, CMC (Walther, 2007; Baron, 1984) as well as sociolinguistic theories (Biber and Conrad, 2019).

The outcomes related to *Thoroughness* underscored that social perceptions of language affect communication with chatbots, which supports the claim that chatbot design would potentially benefit from developing language that complies with a particular social role.

2.3 Chatbot Language Design

Interaction with technology using natural language is "becoming increasingly feasible and potentially very significant" (Dale, 2016). Major technology companies such as Microsoft, IBM, and Facebook have invested in providing platforms to integrate third-party chatbots into their instant messaging tools. According to Følstad and Brandtzæg (2017), along with social networks, mobile messaging applications are the main user interface to the Internet, and from 2016-2017 thousands of chatbots have been developed for popular instant messaging tools (Brandtzaeg and Følstad, 2017). Despite the recent popularity,

there is evidence that users are unsatisfied with their experiences with chatbots (Kiseleva *et al.*, 2016; Luger and Sellen, 2016). Dale (2016) states that interacting with currently available chatbots conveys the impression of *"being managed through a tightly controlled dialog flow"* with reduced interactivity, which turns users into an option-selector rather than a conversational partner.

Making chatbot's conversations appear more natural to users is challenging. Although previous studies have shown that people do not talk to chatbots in the same way they talk to other humans (Hill *et al.*, 2015; Mou and Xu, 2017; Raij *et al.*, 2007; Shechtman and Horowitz, 2003), the literature has also shown that people expect chatbots to reflect social and conversational protocols (Brahnam and De Angeli, 2012; Hayashi, 2016; Lee and Choi, 2017; Marino, 2006; Silvervarg *et al.*, 2012). For example, Lee and Choi (2017) highlight that interactions with self-disclosure and reciprocity are more satisfactory than interactions without these characteristics. Hayashi (2016) found that people in a group interacting with chatbots were influenced by the emotional state that the chatbots constructed while performing a problem-solving task. Some studies showed that topic selection and verbal abuse by the user are gender-related characteristics (Brahnam and De Angeli, 2012; Silvervarg *et al.*, 2012). In contrast, Hill *et al.* (2015) claim that people used more–but shorter–messages, and a more restricted vocabulary when communicating with chatbots.

An explanation for these contrasts is that message interactivity is dependent on the identity of the interlocutor (Sundar *et al.*, 2016), i.e., in human-chatbot interactions, people direct their messages to an artificial agent. Sundar *et al.* (2016) discuss that people usually send challenging questions to a chatbot to test their responsiveness, while they would not do the same with a human interlocutor. Chatbots are typically designed to mimic the social roles usually associated with a human conversational partner, for ex-

ample, a buddy (Thies *et al.*, 2017), a tutor (Tegos *et al.*, 2016b; Dyke *et al.*, 2013), healthcare provider (Montenegro *et al.*, 2019; Fitzpatrick *et al.*, 2017), a salesperson (Gnewuch *et al.*, 2017; Zhu *et al.*, 2018), a hotel concierge (Lasek and Jessa, 2013), or, as in this research, a tourist assistant (Chaves and Gerosa, 2018; Chaves *et al.*, 2019b). Research on mind perception theory (Lee *et al.*, 2019; Heyselaar and Bosse, 2019; Keijsers and Bartneck, 2018) suggests that although artificial agents are presumed to have sub-standard intelligence, people still apply certain social stereotypes to them. It is reasonable, then, to assume that "machines may be treated differently when attributed with higher-order minds" (Lee *et al.*, 2019). As chatbots enrich their communication and social skills, the user expectations will likely grow as the conversational competence and perceived social role of chatbots approach the human profiles they aim to represent. A variety of factors influence how people perceive chatbot communication skills (Chaves and Gerosa, 2020; Feine *et al.*, 2019; Tariverdiyeva, 2019) and, as user expectations of proficiency increase, one important way to enhance chatbot interactions is by carefully planning their use of language (Kirakowski *et al.*, 2009; Go and Sundar, 2019).

Most linguists agree that the language choices made by humans are systematic (Kilgarriff, 2005), and previous research has provided ample evidence that variation within a language can often be accounted for by factors such as individual author/speaker style (e.g., (Argamon *et al.*, 1998; Leech and Short, 2007)), dialect (e.g. (Labov *et al.*, 2005; Szmrecsanyi, 2011)), genre (e.g., (Kamberelis, 1995; Paltridge, 1994)), and register (e.g., (Biber, 1988; Biber and Conrad, 2019)). Among these factors, style has particularly captured the attention of researchers on conversational agents (Feine *et al.*, 2019; Thomas *et al.*, 2018; Niederhoffer and Pennebaker, 2002; Jakic *et al.*, 2017; Lin and Walker, 2017), with explorations ranging from consistently mimicking the style of a particular character (Lin and

Walker, 2017; Syed, 2020) to dynamically matching the style to the conversational partner (Niederhoffer and Pennebaker, 2002; Hoegen *et al.*, 2019).

A number of studies have sought to empirically evaluate the influence of conversational style on user experiences with chatbots (Elsholz *et al.*, 2019; Araujo, 2018). For example, Elsholz *et al.* (2019) compared interactions with chatbots that use modern English to those that use a Shakespearean language style. Users perceived the chatbot that used the modern English style as easy to use, while the chatbot that used Shakespearean English was seen as more fun to use. Araujo (2018) evaluated the influence of anthropomorphic design cues on user perceptions of companies represented by a chatbot, where perceptions include attitudes, satisfaction, and emotional connection with the company; one cue, for instance, was the use of an informal language style. Results showed that anthropomorphic cues resulted in significantly higher scores for adjectives like likable, sociable, friendly, and personal in user evaluations of the interactions (though the relative impact of individual anthropomorphic cues on the outcomes was not evaluated). Similarly, based on exploratory analysis, Tariverdiyeva (2019) concluded that "appropriate degrees of formality" (renamed "appropriate language style" in a subsequent work (Balaji, 2019)) directly correlates with user satisfaction. We note that these studies define "appropriate language" as the "ability of the chatbot to use appropriate language style *for the context*." This linkage of perceived appropriateness of language to context is important and reflects clear evidence that appropriateness of language is not absolute, but rather influenced by the user's specific expectations concerning the chatbot's communicative behavior and the stereotypes of the social category (Jakic *et al.*, 2017; Krauss and Chiu, 1998). For example, when assessing the effects of language style on brand trust in online interactions with customers, Jakic *et al.* (2017) concluded that the perceived language fit

between the brand and the product/service category increases the quality of interaction. Proficiency in human-like language style may also influence the user perceptions of chatbot credibility. Jenkins *et al.* (2007) observed that chatbots are deemed sub-standard when users see them "*acting as a machine*"; similarly, in analyzing the naturalness of chatbots, Morrissey and Kirakowski (2013) found that correct language usage was a determinant in perceived chatbot quality. The failure to convey linguistic expertise compromises credibility (Zumstein and Hundertmark, 2017), i.e., the chatbot's ability to convey believability and competence (Sweeney and Swait, 2008; Mack *et al.*, 2008).

Although some scholars define style as "the meaningful deployment of language variation in a message" (Feine *et al.*, 2019), sociolinguistics define style as a set of linguistic variants that reflect aesthetic preferences, usually associated with particular speakers or historical periods (Biber and Conrad, 2019) (e.g., Shakespearean vs. modern English). Sociolinguistic studies also emphasize that the *"core linguistic features like pronouns and verbs are functional"* rather than aesthetic (Biber and Conrad, 2019), which points to register. Register theory states that for each interactional situation, there is a subset of norms and expectations for using language to accomplish communicative functions (Biber and Conrad, 2019). In a conversation, every utterance is influenced by the social atmosphere (Bakhtin, 2010; Jabri *et al.*, 2008), which is represented in the form of *situational parameters*, such as the relationship between participants, the purpose of the interaction, and the topic of the conversation (Kamberelis, 1995; Biber and Conrad, 2019). This results in the emergence of situationally-defined language varieties, which ultimately determine the interlocutor's linguistic choices (Biber, 2012; Biber and Conrad, 2019).

Although the relevance of register in human-human communication has been extensively demonstrated (Biber, 2012), the extent to which this theory applies to human-

chatbot interactions has yet to be widely investigated. There is some evidence suggesting that chatbots should use language appropriate to the service category that the chatbot represents (Balaji, 2019). Still, there has been no systematic analysis of how user perceptions might be influenced by expectations regarding chatbot language, or exploration of specific core linguistic features that determine the appropriateness of language fit. In the next section, we focus on how the register theory applies to human-human communication and why it should be considered in chatbot interactions.

2.4 Register and Linguistic Variation

Register theory posits that every utterance in a conversation is influenced by the social atmosphere (Bakhtin, 2010; Jabri *et al.*, 2008), which is represented in the form of situational parameters, such as the relationship between participants, the purpose of the interaction, and the topic of the conversation (Kamberelis, 1995; Biber and Conrad, 2019). The influence of these parameters results in the emergence of situationally-defined language varieties (Biber, 2012; Biber and Conrad, 2019). Hence, the register can be interpreted as the distribution of the *linguistic features* in a conversation, given the *context*; the linguistic features consist of the set of words or grammatical characteristics that occur in the conversation, and the context consists of a set of situational parameters that characterize the situation in which the conversation occurs, e.g., the participants, the channel, the production circumstances, and so on. Figure 2.1 illustrates the relationship among linguistic variation, register, and function (Egbert and Biber, 2016).

Several recent studies have shown that register is crucial for linguistic research on language variation and use: most linguistic features are adopted in different ways and to varying extents across different registers. For example, some studies have focused on

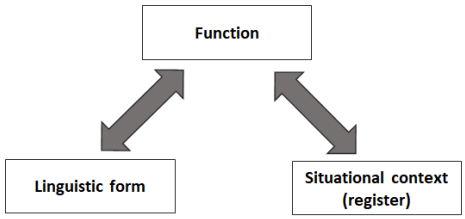


Figure 2.1: Characteristics of register

describing register variation in the use of a narrower set of features, such as grammatical complexity features (Biber *et al.*, 2011; Biber and Gray, 2010), lexical bundles (Biber *et al.*, 2004; Hyland, 2012), and evaluation (Hunston *et al.*, 2010). The *Longman Grammar of Spoken and Written English* (Biber *et al.*, 1999) documents the systematic patterns of register variation for most grammatical features in English, exposing the power of register as a predictor of linguistic variation. Indeed, we draw on this grammar to support the discussion surrounding the use of particular linguistic features in the context of tourist assistant discourse (see Chapter 6).

The value of register for understanding conversational structure is further emphasized by studies showing that *failing* to account for register in linguistic analyses and computational language models can, and often does, result in incorrect conclusions about language use. Biber (2012), for instance, offers many examples of how failing to account for conversational register in a linguistic analysis can result in faulty conclusions.

Given the crucial role of register in shaping human-human communication, we suggest that register must be accounted for in the design of a chatbot's language; the user's perceptions of a chatbot as competent and trustworthy conversational partners depends on the chatbot's correct use of register. This research works to provide a practical cornerstone for this broad endeavor by exposing associations between the frequency of core linguistic features (which comprise the conversational register) and user evaluations of the

quality of the chatbot interactions. We focus the study on the domain of "tourist information search" to (a) analyze the linguistic features relevant to characterizing conversational register in this domain, and (b) show how adherence or failure to adhere to that register impacts user perceptions of conversational quality. To identify the typical patterns of language present in the interactions, we applied register analysis, as developed by Biber and Conrad (2019), consisting of three main steps. First, *situational analysis* aims to characterize the interactions in the target corpus using a conversational taxonomy based around seven situational parameters: participants, relationship, channel, production, setting, purpose, and topic. Second, *register characterization* analyzes and aggregates the results of the situational analysis to yield a statistical characterization of the linguistic features typically used in domain interactions (in this case, within the tourist assistant domain); the result is a register model for the domain, i.e., a concrete representation of the appropriate register for the given domain. Finally, we interpret and discuss the functional link between situation and the user's preferences of a chatbot's language use.

In the following section, we briefly discuss the choice of the tourist information search domain as the testbed for exploring the role of conversational register in chatbot design.

2.5 Chatbots in the tourism domain

Tourism is one of the fastest-growing economic sectors in the world and is a major category of international trade in services (UNWTO, 2017). As the sector grows, the demand for timely and accurate information on destinations also increases. Because people want to reduce uncertainty in travel, they engage in an information search process that enriches the quality and reduces the perceived risks of their trip (Fodness and Murray, 1997). A recent survey (Loo, 2017) revealed that the Internet is the top source for travel

planning. As the penetration of smartphone devices with data access has increased, travelers increasingly search for information and make decisions en-route (Buhalis and Jun, 2011; Think with Google, 2016; Wang *et al.*, 2016). However, conducting en-route travel information searches can be an overwhelming experience (Lang, 2000; Pawłowska, 2016), due to the information overload compounded by a lack of reliable mechanisms for finding accurate, trustworthy, and relevant information (Lang, 2000).

Radlinski and Craswell (2017) suggest that complex information searches could benefit from conversational interfaces and, indeed, several conversational agents with a wide range of characteristics have been developed to improve tourism information search and travel planning (Alexis, 2017; Ivanov and Webster, 2017; Lee *et al.*, 2010; Linden *et al.*, 1997). Loo (2017) claims that over one in three travelers across countries are interested in using digital assistants to research or book travel and that travel-related searches for "tonight" and "today" have grown over 150% on mobile devices in just two years. In particular, a report on the chatbot market (Grand View Research, 2017) places the Travel & Tourism sector as one of the top five markets with the best revenue prospects by 2025. In response to this trend, the number of chatbots within the online tourism sector has increased. The BotList website, for example, lists nearly a hundred available chatbots under the travel category; some examples include the Expedia [1] and Marina Alterra [2] virtual assistants.

Travel planning is also a domain in which perceived competence and trustworthiness is central to user experience; advice that is not appropriately presented is unlikely to be trusted and utilized. From a more pragmatic perspective, the popularity of tourist assistants (both human and chatbot) means that a growing corpus of conversations in

[1] https://botlist.co/bots/expedia

[2] https://botlist.co/bots/marina-alterra

this domain exists and can serve as the seed for the analysis.

In sum, we selected the tourism advising domain for developing a practical framework for including register in the design of chatbot conversational engines because proper use of conversational register is likely to be particularly critical for user experience in this domain, there is a real-world demand for chatbots for travel advice, and several corpora of human and chatbot interactions in this domain are available.

Chapter 3

DATA COLLECTION

To investigate how language varies across registers within a particular domain, we need to understand how humans in that situation would build their utterances. Because we are interested in helping tourists to gather en-route information about a destination through an instant message tool, we reproduced this interactional situation by collecting a baseline corpus of human-human conversations between tourist assistants and tourists which is introduced ins Section 3.1. Then, since register is characterized by comparing varying contexts, we selected a second corpus of conversations in the tourism domain that is available online and is commonly used in natural language research. This corpus is introduced in Section 3.2.

To identify the situational parameters in which the conversations in each corpus occurred, we performed a situational analysis to place the register within a broad taxonomy of situational features that define the interactional context. The situational analysis supports the qualitative interpretation of the communicative functions of a linguistic feature within a register, and it is discussed in Section 3.3.

3.1 The FLG corpus of tourist conversations

To collect the FLG corpus, we hired three experienced professionals from the Flagstaff Visitor Center in Flagstaff, Arizona, USA, to answer tourist questions about the city and nearby tourist destinations during summer 2018. The official government website reports that Flagstaff receives over 5 million visitors per year (Flagstaff, 2019), including in-state,

out-of-state, and international visitors. According to the 2017-2018 Flagstaff Visitor Survey (Thomas Combrink, 2018), Flagstaff is the central hub for visiting tourist destinations such as Grand Canyon National Park, Arizona Snow Bowl, the Navajo and Hopi reservations, and many other local attractions. Regional tourism is significant as well, with a large number of visitors seeking to escape the heat and crowding of the Phoenix metropolitan area.

The three tourist assistants were native English speakers, female, had some post-secondary education, and had four or more years of experience as tourist assistants. Two of them were 25-34 years old; the other was in the 35-44 age range. Although they had more than four years of experience in providing tourist information in in-person conversations at the Visitor Center, they had never professionally provided information through an online platform.

To recruit tourists to interact with the tourist assistants, we advertised the free tourist assistant online service in the city of Flagstaff through flyers and intercepted tourists at the Flagstaff Visitor Center in Historic Downtown, directing them to a booth to use the service. About 30 tourists participated in the interactions. We also collected tourism-related questions about Flagstaff from websites such as Quora, Google Maps, and TripAdvisor, and a researcher posed these questions to the tourist assistants. The tourist assistants were unaware of the origin of the questions and thought they were always interacting with real tourists.

The tourist advising conversations were performed through a Facebook Messenger account (Facebook, 2018) over the summer of 2018. The human tourist assistants participated in the study from the research lab. Before the first interaction, the tourist assistants participated in a training session, in which we presented the environment and the tools.

During the study, a researcher observed the interactions and took notes on comments made by the tourist assistants. Because we wanted to understand the natural linguistic variation in tourism-related interactions, both tourists and tourist assistants were free to interact according to their needs, interests, and knowledge. No tasks were proposed to the tourists, nor were any scripts provided to the tourist assistants. The textual exchanges were exported from Facebook Messenger and archived to create the FLG corpus; the corpus comprises 144 interactions with about 540 question-answer pairs. To analyze the register of the conversation, we only used the answers from the tourist assistants.

3.2 The DailyDialog Corpus

The second corpus we selected is $DailyDialog$ (Li *et al.*, 2017), which is a corpus available online and used as a reference for research on natural language generation in the tourism domain. DailyDialog consists of conversations about daily life automatically extracted by web crawling utilities from websites for English language learning, with topics ranging from ordinary life to politics, health, and tourism. Similar to someone who would use this corpus to train a chatbot, we filtered the original corpus to select only conversations that were originally labeled as "tourism" and that focused on customer service interactions, e.g., hotel guest-concierge, business person-receptionist, tourist-tour guide, etc. We chose $DailyDialog$ because it contains a large set of conversations in the tourism domain, and it is likely to be used as a baseline model for chatbot conversations (Galitsky *et al.*, 2019).

We downloaded the $DailyDialog$ corpus from its website [1] . After filtering to focus on tourism-related interactions, the subset of $DailyDialog$ used in this research comprises

[1] http://yanran.li/dailydialog

999 interactions. Because we are only interested in the utterances produced by the service providers ($DailyDialog$) and tourist assistants (FLG), we edited the conversations to remove the tourists' utterances.

3.3 Corpora characteristics

We followed the situational analytical framework proposed by Biber and Conrad (2019) to identify the situational parameters in which the conversations took place. The primary outcome of the situational analysis is presented in Chaves *et al.* (2019b) and summarized in Table 3.1.

Table 3.1: Situational analysis. Situational parameters are extracted from the situational analytical framework (Biber and Conrad, 2019)

Situational parameter	**DailyDialog**	**FLG**
Participants	Customer and service providers	Tourists and tourist assistants
Relationship	Role, power, and knowledge relations vary	Tourist assistant and tourist, the former owns the knowledge
Channel	Human-written, representing face-to-face	Written, instant messaging tool
Production	Planned	Quasi-real-time
Setting	Private, shared time, and mostly physically shared place	Private, shared time, virtually shared place
Purpose	Provide a service or information	Information search
Topic	Varies within the context of tourism	Local information (e.g., activities, attractions)

According to register theory (Biber and Conrad, 2019), differences in situational parameters result in a varying register; people use different patterns of language depending on the context. $DailyDialog$ presents larger variability in terms of situational parameters (e.g., participants, purpose, and topic) than FLG. Given differences in the situational parameters, we expect that the language characteristics in FLG differ from the language characteristics in $DailyDialog$. We investigate this claim in the next chapter, where we discuss the register characterization analysis, which identifies the linguistic features that

determine the register of each corpus and how the typical language varies among different situations in the tourism domain.

Chapter 4

REGISTER CHARACTERIZATION

To characterize the varying conversational registers used in the two corpora conversations, we performed a register analysis (Biber, 1988). Register analysis consists of identifying the linguistic features typically used in a corpus, which is based on tagging and counting the linguistic features present in the utterances and interpreting them according to their function in the sentence (Biber and Conrad, 2019; Biber, 1988). We performed register analysis for both FLG and $DailyDialog$ corpora and then compared the outcomes to identify the variations in language use across corpora. The following sections present this analysis and its outcomes in detail.

4.1 Procedures

The register analysis relied on information from the Biber grammatical tagger Biber (2017) to identify the linguistic variation present in each corpus. Given a set of texts, this tool tags and counts linguistic features present in each text, and returns the counts normalized per 1,000 words. The tagger also calculates *dimension scores* for each text, which are based on aggregation of features derived using factor analysis Biber (2017). The dimension scores reveal the prevailing characteristics of the register (i.e., the levels of personal involvement, narrative flow, contextual references, persuasion, and formality present in the text) (Biber, 1988). For example, a positive score for Dimension 1 indicates that the discourse is involved and interactive while a negative score for that dimension indicates that the discourse is more informational dense. Appendix C contains a glossary

of the dimensions and the 49 linguistic features that compose these dimensions, including examples. Details about the tagger can be found elsewhere (Biber, 2017, 1988).

We first analyzed the dimension scores to understand the linguistic characteristics and variation in the discourse in each corpus. Following Biber (1988), we applied a one-way multivariate analysis method (MANOVA) to generate a statistical comparison of the dimension scores across corpora, where the dependent variables are the values of the five dimension scores, and the independent variables are the $DailyDialog$ (control group) and the three tourist assistants from the FLG corpus are $TA1$, $TA2$, and $TA3$ (experimental groups). Each *text* corresponds to one observation in the model, where a text is a set of one or more contiguous sentences produced by an interlocutor (i.e., one answer). Given the significant overall MANOVA test, we also performed a one-way univariate analysis ($df = 3, 1139$) for each of the five dimensions to identify the individual dimensions that influence the prevailing register characteristics.

Finally, for each of the 49 linguistic features used to calculate the dimension scores, we performed an ANOVA statistical test ($df = 3, 1139$) where the dependent variables are the frequency of occurrence of a feature normalized per 1,000 words, and the independent variables are the two corpora: the control group is $DailyDialog$, and the experimental groups are each of the three tourist assistants from the FLG corpus. All the reported statistics use a 5% significance level ($\alpha = 0.05$).

4.2 Results

The MANOVA revealed that the three tourist assistants' dimension scores are significantly different from the average $DailyDialog$ discourse ($Wilks = 0.92, F = 6.23, p < 0.0001$). Table 4.1 summarizes the univariate analysis per dimension.

Table 4.1: Univariate analysis of dimension scores ($df = 3, 1139$). For each dimension, the table shows the estimated dimension score $\pm$, the standard error per group ($DailyDialog$, $TA1$, $TA2$, $TA3$), and the corresponding *F*- and *p*-values.

	DailyDialog	**TA1**	**TA2**	**TA3**	***F***	***P*-value**
Dim. 1: Involvement	30.50 $\pm$ 1.10	14.73 $\pm$ 5.06	5.69 $\pm$ 5.12	-5.02 $\pm$ 4.86	25.58	$<$0.0001
Dim. 2: Narrative flow	-4.10 $\pm$ 0.09	-4.45 $\pm$ 0.41	-4.31 $\pm$ 0.41	-4.79 $\pm$ 0.39	1.30	0.2978
Dim. 3: Contextual ref.	-6.28 $\pm$ 0.33	-3.33 $\pm$ 1.52	-1.75 $\pm$ 1.54	-2.65 $\pm$ 1.46	5.44	0.0010
Dim. 4: Persuasion	2.43 $\pm$ 0.30	1.98 $\pm$ 1.40	-0.02 $\pm$ 1.41	1.93 $\pm$ 1.34	1.01	0.3877
Dim. 5: Formality	0.36 $\pm$ 0.26	-0.81 $\pm$ 1.20	-1.70 $\pm$ 1.21	-2.10 $\pm$ 1.15	2.41	0.0658

The dimensional analysis revealed that $DailyDialog$ is characterized by an oral discourse while tourist assistants in FLG are more literate and informational than involved (Dimension 1), which can be explained by both the face-to-face nature of the conversations in $DailyDialog$ and the variation in the role and power of participants. $DailyDialog$ also has a more extreme negative score for contextual references (Dimension 3), which might be explained by the shared space and common ground provided by face-to-face interactions. $DailyDialog$ also has a slightly more formal discourse and elaborated language, although this difference is not significant. Both corpora show descriptive rather than narrative language (negative estimates for Dimension 2) and slightly persuasive language (positive estimates for Dimension 4).

Since the ultimate goal is to reproduce the patterns of language (i.e. register) of $DailyDialog$ within the conversations of the FLG corpus to produce the parallel corpora, we need to identify not only the main linguistic characteristics present in the discourse (e.g., how involved or persuasive is the discourse), but also how these characteristics emerge in each corpus. Although the dimensional analysis reveals the overall register characterization, it is not sensitive enough to identify the prevailing linguistic features that influence the overall discourse. For example, although Dimension 4 is not significantly different, the individual linguistic features that contribute to the dimension score varied across corpora.

Thus, we statistically compared the occurrences of every linguistic feature per dimension. Figure 4.1 depicts the linguistic features that vary significantly between the two initial corpora (FLG, in red, and $DailyDialog$, in blue), as revealed by the ANOVA analysis per feature. The estimates are represented by the dots and the lines represent the standard error. A table that lists these number, as well as the *F*- and *P*-values for every analyzed linguistic feature (including the non-significant features) is presented in Appendix D.

In this chapter, we showed that the core linguistic features (collectively representing conversational register) in FLG vary significantly from $DailyDialog$. We discussed the prevailing characteristics of tourist assistants' discourse for each corpus, and what linguistic features determine their typical register. As indicated in Figure 4.1, the register analysis reveals 24 linguistic features that vary significantly across corpora through all of the five register dimensions. As we anticipated, differences in the situational parameters influenced the patterns of language observed in the corpora. These results align with the literature in linguistics (Biber, 1988; Biber and Conrad, 2019), supporting the claim that linguistic features are used to accomplish a function and that the differences in situational parameters result in conversations with varying linguistic form, even for conversations within the same domain, in this case, the tourism.

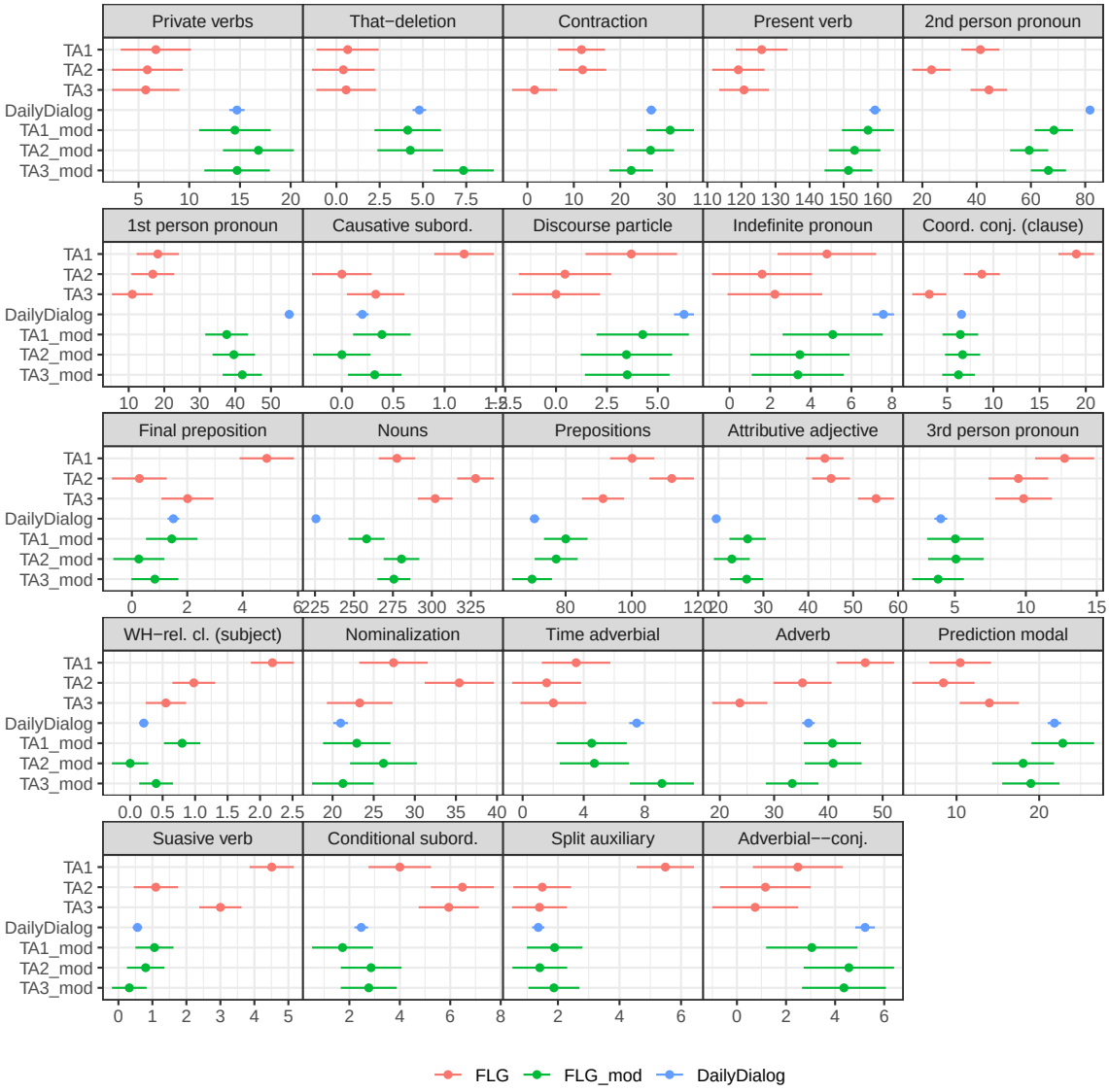


Figure 4.1: Visualization of ANOVA results for individual features comparison between $DailyDialog$ and both original (FLG) and modified (FLG_{mod}) corpora. For each significantly different feature, the estimates and standard errors are represented as a dot and a horizontal line, respectively. The colors represent each corpora (red for FLG, blue for $DailyDialog$ and green for FLG_{mod}. F- and P-values, as well as the raw numbers for all the linguistic features analyzed are presented in Appendix D

Chapter 5

TEXT MODIFICATION

Having characterized the differences in register between the FLG and $DailyDialog$ corpora, the next step was to clone the FLG corpus and then use the register characterization to artificially alter its conversational register. Specifically, we modified the individual utterances found in the FLG corpus to mimic the register characteristics observed in the $DailyDialog$ corpus. In psycholinguistics, linguistic modification consists of changing the language of a text while preserving the text's content and integrity, which includes using familiar or frequently used words (Sato, 2007; Bosher and Bowles, 2008; Long and Ross, 1993). We applied this technique to alter the linguistic features in FLG to approximate its language to the patterns presented in $DailyDialog$. We then performed a validation study to verify whether the modifications preserved the essential content and the naturalness of the original text. The new corpus, which we call FLG_{mod}, is paired with the original FLG corpus to form a pair of parallel corpora equivalent in topic, participants, and informational content, but expressed in varying patterns of language. Details of this modification process are presented in the following subsections.

5.1 Modification process

We manipulated the answers in FLG to approximate the estimate values (depicted in Figure 4.1) of a particular feature to the corresponding estimate in $DailyDialog$. For example, the estimate for *private verbs* in $DailyDialog$ is 14.70 occurrences per 1,000 words, whereas in FLG the greatest rates for *private verbs* is 7.29; therefore, we want to

increase the occurrences of *private verbs*, until we reach an estimate that is closer, and not significantly different from 14.70 for every tourist assistant in FLG corpus.

5.1.1 Procedures

Inspired by previous studies on chatbot language use (e.g., (Elsholz *et al.*, 2019)), modifications were performed semi-manually, using the AntConc tool (Anthony, 2005) and a Python script as support tools.

The first step of the text manipulation was to clone the FLG corpus to create the initial FLG_{mod}. All the modifications were performed over this cloned version and FLG was kept intact. We loaded the tagged files into the AntConc tool to facilitate inspection. Figure 5.1 shows an example of a feature's search using AntConc. Then, for each of the 24 linguistic features that shows statistical differences between FLG and $DailyDailog$ (see Figure 4.1), we inspected the $DailyDialog$ corpus to understand how the feature was used in order to reproduce the patterns of language more accurately.

After learning how one particular feature is used in $DailyDialog$, we reproduce that use in FLG using a Python script. The Python script takes a list of paired inputs, where the first parameter is a text present in the original data that needs to be changed (target); and the second parameter is the text that will replace the original (goal). The script searches for all occurrences of the target in the FLG corpus and replaces them with the goal in FLG_{mod}. This process is repeated until there are no more target-goal entries in the list. For example, Table 5.1 shows a subset of target-goal entries in the Python script. Consider the following sentence from FLG corpus:

> "Well {there are always} {*a lot of*} {*great options*} {*in downtown Flagstaff*} {*for live music*}" [$TA1$, brackets included for the example only]

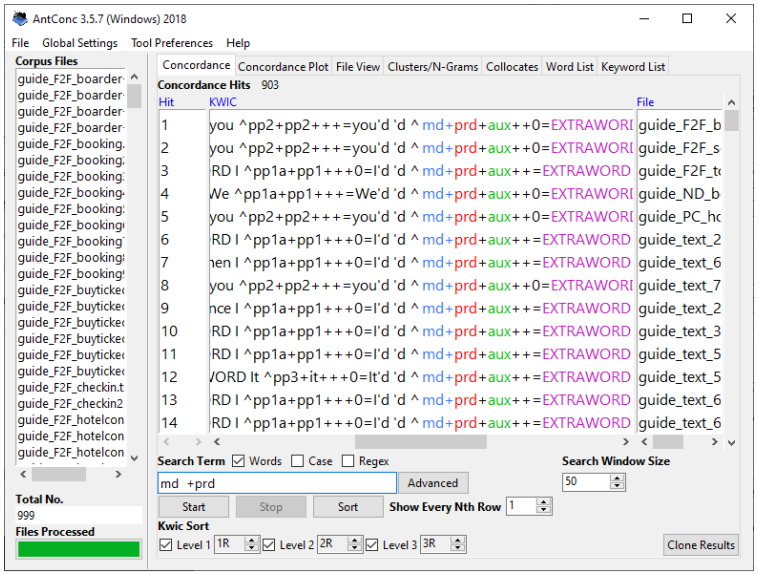


Figure 5.1: Example of feature inspection using AntConc (Anthony, 2005) tool. The example shows a search for "md +prd" (prediction modals) and the results are shown in the "Concordance" tab. The search results allow us to see what words are tagged with the searched feature (for example, the results show the occurrences of "*'d*", which represents the contracted form of *would*)

After performing the modifications listed in Table 5.1, the quote in FLG_{mod} reads as follows:

> "Well {*we always have*} {*places you can visit that have live music*}" [$TA1_{mod}$, brackets included for the example only]

After running the Python script, the produced FLG_{mod} is submitted to the Biber's tagger for a new register analysis, as described in Section 4. If the features of interest in FLG_{mod} were still statistically different from the means in $DailyDialog$, the process was repeated until the cumulative changes gathered in the modified corpus yielded a register analysis result that was as similar as possible to the register profile of $DailyDialog$. We performed a total of six rounds of modifications until we find statistically non-significant differences.

Table 5.1: Example of paired inputs for the Python script. The column TARGET lists excerpts from FLG corpus that need to be changed. The column GOAL lists the new texts that replaces the target to produce language that conforms with the patterns of language in $DailyDialog$. For example, every occurrence of the words "great option" in FLG is replaced with "option" in FLG_{mod}

TARGET	GOAL
⋮	⋮
"great options"	"options"
"going on at"	"happening"
"a lot of"	""
"in downtown flagstaff"	""
"there are always"	"we always have"
"options for live music"	"places you can visit that have live music"
"at bars like"	"i'd suggest bars like"
⋮	⋮

5.1.2 Results

Figure 4.1 shows the comparison between $DailyDialog$ and FLG_{mod}. The figure depicts the estimates for each tourist assistant in FLG_{mod} (in green) compared to the $DailyDialog$ (in blue). For three features, namely *nouns*, *first-person pronouns*, and *second-person pronouns*, linguistic modification did not reach a non-significant difference, although we substantially reduced the *F*-values, which is clearly depicted in Figure 4.1. The problem with these features is that differences between $DailyDialog$ and FLG were so extreme that forcing them to non-significant levels in the modified corpus could affect the distribution of the co-occurring features (e.g., increasing verbs associated with pronouns) or produce artificial changes that could harm the content preservation and the naturalness of the resulting answer. For all the other features, the modifications reached the goal of non-significant differences between the two corpora. Table 5.2 shows an example of a modified answer to illustrate this process. Although both the counts for the co-occurring

features and the length of the answers (which influences the normalized counts) changed due to the modifications, features that were not statistically significant when compared to FLG were still not significant when compared to FLG_{mod}. A table with the estimates, standard errors, *F*-values and *p*-values for all the analyzed features can be found in Appendix D.

Table 5.2: Example of a modified answer. The left side shows the answer provided by a tourist assistant in the original data collection. The right side shows the corresponding answer, modified to portray features that mimics $DailyDialog$ linguistic form. Modified words are highlighted in bold and the tags attributed to the words are between square brackets.

Original answer (FLG corpora)	**Modified answer (FLG_{mod} corpora)**
Well there are always a lot **of** *[preposition]* **great options** *[attributive adjective, noun]* **in downtown Flagstaff** *[preposition, nouns]* **for** *[preposition]* live music at bars like The State Bar and the Hotel Monte Vista. You can see what **activities** *[nominalization]* are going **on** *[preposition]* at www.flaglive.com **for music and events** *[preposition, nouns]*.	Well **we always have** *[1st person pronoun, present verb]* places **you** can **visit** *[2nd person pronoun, present verb]* that **have** *[present verb]* live music. **I'd suggest** *[1st person pronoun, contraction, prediction modal, suasive verb]* bars like The State Bar and the Hotel Monte Vista. You can see what music and events are happening at www.flaglive.com.

In summary, for each question-answer pair in the original FLG, there is a corresponding question-answer in FLG_{mod}, where the answer has equivalent informational content but is expressed in a different register. The patterns of language use in FLG_{mod} mimic those in $DailyDialog$, which is, on average, more personally involved and oral, with additional features for persuasion and formality. To evaluate the quality of the modifications, we asked human subjects to compare the content in the original and modified versions of the answers and assess the naturalness of the modified answers, as detailed in the following section.

5.2 Validation of Modifications

One important aspect of text modification is that both the informational content and basic linguistic integrity of the text should be preserved (Sato, 2007; Bosher and Bowles, 2008; Long and Ross, 1993). Therefore, we performed a validation study where participants proof-read the original and modified answers and assessed the modifications in terms of content preservation and quality attributes, such as naturalness and meaningfulness.

5.2.1 Procedures

We randomly selected 54 (10%) question-answer pairs from the parallel corpora and asked participants to judge the content preservation and naturalness of the answers. We collected data via an online questionnaire, where each participant assessed two blocks of questions, presented in random order. In one block, participants were presented with a tourist question and two possible answers (A and B), which correspond to the original answer from FLG and modified version from FLG_{mod}. Participants were unaware of how the answers A and B were produced. For each question-answer pair presented on the screen, participants were invited to rate how similar the information provided in the answers was, regardless of how the messages were written. Participants used a slider to select the similarity level, where the extremities of the sliders were labeled with "Completely different" (0) and "Exactly the same" (100). One example of a content preservation question is presented in Figure 5.2. Each participant assessed 10 randomly selected question-answer pairs for content preservation.

In the second block of the assessment, we were interested in the naturalness of the answers. We selected a subset of 27 question-answer pairs from the original set of 54.

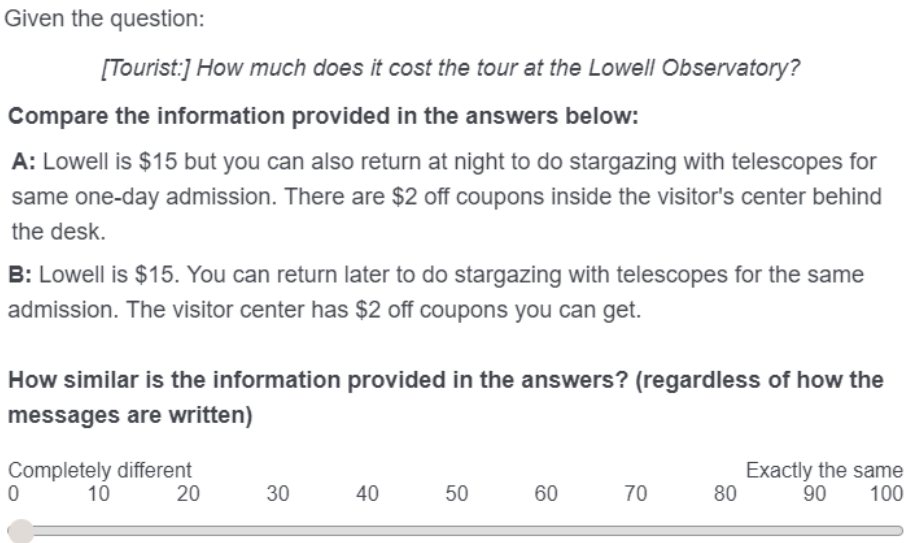


Figure 5.2: Example of a content preservation question. Participants selected their answer to the question using the slider, where 0 represents that the content in the answers is completely different, while 100 represents that the content in the answers A and B is exactly the same.

Participants were presented with a question and one single answer, either from FLG or FLG_{mod}, at a time. Considering the question-answer on the screen, participants were invited to use a seven-point Likert scale (1:completely disagree/7:completely agree) to rate the answer on four dimensions: natural, complete, meaningful, and well-written. Each participant rated nine question-answer pairs randomly selected from the 54 possible (27 answers from FLG and 27 answers from FLG_{mod}), as well as one attention check.

To evaluate the content preservation, we fitted an intercept-only, mixed effect linear model (Bates *et al.*, 2015) with the score for content similarity as the dependent variable. The random effects are the questions and the participant identification. We considered the content preservation for a modification to be reasonable (i.e., content essentially the same) if the estimate for the intercept (taking into account the standard error) stayed above the upper quartile ($\theta \pm \sigma > 75$).

The naturalness was evaluated using a Cumulative Link Mixed Model (CLMM) for ordinal data (Christensen, 2019). The four rated items were individually assessed, and the ratings per item were combined into negative ratings (1-3), neutral (4), and positive ratings (5-7). We fitted a model with the rates for each item as the dependent variable, the corpora as the independent variable (original vs. modified) and three random effects, namely the participants, the questions, and the items per question.

5.2.2 Participants

Participants were recruited through Prolific [1] , in February 2020. Prolific is an online recruitment service explicitly designed for the scientific community to enable large-scale recruitment of willing research participants (see (Palan and Schitter, 2018)). We recruited a total of 90 participants, but two were later discarded due to failure to answer the attention check ($N = 88$). Most participants were female (56), and the age range was 18-77 ($\mu = 33.6$ years-old, $\sigma = 11.3$). All the participants claimed English as their first language and were located in USA territory.

5.2.3 Results

The content preservation dataset contains 880 observations (10 observations per participant). Each question-answer pair was evaluated from 14 to 18 times.

The model shows that the estimated mean for content similarity is 86.78 (SE=1.39, df=99.5), and the confidence interval is $(84.1, 89.5)$, which is reasonably above the upper quartile (75). Both random effects are significant, and the effect of participants has the largest variance. Hence, we conclude that the linguistic modifications made in producing

[1] https://www.prolific.co

the FLG_{mod} corpus reasonably preserved the content of answers in FLG.

Table 5.3: CLMM random effects. The variance explained by the participants is larger than the residual variance, which shows that a significant portion of the variation is influenced by participant biases.

Groups	Variance	Std.Dev.
PID (Intercept)	143.674	11.986
Question (Intercept)	9.112	3.019
Residual	102.486	10.124

Regarding naturalness, the number of positive ratings is consistently higher than negative ratings for both corpora (see Table 5.4). The CLMM models show that the scores for every item do not significantly vary, as presented in Table 5.5. Additionally, the participant biases account for a lot of variance (see Appendix E for random intercepts results). Hence, we conclude that the answers in FLG_{mod} are not significantly different in terms of naturalness from the answers in FLG.

In summary, the text modification process produced a parallel corpus, FLG_{mod}, with equivalent informational content as the FLG corpus, but with the register characteristics of the $DailyDialog$ conversations. Validation study indicates that the modifications introduced to generate the FLG_{mod} corpus preserved the content and the naturalness expressed in FLG.

Table 5.4: Counts of scores per group. The table reports the number of times the original and modified answers received a negative, neutral, and positive scores. Both original and modified answers consistently received more positive than negative scores.

group	Negative (1-3)	Neutral (4)	Positive (5-7)
Original (FLG)	180	84	1317
Modified (FLG_{mod})	255	105	1210

Table 5.5: CLMM results per evaluated item. The table shows the estimate, standard error, Z-values, and p-values for each item.

Item	Estimate	SE	z	Pr ($>|z|$)
Natural	-0.61	0.32	-1.90	0.06
Meaningful	0.07	0.051	1.42	0.16
Complete	-0.10	0.29	-0.34	0.74
Well-written	-0.58	0.35	-1.66	0.10

Chapter 6

USER STUDY 1: THE USER PERCEPTIONS

In this chapter, we finally start addressing the motivating question that drives my effort: investigating whether users are sensitive to changes in conversational register and how such differences in register impact perceived quality of the interaction and overall user experience. Because user experience is a broad concept, we begin this chapter by discussing user experience in the context of this research (see Section 6.1). After that, in Section 6.2, we present the procedures and results of the study of user perceptions, which focuses on evaluating whether there is an association between the frequency of core linguistic features and user preferences with respect to language use.

6.1 User Experience

User experience (UX) refers to the overall experience of a person using a software product. According to the ISO 9241-11 (ISO 9241-11, 2018), user experience includes the user perceptions and responses that result from the use of a system, including emotions, beliefs, preferences, perceptions, comfort, behaviors, and accomplishments that occur before, during, and after use. Because user experience is a very broad concept, it is crucial to delimit the scope of the term for this particular study.

User experience is often measured in terms of usability metrics, such as effectiveness, efficiency, and satisfaction (McNamara and Kirakowski, 2006; Radziwill and Benton, 2017; Finstad, 2010) [1] . In this research, however, we controlled for the chatbot's technology

[1]Effectiveness, efficiency, and satisfaction are part of the general definition of usability, according to the

and knowledge as well as the user's tasks, since the main goal was to evaluate the user perceptions regarding the varying patterns of language use. Thus, the usability constructs, such as task success, robustness, and ease of use, was equivalent across treatments.

Thus, for this research, user experience is defined in terms of attitudinal metrics. Specifically, we measured specific attributes that are potentially influenced by the user's expectations and perceptions of the chatbot's language use. Since the conversation's participants and the relationship among them are pointed out in the situation analysis framework (Biber and Conrad, 2019) as characteristics that influence the register, we evaluated whether the register variation influence how *appropriate* is the language, given the chatbot's social role. The perceived appropriateness of a tourist assistant's language is influenced by the tourist's expectations concerning the assistant's communicative behavior and the stereotypes of the social category (Jakic *et al.*, 2017; Krauss and Chiu, 1998).

The perceived communicative behavior might also influence how the users perceive the chatbot's *credibility*. As we discussed in Chapter 2, the failure to convey expertise compromises credibility (Zumstein and Hundertmark, 2017), which is the ability of the service provider to be believable and trustworthy (Sweeney and Swait, 2008; Mack *et al.*, 2008). According to Corritore *et al.* (2005), credibility consists of four factors: honesty, expertise, reputation, and predictability. In this study, we focus on the expertise and honesty factors, which represent a chatbot's perceived competence and believability. Since the participants have no previous experiences with the studied chatbot, reputation and predictability factors did not apply.

ISO 9241-11 (ISO 9241-11, 2018). Usability, in turn, is part of the overall users experience and, in many cases, the two terms are used interchangeably (Tullis and Albert, 2013). In this research, we differentiate usability from user experience, where usability is the ability to carry the task successfully and user experience focuses on the user perceptions and behavior resulting from the interaction (Tullis and Albert, 2013).

The literature also shows that using appropriate language might increase the chatbot's human-likeness (Chaves and Gerosa, 2020) and enhance the feeling of being together (Katkute *et al.*, 2017). Hence, we expect that register-appropriate language positively influence a chatbot's *social presence* (Ciechanowski *et al.*, 2018; Go and Sundar, 2019; Katkute *et al.*, 2017). In CMC (Computer-Mediated Communication) fields, social presence describes the degree of salience of an interlocutor (Short *et al.*, 1976; Gunawardena and Zittle, 1997), in this case, the chatbot, and how it can project itself as an individual. The chatbots in this research were introduced as such, and we did not intend to deceive the participants into thinking that they were talking to a human. Hence, the only anthropomorphic cue we manipulated was the language use, and we did not measure the chatbot's perceived humanness (i.e., whether the participant believes that the interactional partner is a human). Instead, we focused on identifying whether anthropomorphic cues are more evident to the participants when the chatbot uses appropriate language.

Finally, because the social orientation toward chatbots is a determinant of the overall user experience (Liao *et al.*, 2016; Go and Sundar, 2019; Ashktorab *et al.*, 2019), we also expect that the a participant's preference regarding human-like social interactions influence the overall user experience, although we did not anticipate what registers would be preferred for each individual profile.

6.2 User perceptions

To explore whether users are sensitive to changes in conversational register and how differences in register impact perceived quality of the interaction, we designed a user study where we asked participants to compare answers from the parallel corpora (FLG and FLG_{mod}) and express their preferences in terms of appropriateness, credibility, and

overall user experience. In this study, participants compared the conversational excerpts by reading the question-answer pairs presented side-by-side. Because there was no interactivity, the anthropomorphism and social presence metrics were left out.

Considering that human tourist assistants produced the language in FLG, and therefore it is likely to be register-appropriate to the proposed interactional situation, we hypothesized that findings would show a preference for answers from FLG; the answers from FLG_{mod} should score lower, as we have artificially modified them to introduce a register that is less likely to be appropriate to the situational parameters in FLG. The analysis additionally included variables to represent the individual interlocutors (both assistants and participants) to compare the strength of these variables when compared to the register variation expressed in the corpora. Considering that *what* is said is similar between FLG and FLG_{mod}, user experience should be mostly influenced by *how* it is said. In the following subsections, we present the method used to evaluate user perceptions of the two parallel corpora and its outcomes.

6.2.1 Procedures

We selected 10% of the question-answer pairs from the parallel corpora to be evaluated. Considering the semi-manual process, in some cases the answers to a given question were very similar between the two corpora, i.e., one particular answer may not have been modified at all as part of the register-shifting process. To focus the comparison on answers with distinct differences in register, we selected answers that had been substantially modified: we calculated the Levenshtein distance (Kessler, 1995; Wikibooks, 2020) between the pairs of original and modified answers. The Levenshtein distance measures the difference between two strings by calculating the minimum number of single-character edits (Wik-

ibooks, 2020) that separate the two strings from each other. We selected the 54 question-answer pairs with the highest values of Levenshtein distance (see Appendix E for the evaluated question-answers pairs and the corresponding Levenshtein distance values).

We collected participant responses via an online questionnaire. After reading and agreeing with the informed consent, participants were introduced to the task: given a tourist question, identify the answer that best represents a tourist assistant's discourse. For this experiment, participants were told that the answers would be provided by a chatbot. For each tourist's question, presented in the screen one at a time, participants could choose one out of three options: the original answer (from FLG), the modified version (from FLG_{mod}), and "I don't know" (see an example in Figure 6.1). Original and modified answers were presented in a randomized order, whereas "I don't know" was always the last option on the list. Appendix F contain a sample of this instrument for the three measured constructs.

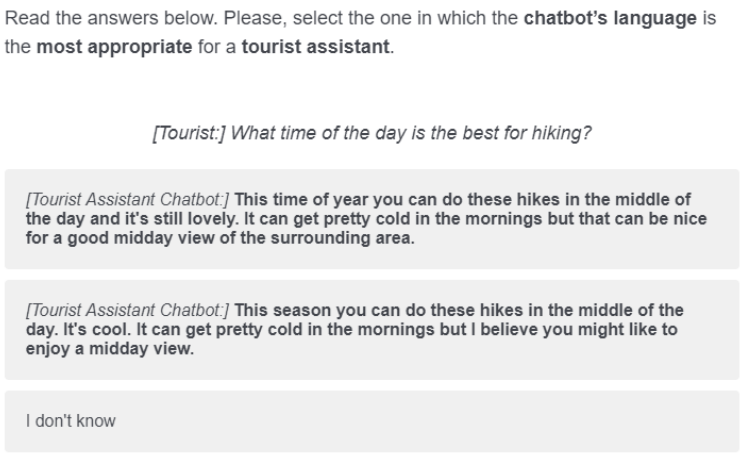


Figure 6.1: Example of a question from the user study 1. In this example, the participant was invited to select the answer that portrays the most appropriate language. Participants selected their responses by clicking on their preferred answer or on the "I don't know" option.

Constructs were also evaluated one at a time. Thus, we first showed a definition of

the construct of interest (e.g., appropriateness), and then the ten question-answer pairs to be evaluated for that construct, which were nine question-answer pairs extracted from the corpora and one attention check. In total, each participant evaluated 27 different question-answer pairs (9 per construct, without repetition across constructs) randomly selected from the possible 54. The order of the constructs was also randomized. In the end, participants answered the demographics and social orientation questionnaire. We used the social orientation items proposed by Liao *et al.* (2016); the social orientation toward chatbots determines the participant's preferences regarding human-like social interactions with chatbots (Liao *et al.*, 2016).

The outcome of the questionnaire consists of the user preferences regarding which answer (original vs. modified) is the most appropriate and credible for a tourist assistant, as well as the answer that would result in the best user experience.

6.2.2 Participants

Participants were recruited through Prolific, in March 2020. We received a total of 193 submissions, 15 of which were discarded due to either technical issues in the data collection or failure to answer the attention checks ($N = 178$). All the participants claimed English as their first language and were located in the USA. Most participants had either a four-year bachelor's degree (59) or some college, but no degree (49). 21 participants had Master's degree, and the other 21 graduated from high school. Common educational backgrounds were STEM (54), Arts and Humanities (42), and Other (32). Three participants had non-binary gender, 86 declared themselves as female, and 89 as male. The age range is 19-73 ($\mu = 33.10$ years-old, $\sigma = 11.08$).

168 participants declared that they search for travel information online, but 99 declared

that they had never used online assistance when traveling. Only 21 participants had never interacted with chatbots before, and 82 said they had interacted five or more times. Appendix F contains plots that show the distribution of participants for demographics and profile variables.

We also recruited 12 participants from the Northern Arizona University community (students and staff) to perform the study in the lab. They answered the same online questionnaire as the participants recruited from Prolific; however, we invited them to think aloud to explain the reasoning behind their preferences. Two researchers sat with participants in the lab during these sessions and independently took notes on the participant's comments. After every session, the researchers debriefed about their notes and merged the outcomes. We used these notes to support the discussion presented in Section 6.3, pointing out quotes from participants that are relevant for understanding the user perceptions of particular linguistic features. The 12 answers to the questionnaire were not included in the statistical analysis to avoid data source bias.

6.2.3 Analysis of the linguistic features

To model the user preferences between original and modified versions of FLG, we fitted a generalized linear model (GLM) for two-class logistic regression, using the $glmnet$ package in R (Friedman *et al.*, 2010). Because we are interested in the difference between original and modified versions of FLG for each linguistic feature of interest (represented in Figure 4.1), we calculated the original $-$ modified counts, and input this difference into the model. For example, suppose that one particular answer provided by the tourist assistants in the original FLG corpus has 31.3 *private verbs* (normalized per 1,000 words), and that after linguistic modification, the corresponding answer has 62.5 *private verbs*.

Thus, for this particular answer, the value for *private verbs* input into the model is $31.3 - 62.5 = -31.2$. A negative value means that the occurrences of that feature were increased in the modification process for that particular answer. In contrast, a positive value for a feature means that the occurrences of that feature were reduced in the modification process.

Problem Definition

Consider a model with the response variable $Y = \{0, 1\}$ where 0 represents the original answers (FLG) and 1 represents the modified answers. The L1-regularized logistic regression algorithm models class-conditional probabilities through a linear function of the predictors (Friedman *et al.*, 2010). The prediction function is defined as:

$$f(x) = w^T x + \beta$$

where x is a feature vector of p real numbers representing (i) the difference between original and modified counts per linguistic feature; (ii) variables representing the participant who answered the question (1 if the observation was answered by that participant, 0 otherwise), the participant's self-assessed social orientation (1 to 7), and the author of the answer (1 if the answer was authored by that tourist assistant, 0 otherwise). We want to learn a p vector of w weights and a real scalar intercept β. The L1-regularization ensures that the learned model has a sparse/interpretable w (some entries will be exactly zero; these entries correspond to features that are not used/important for prediction). The prediction function $f(x)$ gives real-valued predictions for the given feature vector x. The logistic link function that finds the predicted probability in $[0, 1]$ is

$$p(x) = \frac{1}{1 + \exp(-f(x))}$$

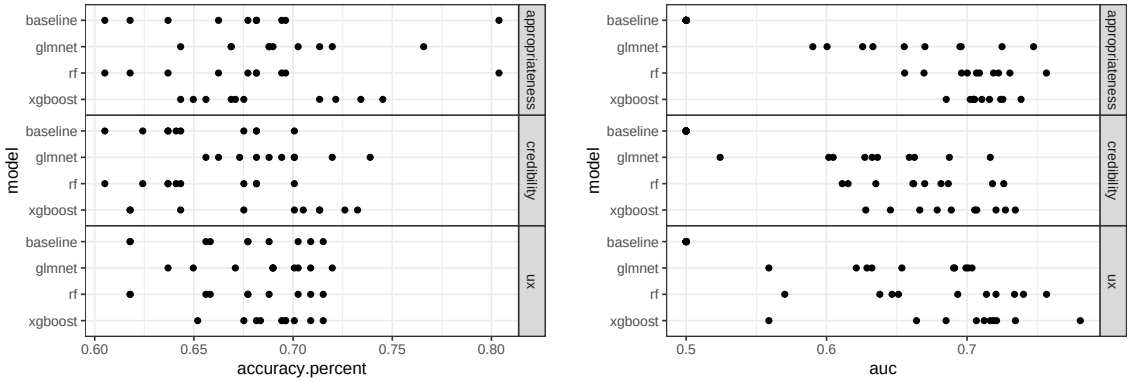
The function predicts the negative answer (i.e., 0:original) when $f(x) < 0$ and $p(x) < 0.5$, while the positive answer (i.e., 1:modification) is predicted when $f(x) > 0$ and $p(x) > 0.5$. For comparison purposes (to determine an upper bound on prediction accuracy), we fitted two non-linear learning models: random forest and gradient boosting. For the random forest, we used the `party` package in R, which provides an implementation of random forests with conditional inference trees (Hothorn *et al.*, 2006; Strobl *et al.*, 2007, 2008). For the gradient boosting, we used the `xgboost` package in R (Chen and Guestrin, 2016), which is an efficient implementation of gradient tree boosting. The measures were also compared to a baseline model, which always predicts the most frequent class in the training data (and provides a lower bound for prediction accuracy). The evaluation metrics were the accuracy, the ROC curve, and the area under the curve (AUC). To calculate these measures, we used 10-fold cross-validation. First, we randomly assigned every observation in the data set to one out of $k = 10$ folds. For each fold ID from 1 to 10, we created a test set comprising all observations with matching fold ID and used all other observations for a training set. We then used the train set to learn model parameters, and we used the test set to evaluate prediction accuracy. The cross-validation pseudo-algorithm is available in Appendix F, and the R code and datasets are available on GitHub (Chaves, 2020a).

6.2.4 Results

The evaluation dataset started with a total of 4,806 observations (178 participants, 27 evaluations per participant). From this total, participants skipped the question without

answering for 11 observations. In 77 others, the participants signaled that they did not have a preference (the "I don't know" option). These observations were discarded from the analysis, resulting in a dataset with 4,718 observations. Each question-answer pair was evaluated from 24 to 35 times per construct. As we expected, participants overall preferred the answers from the original corpus, although the modified version was preferred for a few answers. A table with the number of votes per question is presented in Appendix F.

Figure 6.2 shows the prediction accuracy and AUC plots for the four fitted models. Since participants generally preferred the original FLG corpus answers, the prediction threshold is close to always predicting the most frequent class (original), which is particularly true for the random forest model. The prediction accuracy of `glmnet` and `xgboost` are only slightly better than the baseline. Nevertheless, the AUC plot indicates that the models are learning something important (the ROC curve plot is available in Appendix F), as the AUC values are consistently better than the baseline. Additionally, the `glmnet` model consistently selects the same variables to the k-folds.



(a) Accuracy (b) AUC

Figure 6.2: Accuracy (a) and AUC (b) results per model for each construct (appropriateness, credibility, and user experience). The baseline represents a model that always predicts the most frequent class (original). Accuracy percentage shows that glmnet, random forest, and xgboost perform only slightly better than the baseline model. AUC, however, is reasonably better than the baseline for the three models.

Note that the non-linear models are not considerably more accurate than the linear model. Although there seems to be some non-linear trend in the data, it does not justify the complexity of the non-linear models. Hence, we used the results from `glmnet` model to interpret the coefficients and identify the linguistic features that determine user preferences.

Coefficients

Table 6.1 presents the coefficients of the linguistic features selected in six or more folds. The first and second columns indicate, respectively, the linguistic feature of interest and the sign of original − modified calculation, which indicates whether one particular feature was increased or decreased in the text modification process. A positive sign (+) for a feature f_i indicates that $\text{count}_{\text{original}}(f_i) > \text{count}_{\text{modified}}(f_i)$, while a negative sign (-) indicates that $\text{count}_{\text{original}}(f_i) < \text{count}_{\text{modified}}(f_i)$. The following three columns present the mean of the coefficients and the standard deviation for each construct. Features with negative coefficients increase the likelihood of the model predicting the original class. In contrast, features with positive coefficients increase the likelihood of the model predicting the modified class. Appendix F include plots of coefficients for each construct.

Original answers have significantly more *coordinating conjunction* clauses and *attributive adjectives* than the modified versions. These features have a negative coefficient for all the three constructs, which indicates that frequent occurrences of these features increase the likelihood of original answers being chosen; participants are more likely to prefer answers in which these features are more frequent. The same conclusion applies to *suasive verbs* and *causative subordination*, although these features shows up as relevant only for the user experience construct.

Table 6.1: Coefficients and standard deviation of the non-zero variables per construct. Only linguistic features were selected as relevant variables for predicting users choices, and most of the selected linguistic features are relevant for all the three constructs. The columns original − modified represents whether the original answers have more (+) or less (−) occurrences of that particular feature. The dots indicate that the corresponding feature was not selected for that particular construct.

		Mean of coefficients ± Std. Deviation		
Linguistic features	orig. − mod.	Appropriateness	Credibility	User Experience
Coordinating conj. clause	(+)	-0.032 ± 0.003	-0.031 ± 0.005	-0.038 ± 0.005
Attributive adjective	(+)	-0.006 ± 0.001	-0.006 ± 0.001	-0.007 ± 0.002
Conditional subordination	(+)	-0.002 ± 0.002	0.005 ± 0.001	·
Nouns	(+)	0.002 ± 0.001	0.004 ± 0.001	0.003 ± 0.001
Prepositions	(+)	0.002 ± 0.001	0.001 ± 0.001	-0.001 ± 0.001
WH-relative clause (subj. pos.)	(+)	0.024 ± 0.004	0.034 ± 0.004	0.024 ± 0.004
Final preposition	(+)	0.035 ± 0.007	0.038 ± 0.010	0.017 ± 0.012
Suasive verbs	(+)	·	·	-0.015 ± 0.003
Causative subord.	(+)	·	·	-0.007 ± 0.005
Third person pronoun	(+)	·	·	0.012 ± 0.002
Adverbial–conjuncts	(-)	-0.029 ± 0.007	-0.017 ± 0.004	-0.056 ± 0.003
Prediction modals	(-)	-0.008 ± 0.001	-0.006 ± 0.002	-0.011 ± 0.001
Contractions	(-)	-0.002 ± 0.001	-0.008 ± 0.001	·
Second-person pronoun	(-)	-0.002 ± 0.001	-0.002 ± 0.001	·
First-person pronoun	(-)	0.007 ± 0.001	0.006 ± 0.002	0.009 ± 0.004
Private verbs	(-)	0.006 ± 0.001	·	·
Present verbs	(-)	0.003 ± 0.001	0.005 ± 0.001	0.003 ± 0.001
That-deletion	(-)	0.008 ± 0.003	·	0.004 ± 0.003
Time adverbials	(-)	0.018 ± 0.005	0.005 ± 0.003	0.007 ± 0.005
Indefinite pronoun	(-)	·	0.006 ± 0.004	·

Original answers also have significantly more *nouns*, *WH-relative clause (subject position)*, and *final preposition*. These features have a positive coefficient for all three constructs, which indicates that frequent occurrences of these features increase the likelihood of modified answers being chosen. This outcome suggests that participants are more likely to prefer answers in which these features are less frequent. The same conclusion applies to *third person pronouns*, but this feature is relevant only for the user experience construct.

Two features showed inconsistent outcomes across constructs. *Conditional subordination* has a negative coefficient for appropriateness, but a positive coefficient for credibility.

This outcome suggests that frequent occurrences of these features influence appropriateness positively while influencing credibility negatively. Additionally, *prepositions* have a positive coefficient for both appropriateness and credibility, indicating that participants are more likely to choose the answers where these features are less frequent. In contrast, *prepositions* have a negative coefficient for user experience, indicating that participants tend to choose answers where this feature is more frequent for this construct. However, when we aggregate the estimate to the standard deviation, it sums up to zero, suggesting that this outcome may be noise.

Modifications have significantly more *adverbial–conjuncts* and *prediction modals* than the original answers. These features have a negative coefficient for all three constructs, which indicates that frequent occurrences of these features increase the likelihood of original answers being chosen. This outcome suggests that participants are more likely to prefer answers in which these features are less frequent. The same inference applies to *contractions* and *second-person pronouns*, although these features did not show up as a relevant feature for user experience.

Modifications also have a larger number of *first-person pronouns*, *present verbs*, and *time adverbials*. These features have a positive coefficient for all three constructs, which indicates that increasing the occurrences of these features increased the likelihood of modified answers being chosen. This outcome highlights the participant's preferences for answers in which these features are more consistently present. The same conclusion applies to *private verbs*, *that-deletion*, and *indefinite pronouns*. However, *private verbs* help to predict appropriateness only; *that-deletion* predicts appropriateness and user experience only; and *indefinite pronoun* is relevant for credibility only.

In conclusion, the results clearly show that *the use of register-specific language has a*

significant impact on user perceptions of conversational quality for tourist assistant chatbots. There is an association between register-specific use of particular linguistic features and perceived quality; linguistic features are stronger predictors of appropriateness, credibility, and user experience than individual characteristics of interlocutors (either assistants or users). The variables representing the tourist assistants who produced original answers, as well as those representing the individual participants and their social orientation, were not selected as predictors of user preferences regarding chatbot language use, since these factors were not relevant.

6.3 Discussion

The results of the user perceptions study presented in this chapter show that users are sensitive to variation in the linguistic register and, specifically, that register has a significant impact on user perceptions of conversational quality; a chatbot that adopts the wrong conversational register risks losing credibility and acceptance by users. In this section, we discuss the findings presented in this chapter. We use quotes from the participants of our in-lab data collection to support our findings.

6.3.1 Certain linguistic features are preferred when efficiency matters

Like other information search interactions, the interactions in the corpora used in this research are goal-oriented. Particularly for en-route tourist information searches (see Section 2), users are often pressed for time; efficiency in finding the target information is critical (Think with Google, 2016; Tussyadiah, 2020; Wang *et al.*, 2016). Efficiency is also a priority in other task-oriented domains where chatbots operate, such as customer service (Brandtzaeg and Følstad, 2018). The statistical modeling revealed several linguistic

features as particularly important in supporting compact, efficient information sharing. First, the linguistic feature *coordinating conjunction* had the largest coefficient in the analysis; it is common in conversations due to real-time production constraints. Complex sentences in a conversation are often a linear combination of short clauses with a simple grammatical structure (Biber *et al.*, 1999), usually connected by *coordinating conjunctions*. FLG_{mod} mimics $DailyDialog$ form, which portrays language that is more elaborated and carefully planned to achieve educational goals. The elaborated complexity leads to varying grammatical structure levels, which can reduce efficiency when providing information. Similarly, participants also often preferred answers with more verbs and fewer *nouns*, which is a pattern present in the modified versions of the answers (FLG_{mod}). This outcome indicates a preference for active language rather than descriptive (Biber, 1988). For example, when comparing the answers

> FLG: There are $2 off coupons *inside the visitor center behind the desk.*
> FLG_{mod}: The visitor center has $2 off coupons you can get.

a participant of the in-lab session stated that the first one (FLG) "gives more information, but it's unnecessary" and "takes more time" [LabP2].

Participants were also more likely to choose answers with fewer *WH-relative clauses* and *adverbial–conjuncts*. *WH-relatives* are "often an extra-piece of information that might be of interest" (Biber *et al.*, 1999), for example, regarding the answers

> FLG: There is the Discovery Map, **which** is more geared toward visitors [...]
> FLG_{mod}: The Discovery Map is more geared toward visitors

[LabP8] stated that FLG_{mod} version was "more clear" with "less fluff to the sentence." *Adverbial–conjuncts* are used to connect sentences in discourse (Biber *et al.*, 1999). Some are more frequent in face-to-face conversations (e.g., so, then, anyway), while others are

more common in written language (e.g., however, therefore, although) (Biber *et al.*, 1999). As FLG_{mod} mimics $DailyDialog$ which focuses on language learning, the most common *adverbial–conjuncts* align with the ones that are common in written registers. These conjuncts increase the formality of the answers (Biber, 1988), and interactive chatbot users may see these words as unnecessary, filler words. In the lab sessions, participants stated that "when there is a lot of information, some filler words can be left out" [LabP1]. For example, when comparing the answers

> FLG: You cannot leave it in 15-minute parking for an extended period of time. On the Amtrak side of the building, there is a paid parking lot.
>
> FLG_{mod}: You cannot leave it in 15-minute parking for more than that. **However**, the Amtrak side of the building has a paid parking lot you could use.

a participant mentioned that the reason for the preference is that "they don't have the 'however,' and lead directly to the next sentence" [LabP8]. Additionally, these conjuncts imbue a style of formality that may create distance between the chatbot and the user.

Finally, the preference for *that-deletion* shown by the analysis is likely influenced by its frequent co-occurrences with *suasive* and *private verbs*. However, *that-deletion* "has colloquial associations and it is therefore common in conversations" (Long and Ross, 1993), since conversations favor the omission of unnecessary words to accommodate real-time production constraints. Hence, user preferences for *that-deletion* in the data could be associated with the preference for efficiency in conversations.

6.3.2 Certain linguistic features impact the perception of human-likeness

The literature on human-chatbot interactions has extensively explored the need for chatbots to be "human-like." On the one hand, scholars grounded in media equation theory (Nass *et al.*, 1994; Fogg, 2003) have shown that people prefer agents who reflect human

social and conversational protocols, e.g., conform to gender stereotypes associated with tasks (Forlizzi *et al.*, 2007); self-disclose and show reciprocity when recommending (Lee and Choi, 2017); demonstrate a positive mood (Hayashi, 2016), and so on. On the other hand, overly humanized agents can create inaccurate expectations in users (Gnewuch *et al.*, 2017) and result in the "uncanny effect" (Appel *et al.*, 2020; Ciechanowski *et al.*, 2018). However, the idea of assigning a social role to a conversational agent does not necessarily imply deceiving people into thinking the software is human; a chatbot can be clearly identified as such, but still benefit from approximating its conversational register to the patterns of human-human communication–several specific findings in the current analysis support this observation.

In the Natural Language Generation field, the aggregation of sentences using *coordinating conjunctions* is commonly used to increase fluency and readability (Reiter and Dale, 2000). According to Reiter and Dale (2000), aggregation of sentences can be de-emphasized if the text is obviously produced by a computer; this suggests that participants would not care about slightly stilted language, since they were told that the answer was produced by a chatbot. The analysis reveals, however, that users have a preference for language that is more human-like, with fewer pauses and more coordination.

First-person pronouns can also increase human-likeness, although the plural form is preferable. The singular form ("I") unambiguously indicates the speaker, whereas referring to the speaker's identity in the plural form ("we") varies according to the context (Biber *et al.*, 1999). Choosing between singular or plural forms is a strong indicator of the identity that the chatbot conveys. When the chatbot says "I," it clearly refers to itself, but when it says "we," it can be interpreted as a general reference to its social category. Using "we" softens the role of the chatbot and highlights its representative role of a broader entity

(e.g., professional tourist assistants, visitor center's representatives, Flagstaff's tourism personnel). Both singular and plural forms convey personal involvement, but "we" may demonstrate more credibility because the chatbot is more likely to be recognized as part of a community of knowledgeable individuals. Although both singular and plural pronouns are measured under the same linguistic feature, the evidence presented here shows that participants preferred the plural use of this feature. As often occurs in $DailyDialog$, the singular form ("I") co-occurred with the *prediction modal* "would" in FLG_{mod} to make suggestions or to give advice. Noticeably, participants preferred answers where *prediction modals* are less frequent, which indicates that the singular form of *first-person pronouns* is unlikely to influence the user's preference for this feature. Quotes from lab sessions also support that participants preferred the plural form. For example, regarding the singular form, [LabP1] stated: "I don't like when the chatbot says 'I,' it seems the developer is trying to trick me to think the bot has opinions. When chatbots use 'I' it sounds too much like pandering." In contrast, when comparing the answers

> FLG:"There are 50 miles of trails within Flagstaff [...]"
> FLG_{mod} "We have 50 miles of trails within Flagstaff [...]"

[LabP3] stated that "the use of the word 'we' makes it more personable than simply saying 'there is this' " [LabP3]. As these findings reveal, chatbot language should conform to the expectations of its social category, as previous literature has suggested (Gnewuch *et al.*, 2017). Still, the register of chatbot conversation must also consider its artificial nature, particularly positioning the agent as a representative of a broader entity.

6.3.3 Certain linguistic features impact the perceived level of personalization

Previous literature has shown the benefits of personalized interactions with chatbots (Thies *et al.*, 2017; Shum *et al.*, 2018; Duijvelshoff, 2017), particularly in domains where the chatbot needs to build rapport and trustful relationships with the user, such as in financial services (Duijst, 2017), companion (or buddy) chatbots (Thies *et al.*, 2017; Shum *et al.*, 2018), and recommendation systems (Cerezo *et al.*, 2019). At the same time, users might feel uncomfortable with some aspects of personalized content (Thies *et al.*, 2017). In this study, tourist assistants did not provide personally tailored information, as they had few or no clues about the tourists' identities or preferences due to the text-based nature of the conversations. In this inherently rather impersonal content, results showed that participants preferred general rather than personalized information.

Participants preferred lower frequencies of occurrences of *second-person pronouns* (e.g. "you"), which are used as a direct address to the user (Biber *et al.*, 1999). Giving that information search interactions focus on the assistant providing information without necessarily sharing a personal relationship with the user, it may sound inappropriate for a chatbot to use *second-person pronouns*, rather than simply stating the information. In the lab sessions, for instance, a participant stated that "[the chatbot] saying 'you' implies a lot of personalization but in a pressuring type of way" [LabP6]. *Second-person pronouns* (as well as singular *first-person pronouns*) often co-occur with *contractions*, which may justify the participants preferences for answers with low frequency of *contractions*.

The appropriateness of *conditional subordination* also relates to personalization. The use of conditional clauses as a mechanism for inserting suggestions, requests, and offers is common in conversation (Biber *et al.*, 1999). The tourist assistants in FLG used *conditional*

subordination to offer different options to the users (e.g., *"if you..., you can/will/would")*, since their individual preferences were unknown. In the lab session, [LabP1] mentioned "prefer the [original answer] because it says 'if you stop' rather than 'I'd suggest you stop'." Moreover, *conditional subordination* helps to frame the subsequent discourse (Biber *et al.*, 1999), which was also observed by [LabP1]: "I like the 'if you are looking for maps' as it sets up the scenario that this information would be useful for" [LabP1]. However, the *conditional subordination* is negatively associated with credibility, which may indicate that when the chatbot gives options, it sounds as if it is not confident about the information provided. It is important to observe the impact of the varying communication purposes here. Although the chatbot is presented as an information provider, the tourist assistants interpreted some tourist questions as requests for recommendations (see more in Chaves *et al.* (2019a)) rather than information. Recommendations are more inherently personalized than information search results, and consequently require more personalization in their expression (see e.g., (Gavalas and Kenteris, 2011; Ricci *et al.*, 2011; Komiak and Benbasat, 2006)). Thus, participants may have expected that the tourist assistants would provide more personal, tailored content rather than conditional options. Clearly, the dynamics that shape these interactions are subtle, and the influence of sub-registers, i.e., the variation in language use to match specific communicative purposes, will need to be explored in more detail to evaluate these assumptions.

6.3.4 Avenues for future investigation

The study we presented in Chapter 6 exposed deeper complexities that point to the need for further exploration. For instance, we found inconsistent results regarding *private verbs* when comparing the cross-validation results to the quotes from lab session

participants. The private verbs used in $DailyDialog$, and consequently included in the FLG_{mod} corpus, convey a certain level of uncertainty (e.g. "guess, " "think," or "believe"). The cross-validation model found an association between the high frequency of *private verbs* and appropriateness, with no effect on the other two constructs (credibility and overall user experience). However, several quotes from lab session participants suggest that *private verbs* did make certain answers less credible. For example:

> " 'I guess' is not the tone I want." [Lab1P]
>
> "I don't like the bot saying 'I guess', it sounds passive aggressive" [Lab2P]
>
> "I don't like how the [modified answer] says 'okay, I believe'. This makes it sound like it doesn't know." [LabP4]

Considering such qualitative feedback from these participants, we believe that the lack of any negative influence of *private verbs* on credibility shown by the analysis may be conditioned by their co-occurrences with other features; this needs to be further investigated.

Results also showed a positive association between *attributive adjectives* and all the three evaluated metrics (appropriateness, credibility, and user experience). However, this association may be too coarse-grained, as the way in which *attributive adjectives* were often used in the specific conversational context of tourism advising is somewhat atypical. The typical use of *attributive adjectives* in conversation is to describe some physical attribute of an object (Biber *et al.*, 1999), e.g., "new," "big," or "smelly." In contrast, *attributive adjectives* in the FLG corpus were more often used *to classify* rather than describe the corresponding *nouns*. For example, common *attributive adjectives* are "local business," "national park," and "natural landmark." Using classifying *attributive adjectives* adds detail to the information without loss in efficiency. Participants mentioned that the *attributive adjective* "makes [the answer] more interesting" [LabP7], explaining the association with

quality metrics, but this may apply only to classifying *attribute adjectives*. Here too, further investigation will be needed to clarify the validity of the observed association.

Finally, *suasive verbs* are typically used to express the degree of certainty associated with the information that the sentence communicates (Biber *et al.*, 1999). For example, when the tourist assistant says "I recommend," it represents how much it believes the tourist should take that advice. [LabP3] observed this fact by stating that " 'I would recommend' seems like more of a suggestion." The analysis showed that *suasive verbs* influence the overall user experience, but were not shown to make the language more credible or appropriate. Closer analysis shows that this association, too, deserves further investigation. In the data used in this research, *suasive verbs* co-occurred mainly with the singular *first-person pronoun*. As noted earlier, the use of plural *first-person pronouns* was clearly linked to credibility; further investigation should evaluate whether the use of plural *first-person pronouns* with *suasive verbs* would increase their impact on credibility.

Chapter 7

USER STUDY 2: THE USERS EXPERIENCES

The study presented in Chapter 6 shows that register has significant impact on the user perceptions of appropriateness, credibility, and overall user experience when comparing question-answer pairs side-by-side. One question that remains is whether these perceptions stand when participants evaluate the conversations by actively interacting with the chatbots. To increase the realism of the context within which the participant behaviors are observed, as suggested by McGrath (1981), we replicated the previous study in a different setup, in which participants interacted with chatbots that used different registers–one that answered the user's questions with sentences from the FLG corpora, and another that answers with sentences from FLG_{mod}–, and then evaluated each interaction as a whole instead of looking at individual question-answer pairs.

As discussed in Section 6.1, the exploratory results from the literature suggest that perceiving a chatbot's language as appropriate may lead the user to also perceive the chatbot as credible, socially present, or anthropomorphic. However, there is a need for empirical evidence of whether using register-specific language results in higher levels of these three constructs. In this second study, we examine this question by assessing a structural model that relates the use of appropriate language to credible, socially present,

and anthropomorphic chatbots.

7.1 Procedures

7.1.1 Experimental setup

We employed a within-subjects design study, where each participant evaluated similar interactions with two chatbots. The only variable we manipulated across treatment was the register used by the two chatbots; one chatbot answers questions with sentences extracted from FLG corpus, whereas the other chatbot answers the questions with sentences extracted from FLG_{mod} corpus. The study was performed online: a landing page [1] explained the study and offered a single dialogue interface where participants interacted with both chatbots as well as answered the study's questionnaires. The chatbots were developed using IBM Watson Assistant (Biswas, 2018) and deployed as an integration into WordPress.

In the chatbot interface, the participant was presented with a summary of the instructions and was introduced to the first chatbot, namely **Tourist Assistant Chatbot A**. The participant had to follow the conversation script provided on the study page, which contained six questions about tourism in Flagstaff. After this conversation, the chatbot asked questions to evaluate the interaction.

When this first round was over, the system introduced the second chatbot, namely the **Tourist Assistant Chatbot B**. Once again, the participant followed the conversation script to interact with the Chatbot B, and answered the questionnaire at the end. Finally, the participant was prompted to choose which chatbot provided the best user ex-

[1] The study web page is available at https://rc.nau.edu/touristassistant. To test the chatbots, please use the participant ID *DefaultID*. All the data associated with *DefaultID* is discarded.

perience and answered the profiling and demographic questions. Figure 7.1 depicts the flow described above. The instruments used to measure the constructs are presented in Appendix G.1.

To account for the order effect, we developed two nearly identical systems. In the first one, the **Tourist Assistant Chatbot A** answered with sentences from FLG and **Tourist Assistant Chatbot B** answered with sentences from FLG_{mod}. The opposite setup was developed in the second system. When participants accessed the study link, a script randomly assigned them to one of the two treatments.

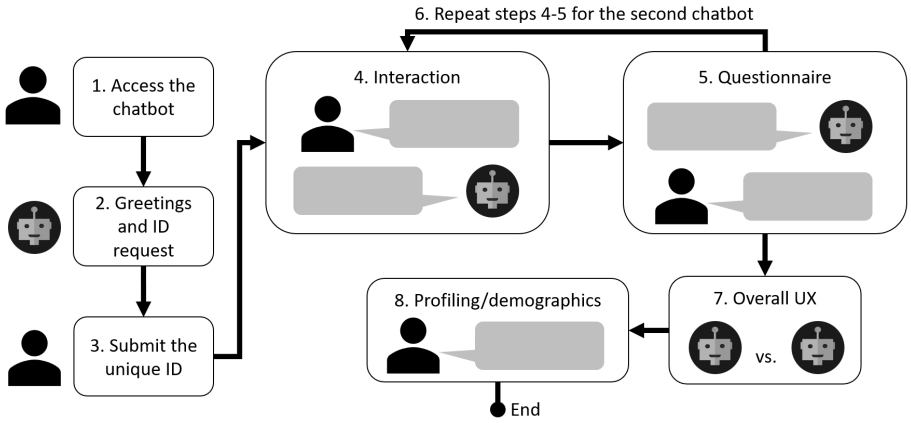


Figure 7.1: Overview of user study 2. Participants interacted with two chatbots; for each of them, they answered the questionnaire about appropriateness, credibility, user experience, social presence, and anthropomorphism. At the end of the two interactions, participants were invited to choose the chatbot that provided the best overall user experience and answered profiling and demographic questions.

7.1.2 Data analysis

We developed a Python script to separate the conversations from the questionnaire responses. The conversation transcripts were manually inspected to check whether participants followed the task instructions; the responses to the questionnaire were then stored for statistical analysis.

The statistical analysis is twofold. Firstly, we investigated whether the inferences from user study 1 stand when participants evaluate an interaction as a whole. In other words, we assessed the impact of using register-specific language on the user perceptions of human-chatbot interactions, in terms of appropriateness, credibility, social presence, anthropomorphism, and overall user experience. To achieve that, we fitted an ordinal regression with Cumulative Link Mixed Model (CLMM) (Christensen, 2019), which compares the responses from the questionnaire (scores) for the two treatments per construct. The CLMM model evaluated the score assigned by the participant (dependent variable) per treatment (FLG vs. FLG_{mod}) per construct. The model includes one interaction effect between the constructs and the treatments as well as two random effects, namely, the construct's indicators (i.e., each Likert-scale item) and the participants.

Secondly, we evaluated whether the perceived appropriateness of language can be considered a predictor of the other constructs. To assess that, we performed an analysis using PLS-SEM (Partial Least Square–Structural Equation Modeling). PLS-SEM is a covariance-based model that relates a set of independent variables to multiple dependent variables (David Garson, 2016). We implemented the PLS-SEM as a path model, which relates the predictors and their associated paths to the response variables (David Garson, 2016). Considering the definition of user experience and the constructs introduced in Section 6.1, we specified the model depicted in Figure 7.2. We evaluated the direct effects of the perceived appropriateness of language (APP) on credibility (CRED), anthropomorphism (ANTHR), social presence (SOC_PR), and overall user experience (UX). Moreover, we also assessed the role of CRED, ANTHR, and SOC_PR in mediating the effect of APP on the overall UX. Grounded on the literature that constantly shows the association between anthropomorphism and social presence Morana *et al.* (2020); Adam *et al.* (2020); Go and

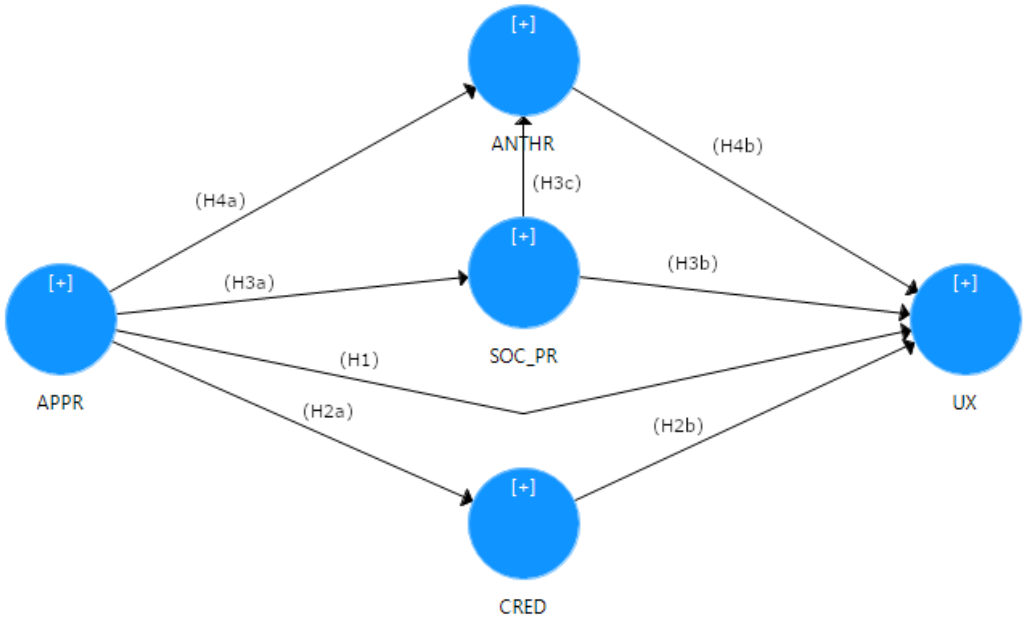


Figure 7.2: PLS-SEM structural model that describes the relationship among constructs. The model was built using SmartPLS3 software (Ringle *et al.*, 2015).

Sundar (2019); Schanke *et al.* (2020), we also evaluated the role of SOC_PR in mediating the effect of APP on ANTHR. In summary, the model evaluated the following hypotheses:

$H1$. The perceived appropriateness of the chatbot's language positively affects the overall user experience.

$H2a$. The perceived appropriateness of the chatbot's language positively affects the chatbot's credibility.

$H2b$. The chatbot's credibility mediates the effect of perceived appropriateness of language on overall user experience.

$H3a$. The perceived appropriateness of the chatbot's language positively affects the chatbot's social presence.

$H3b$. The chatbot's social presence mediates the effect of perceived appropriateness of language on overall user experience.

$H3c$. The chatbot's social presence mediates the effect of perceived appropriateness of language on anthropomorphism.

$H4a$. The perceived appropriateness of the chatbot's language positively affects the chatbot's anthropomorphism.

H4b. The chatbot's anthropomorphism mediates the effect of perceived appropriateness of language on overall user experience.

We used SmartPLS3 (Ringle *et al.*, 2015) software to analyze the structural model. The internal consistency of each latent variable was assessed through the Composite Reliability (CR) and Cronbach's alpha (α) values, which show whether the indicators (i.e., questionnaire items) of a factor are related (see (David Garson, 2016)), and the convergent validity for the constructs were assessed through the Average Variance Extracted (AVE) values, which indicates how much variance of the indicators is explained by the factors (see (David Garson, 2016)).

7.1.3 Participants

Participants were recruited through Prolific [2] in September 2020. We received a total of 35 submissions, one of which was discarded due to technical issues in the data collection ($N = 34$). Most participants were male (21 out of 34), and the age range was 18-51 ($\mu = 30$ years-old, $\sigma = 8.3$). Most participants (22) claimed to have interacted with chatbots five or more times before the study, while only four have never interacted with chatbots before. All the participants claimed English as their first language and were located in the USA.

7.2 Results

We analyzed the results in two steps. First, we compare the differences in scores for each construct (APP, CRED, ANTHR, SOC_PR, and UX) to observe differences across treatments (FLG vs. FLG_{mod}). Subsequently, we assessed the validity of the structural model depicted in Figure 7.2, and evaluated the proposed hypotheses to understand the associ-

[2] https://www.prolific.co

ations between the perceived appropriateness of language and other aspects of human-chatbots interactions.

7.2.1 Chatbot comparison: FLG vs. FLG_{mod}

Figure 7.3 shows the scores per construct across treatments. In order to compare whether participants significantly rated one chatbot better than the other, we fitted the CLMM model. The overall model results indicate that there is a difference (LR Chi-Square=80.9, *p*-value< 0.001) when compared to the null model (intercept-only, no fixed effects). The β coefficient for the treatments (FLG vs. FLG_{mod}) is negative (-0.91), meaning that rating in higher categories is more likely for FLG chatbot. The odds ratio is $exp(\beta_{treatment}) = exp(0.91) = 2.5$, which means that FLG chatbot is more than twice as likely to be rated in a category j or above relative to FLG_{mod}.

The interaction effect is also significant (LR Chi-Square=13.08, *p*-value$= 0.01$), which means that the difference observed across treatments does indeed depend on the construct. To evaluate which constructs are influencing the differences between treatments, we performed pairwise comparisons using the Tukey adjustment method to identify the constructs for which the treatments vary. Table 7.1 presents the statistical results for the pairwise comparison.

As Table 7.1 shows, the scores for APP and CRED are significantly different for the two treatments. The positive estimate indicates that FLG chatbot received higher rates than FLG_{mod} chatbot, which aligns with the outcomes of the previous user study. There is also a marginally significant difference for the SOC_PR scores. Confirming the visual inference from Figure 7.3, the difference in the scores for anthropomorphism is not significant. Finally, the variation observed for the user experience construct is not significant as

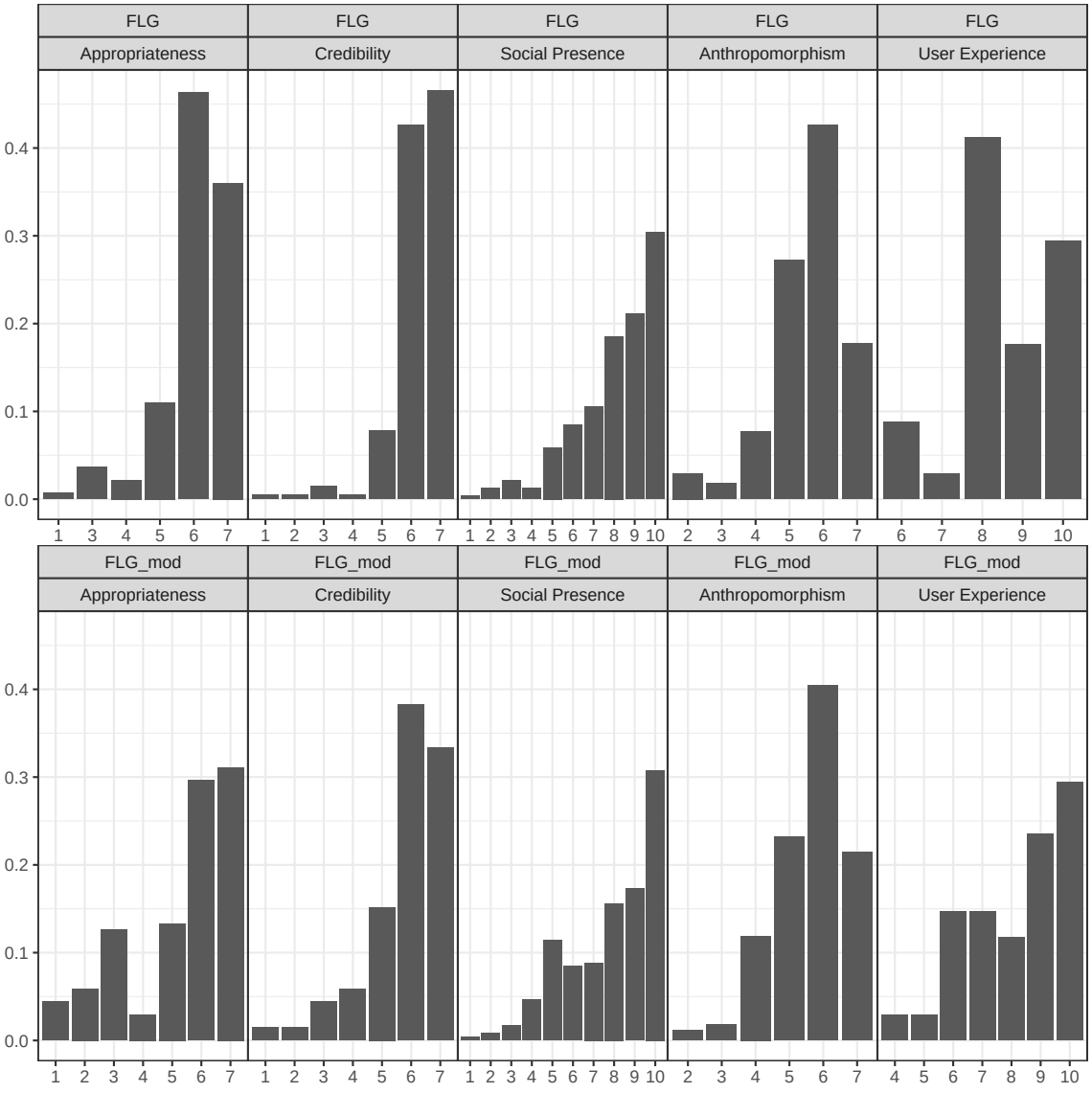


Figure 7.3: Histogram of scores per chatbot per construct.

Contrast (FLG-FLG_{mod})	Estimate	SE	z ratio	p-value
Perc. Approp. Language (APP)	0.91	0.23	4.01	$<.001$
Credibility (CRED)	0.73	0.18	4.13	$<.001$
Social Presence (SOC_PR)	0.48	0.19	2.59	0.01
Anthropomorphism (ANTHR)	-0.08	0.19	-0.45	0.65
User Experience (UX)	0.52	0.45	1.17	0.24

Table 7.1: CLMM results: pairwise comparison of treatments per construct. The differences for perceived appropriateness of language and credibility are supported considering a 5% confidence level.

well. However, the small sample size (N= 34 participants) may be influencing the power to detect differences for this particular construct. The implications of these results are discussed in Section 7.3.

In summary, the results in this section show that the chatbot that uses a register-specific language (FLG chatbot) is more than twice more likely to receive higher scores relative to the chatbot that uses language similar to the patterns of $DailyDialog$ (the FLG_{mod} chatbot). This outcome confirms the inferences from user study 1, where original sentences (FLG) were consistently preferred over modified sentences (FLG_{mod}. The results also support that FLG is significantly perceived as portraying a more appropriate language as well as being more credible and socially present. In the next section, we empirically investigate how these constructs are associated with each other and whether they determine the overall user experience.

7.2.2 Structural model

Table 7.2 shows the values for the Cronbach's α, the composite reliability (CR), and the average variance extracted (AVE). According to Hair *et al.* (2019), Cronbach's α and CR values above 0.70 are considered to indicate high levels of reliability and internal con-

sistency. As for AVE, Hair *et al.* (2019) states that AVE of 0.50 or higher indicates that the construct explains at least 50 per cent of the variance of the construct's items, which is considered acceptable. Therefore, we conclude that the convergent validity and reliability of the constructs are established.

Latent variable	# items	Cronbach's α	CR	AVE
Perc. appropriateness of language	4	0.922	0.923	0.751
Credibility	6	0.907	0.909	0.633
Social presence	7	0.886	0.883	0.533
Anthropomorphism	5	0.843	0.849	0.538
User Experience	1	1	1	1

Table 7.2: PLS-SEM model validity. Cronbach's α and CR values are above .70, and AVE is above .50 for all the constructs, which establishes the validity of the model.

Given the satisfactory model assessment, the PLS-SEM results determine the relationship between the constructs, which provide support to the proposed hypothesis. Figure 7.4 presents the PLS-SEM results, where the numbers inside each latent variable represent the construct's coefficient of determination (R^2), which determines the overall effect size measure for the path models (David Garson, 2016). In general, R^2 values of 0.75 can be considered substantial, while 0.25 can be considered weak (Hair *et al.*, 2019). The numbers on the arrows represent the standardized path coefficients, which determines the possible relationship between the latent variables. Standardized path coefficients range between -1 and +1, and weights closest to absolute 1 reflect the strongest paths (David Garson, 2016).

As Figure 7.4 depicts, the R^2 values suggest that there is a substantial effect for CRED, ANTHR, and UX, and a weak effect for SOC_PR. The path coefficients show that APP has a possibly strong direct effect on CRED and a moderate effect on SOC_PR. However, the path coefficients from APPR to ANTHR and UX shows a very small effect, suggesting that

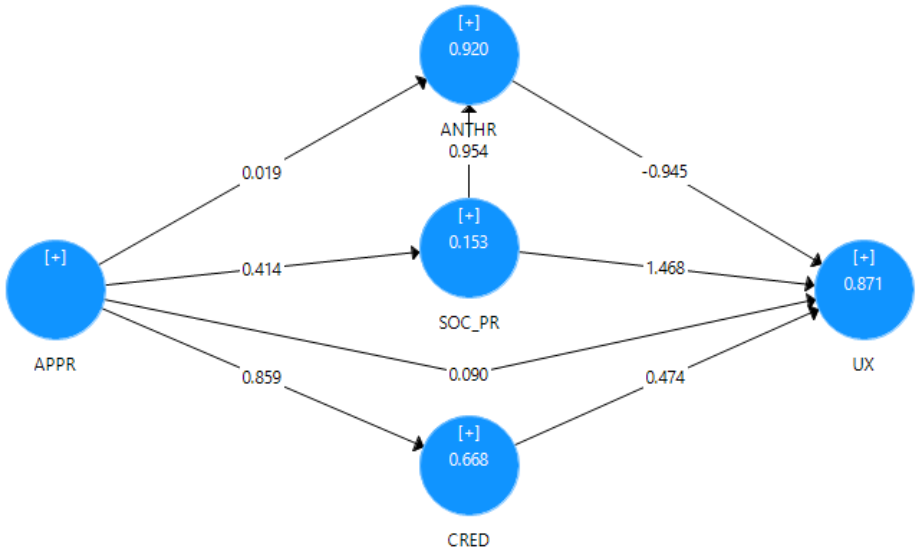


Figure 7.4: PLS-SEM: path coefficients results.

it is possibly not significant. SOC_PR has a strong mediation effect on ANTHR as well. To confirm these inferences, we calculated the PLS bootstrapping with 10,000 samples to assess the significance of the path coefficients and the direct and indirect effects.

The results are shown in Table 7.3. As we can see, the more a user perceives the chatbot's language as appropriate, the more the chatbot is perceived as credible, present, and anthropomorphic, since hypothesis H2, H4, and H6 are supported. However, these constructs do not predict the overall user experience. The implications of these results are discussed in the next section.

	Path	Relationship	Original sample	*T*-statistics	*p*-value	Inference
H1.	APP $->$ UX	direct	0.09	0.01	0.99	Not supported
H2a.	APP $->$ CRED	direct	0.86	15.40	0.00	Supported
H2b.	APP $->$ CRED $->$ UX	mediated	0.41	0.02	0.98	Not supported
H3a.	APP $->$ SOC_PR	direct	0.41	2.95	0.00	Supported
H3b.	APP $->$ SOC_PR $->$ UX	mediated	0.61	0.02	0.98	Not supported
H3c.	APP $->$ SOC_PR $->$ ANTHR	mediated	0.40	2.81	0.01	Supported
H4a.	APP $->$ ANTHR	direct	0.02	0.21	0.83	Not supported
H4b.	APP $->$ ANTHR $->$ UX	mediated	-0.02	0.01	0.99	Not supported

Table 7.3: Hypotheses assessment.

In conclusion, the results presented in this section support the perceived appropriateness of language as a direct predictor of credibility and social presence, as well as its indirect effect on anthropomorphism. It also reveals that the model's paths that lead to the overall user experience were not supported, which may indicate that a more complex model is required to determine how participants perceive their experiences with chatbots.

7.3 Discussion

The outcomes of the across-treatment analysis confirmed the inferences from user study 1: using register-specific language increases the perceptions of language appropriateness and credibility, and there is a positive relationship between these constructs. Therefore, when users perceive the chatbot's language as complying with the expected register, they also perceive the chatbot as more credible, which is desired in many task-oriented human-chatbot interactions. Credibility relates to the ability to convey expertise, believability, and trust (Zumstein and Hundertmark, 2017; Corritore *et al.*, 2005), and the literature has widely shown that these aspects are crucial for successful human-chatbot interactions (Przegalinska *et al.*, 2019). For example, Pillai and Sivathanu (2020) shows that perceived trust affects the adoption intention of chatbots for tourism. Nordheim *et al.* (2019) found that perceived expertise is essential to an user's trust in customer service chatbots. Thus, we can conclude that when credibility is desired, chatbot designers should consider using register-appropriate language.

Additionally, results show that the use of register-specific language positively influences the extent to which a chatbot is perceived as socially present, which leads to higher levels of anthropomorphism. This aligns with other results in the literature. For example, Laban and Araujo (2020) observed a direct association between perceiving a chatbot as

more anthropomorphic and higher levels of social presence and that the perceived chatbot's service performance is mediated by the perceptions of the anthropomorphic cues. Adam *et al.* (2020) demonstrate that anthropomorphism increases the likelihood that users comply with a chatbot's request for service feedback and that social presence mediates the effect of anthropomorphic design cues on user compliance. Also, Schanke *et al.* (2020) found that anthropomorphism plays an important role when sensitive information disclosure is required.

Anthropomorphism and social presence are often considered as measurements of a chatbot's social skills. Some scholars consider the chatbot's patterns of language use as an anthropomorphic clue (Adam *et al.*, 2020; Go and Sundar, 2019; Araujo, 2018; Schanke *et al.*, 2020), usually associating informal, casual language to the human-like conditions. Results confirm the association between language and anthropomorphism, achieved through social presence. More importantly, even with a small sample size, we showed that participants are more likely to attribute higher scores for social presence to the chatbot that uses register-specific language. These results support the claim for a more formal technique to design chatbot's language to comply with the context and the role that it represents.

Finally, we could not find evidence for either direct or mediated association between perceiving the chatbot's language as appropriate and the overall user experience. As we stated in Section 6.1, user experience is a complex construct. Although we highlighted the linguistic aspect throughout the study, we believe that perceptions other than language may have influenced user answers to this item, which would explain the floating score patterns we observed in Figure 7.3. Considering the outcomes from user study 1, if we look back at Table 6.1, appropriateness and credibility are more consistent in terms of the influential linguistics features than the user experience. There are three linguistic features

(*conditional subordination*, *contractions*, and *second-person pronoun*) that influence both appropriateness and credibility, but not the user experience, while there are three other linguistic features that influence user experience only (*suasive verbs*, *causative subordination*, and *third person pronoun*). This distinction points out that the users perceptions of their overall experience vary from their perceptions of other relevant aspects of the interaction, such as the agent's credibility. Credibility, social presence, and anthropomorphism alone did not determine the overall user experience in this study.

Noticeably, the path coefficient between anthropomorphism and user experience is negative, which means that higher levels of anthropomorphism would decrease the scores for user experience. Although this effect is not significant, it points toward a essential aspect of human-chatbots interactions: increased human-likeness results in an increased unpleasant impression about the artificial agent, which the literature calls *uncanny effect* Appel *et al.* (2020); Ciechanowski *et al.* (2018). As Schanke *et al.* (2020) states, increasing anthropomorphism can result in unintended consequences. In that study, the authors found that anthropomorphism is positively associated with customers' willingness to input more personal information, but also resulted in price sensitivity. In the study presented in Chapter 6, we discuss that participants considered linguistic features used by the human tourist assistants inappropriate for a chatbot. For example, participants associated the plural form of first person pronoun ("we") with positive impressions, while the singular form ("I") is associated with more negative impressions. Thus, further research is required to explore whether anthropomorphism is suppressing the effect on overall user experience and, in this case, how one can tailor the register to balance the levels of anthropomorphism conveyed by the chatbot.

Chapter 8

USER PERCEPTIONS: REPLICATION

The study on user preferences, presented in Chapter 6, revealed the association between using register-specific language and user perceptions of the interaction. To evaluate whether this conclusion is supported if we compare FLG to a different corpus, we selected another corpus in the tourism domain, namely $Frames$ (Asri *et al.*, 2017), and replicated the steps *(3) Data collection* through *(6) User study: users preferences* of this research design. This chapter presents the outcomes from this replication.

8.1 Data collection

The $Frames$ (Asri *et al.*, 2017) dataset is a corpus of 1349 human-human, text-based interactions where users asked for assistance to book a trip. The corpus is intended to support research on dialogue flow when decision-making is involved. The conversations were collected in a Wizard of Oz (WoZ) setting, where the human assistant pretended to be a dialogue system. The wizards had access to the travel packages with hotel and round-trip flights, whereas participants were given the constraints to consider the different options and find the best deal. The corpus is available online [1] along with the tasks, descriptive statistics, and other relevant information.

After downloading the $Frames$ corpus from its website, we parsed the conversations to retain only the tourist assistant's utterances. Then, we followed the situational analytical framework Biber and Conrad (2019) to identify the situational parameters in compar-

[1] https://www.microsoft.com/en-us/research/project/frames-dataset/

ison to FLG. The outcome is presented in Table 8.1.

Table 8.1: Situational analysis ($Frames$ vs. FLG).

Situational parameter	Frames	FLG
Participants	Tourist and travel assistant	Tourists and tourist assistants
Relationship	Tourist assistant and tourist, the former owns the knowledge	Tourist assistant and tourist, the former owns the knowledge
Channel	Written, instant messaging tool	Written, instant messaging tool
Production	Quasi-real-time	Quasi-real-time
Setting	Private, shared time, virtually shared place	Private, shared time, virtually shared place
Purpose	Decision-making; book travel packages; user's constraints	Information search
Topic	Travel packages reservation	Local information (e.g., activities, attractions)

Note that the variation in the situational parameters between $Frames$ and FLG is mainly in the purpose and topic parameters; $Frames$ focuses on pre-travel decision-making, while FLG focuses on en-route information search. As a result, we expect the register characteristics to differ across corpora, which we evaluate in the next section.

8.2 Register characterization

We followed the same procedures presented in Chapter 4 to tag and count the linguistic features present in the $Frames$ corpus as well as to calculate the *dimension scores* for each text. We applied a one-way MANOVA to generate a statistical comparison of the dimension scores across corpora ($Frames$ and FLG). The MANOVA revealed that the dimension scores for FLG's tourist assistants are significantly different from the dimension scores for $Frames$ ($Wilks = 0.95, F = 5.71, p < 0.0001$). Given the significant overall MANOVA test, we also performed a one-way univariate analysis ($df = 3, 1489$) for each of the five dimensions to identify the individual dimensions that influence the prevailing register characteristics. Table 8.2 summarizes the univariate analysis per dimension.

Table 8.2: Univariate analysis of dimension scores ($df = 3, 1489$). For each dimension, the table shows the estimated dimension score $\pm$, the standard error per group ($Frames$, $TA1$, $TA2$, $TA3$), and the corresponding *F*- and *p*-values.

	Frames	**TA1**	**TA2**	**TA3**	***F***	***p*-value**
Dim. 1: Involvement	13.18 $\pm$ 0.68	14.73 $\pm$ 3.62	5.69 $\pm$ 3.66	-5.02 $\pm$ 3.48	10.07	$<$0.0001
Dim. 2: Narrative flow	-4.94 $\pm$ 0.04	-4.45 $\pm$ 0.20	-4.31 $\pm$ 0.21	-4.79 $\pm$ 0.20	4.47	0.0040
Dim. 3: Contextual ref.	-1.06 $\pm$ 0.16	-3.33 $\pm$ 0.87	-1.75 $\pm$ 0.88	-2.65 $\pm$ 0.83	3.42	0.0166
Dim. 4: Persuasion	-0.09 $\pm$ 0.15	1.98 $\pm$ 0.81	-0.02 $\pm$ 0.82	1.93 $\pm$ 0.78	4.12	0.0064
Dim. 5: Formality	-0.49 $\pm$ 0.14	-0.81 $\pm$ 0.74	-1.70 $\pm$ 0.75	-2.10 $\pm$ 0.71	2.42	0.0642

The dimension score reveals that the purpose (decision-making vs. information search) likely impacted the narrative flow (dimension 2), and the persuasion (dimension 4), since the tourist assistant in $Frames$ focused on describing the options rather than arguing on what were the best deals. However, as we discussed in Chapter 4, the dimension scores are not sensitive enough to identify how the register characteristics emerge in each corpus; thus, we compared the occurrences of every linguistic feature per dimension. The left side of Figure 8.1 depicts the linguistic features that vary significantly between the two corpora (FLG and $Frames$), as revealed by the ANOVA analysis per feature. The figure shows the estimates for $Frames$ (control group, in blue) and each tourist assistant in FLG (TA1, TA2, and TA3, in red). The horizontal line represents the standard error. A table with the numbers, including the non-significant, linguistic features is presented in Appendix H.

We found that 29 out of 48 linguistic features were significantly different across corpora. The next step was to modify the conversations to produce another parallel corpus, FLG_{mod_2}, which has equivalent content as FLG, but this time mimics the register we identified in the $Frames$ corpus.

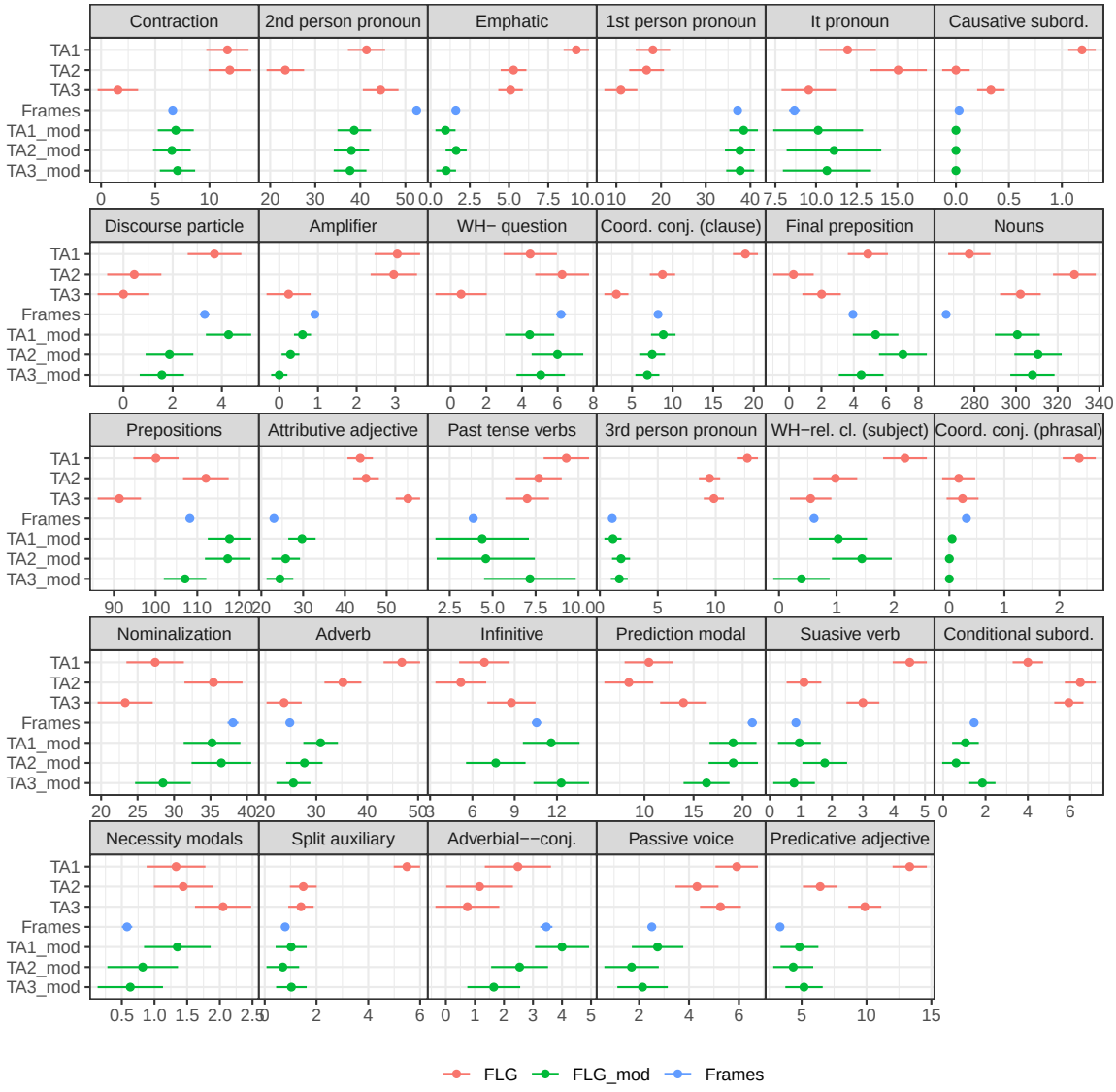


Figure 8.1: Visualization of ANOVA results for individual features comparison between $Frames$ and both original (FLG) and modified (FLG_{mod}) corpora. For each feature, the estimates and standard errors are represented as a dot and a horizontal line, respectively. The colors represent each corpora (red for FLG, blue for $Frames$ and green for FLG_{mod}. All the statistics are calculated with $df = 3, 1489$

8.3 Text modification

Text modification process followed the same procedures as described in Chapter 5. We created a new list of paired inputs and passed it to the Python script, which applied the changes to a copy of the original FLG conversations to create the FLG_{mod_2} corpus. Figure 8.1 shows the comparison between $Frames$ and FLG_{mod_2}. The figure shows the estimates for $Frames$ (control group, in blue) and each tourist assistant in FLG (TA1, TA2, and TA3, in green). The horizontal line represents the standard error. Once again, we could not reach non-significant levels for second person pronouns and nouns, although we reduced the *F*-values substantially. Table 8.3 shows an example of a modified answer. Once more, features that were not statistically significant when compared to FLG were still not significant when compared to FLG_{mod_2}. A table with the estimates, standard errors, *F*-values, and *p*-values for all the analyzed features can be found in Appendix H.

Table 8.3: Example of a modified answer (FLG vs. FLG_{mod_2}). The left side shows the answer provided by a tourist assistant in the original data collection. The right side shows the corresponding answer, modified to portray features that mimics $Frames$ linguistic patterns. Modified words are highlighted in bold and the tags attributed to the words are between square brackets.

Original answer (FLG corpora)	**Modified answer (FLG_{mod_2} corpora)**
There is a self-guided Rte 66 tour that starts in the **Historic Train** *[attributive adjective]* Center on 1 E. Rte. 66. In the visitor's center there is a **self-guided map** *[attributive adjective]* that shows the original alignment through the redeveloped Southside Historic District and passes by **classic drive-in** *[attributive adjective]* motels and **Flagstaff** *[noun]* landmarks of old. Let me know **if** *[conditional subordination]* you have further questions.	**We** *[first person pronoun]* **offer** *[present verb]* a self-guided Rte 66 tour **for you** *[preposition, second person pronoun]* that starts in the Train Center on 1 E. Rte. 66. In **our** *[first person pronoun]* visitor center, a map **has** *[present verb]* the original alignment through the redeveloped Southside Historic District and passes by motels and landmarks of old. Tell me your other questions.

Once the FLG_{mod_2} parallel corpus was developed, we followed the same procedures described in Chapter 6 and the results are presented in the next section.

8.4 User perceptions study

In this section, we present the results of the replicated user perceptions study.

8.4.1 Participants

Participants were recruited through Prolific in September, 2020. We received a total of 174 submissions, 29 of which were discarded due to either technical issues in the data collection or failure to answer the attention checks ($N = 145$). All the participants claimed English as their first language and were located in the USA. Additionally, we configured Prolific to recruit only participants who did not participate in the user study 1 (Chapter 6), to avoid biases in the data collection. Most participants had either a four-year bachelor's degree (49) or some college, but no degree (42). 20 participants graduated from high school, and other 17 had Master's degree. Common educational backgrounds were STEM (34), Arts and Humanities (28), and others (29). Three participants had non-binary gender, 70 declared themselves as female, and 72 as male. The age range is 18-60 ($\mu = 30.77$ years-old, $\sigma = 10.32$). Appendix H contains plots that show the distribution of participants for demographics and profile variables.

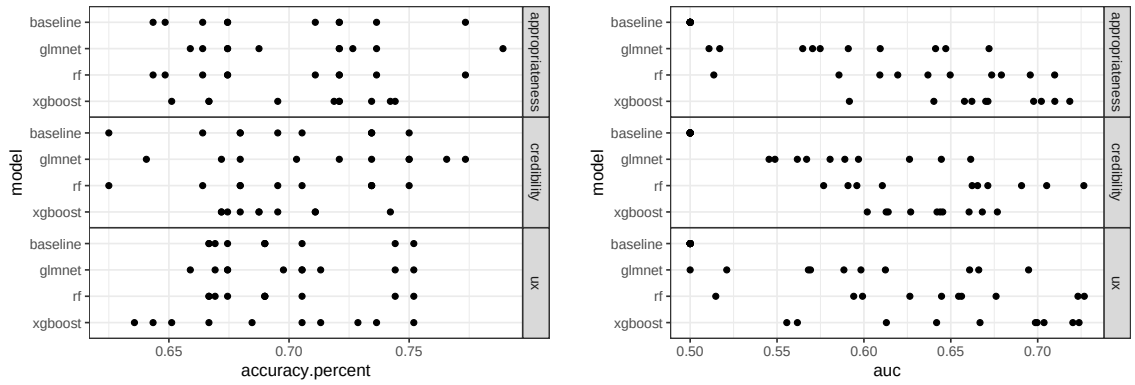
8.4.2 Analysis of the linguistic features

We fitted the generalized linear model (GLM), using the $glmnet$ package in R (Friedman *et al.*, 2010), using the same algorithm presented in Chapter 6. For this round, the evaluation dataset started with a total of 3,915 observations (145 participants, 27 evaluations

per participant). From this total, participants skipped the question without answering for only one observation. In 59 others, the participants chose the "I don't know" option. These observations were discarded from the analysis, resulting in a dataset with 3858 observations. Each question-answer pair was evaluated from 22 to 26 times per construct. As in user study 1, participants overall preferred the answers from the original corpus, although the modified version was preferred for a few answers. Appendix H presents a table that shows the number of times each option (i.e., original, modified, I don't know) was selected per question.

Figure 8.2 shows the prediction accuracy and AUC plots for the four fitted models, which are similar to these measures for user study 1. Since participants generally preferred the original FLG corpus answers, the prediction threshold is close to always predicting the most frequent class (original). The prediction accuracy of `glmnet` and `xgboost` are only slightly better than the baseline. Nevertheless, the AUC values are consistently better than the baseline (the ROC curve plot is available in Appendix H). As in user study 1, the non-linear models are not considerably more accurate than the linear model, which justify the use of $glmnet$ model to identify the linguistic features that determine the users preferences.

Table 8.4 presents the coefficients of the linguistic features selected in six or more folds. The first and second columns indicate, respectively, the linguistic feature of interest and the sign of original $-$ modified calculation, which indicates whether one particular feature was increased or decreased in the text modification process. A positive sign (+) for a feature f_i indicates that $\text{count}_{\text{original}}(f_i) > \text{count}_{\text{modified}}(f_i)$, while a negative sign (-) indicates that $\text{count}_{\text{original}}(f_i) < \text{count}_{\text{modified}}(f_i)$. The following three columns present the mean of the coefficients and the standard deviation for each construct. Features with



(a) Accuracy (b) AUC

Figure 8.2: Accuracy (a) and AUC (b) results per model for each construct (appropriateness, credibility, and user experience). The baseline represents a model that always predicts the most frequent class (original). Accuracy percentage shows that glmnet, random forest, and xgboost perform only slightly better than the baseline model. AUC, however, is reasonably better than the baseline for the three models, which aligns with the results from Chapter 6.

negative coefficients increase the likelihood of the model predicting the original class. In contrast, features with positive coefficients increase the likelihood of the model predicting the modified class. Appendix H include plots of coefficients for each construct.

Original answers have significantly more *split auxiliaries*, *nouns*, and *adverbs* than the modified versions. These features have a negative coefficient for all the three constructs, which indicates that frequent occurrences of these features increase the likelihood of original answers being chosen; participants are more likely to prefer answers in which these features are more frequent. The same conclusion applies to *contractions*, *causative subordinations*, and *predicative adjectives*, although *contractions* show up as relevant for appropriateness and user experience only, *causative subordinations* are relevant only for the appropriateness construct, and *predicative adjective* predicts credibility only.

Original answers also have significantly more *agentless passives* and *third-person pronouns*. These features have a positive coefficient for all three constructs, which indicates

Table 8.4: Coefficients and standard deviation of the non-zero variables per construct. Most of the selected linguistic features are relevant for all the three constructs. The columns original − modified represents whether the original answers have more (+) or less (−) occurrences of the corresponding feature. The dots indicate that the corresponding feature was not selected for that particular construct.

		Mean of coefficients ± Std. Deviation		
Linguistic features	orig. − mod.	Appropriateness	Credibility	User Experience
Split auxiliary	(+)	-0.010 ± 0.003	-0.027 ± 0.004	-0.017 ± 0.007
Adverbs	(+)	-0.009 ± 0.001	-0.002 ± 0.001	-0.005 ± 0.001
Nouns	(+)	-0.001 ± 0.000	-0.003 ± 0.001	-0.001 ± 0.001
Contractions	(+)	-0.003 ± 0.001	·	-0.007 ± 0.002
Causative subordination	(+)	-0.011 ± 0.012	·	·
Predicative adjective	(+)	·	-0.003 ± 0.001	·
Agentless passive	(+)	0.002 ± 0.001	0.006 ± 0.001	0.005 ± 0.003
Third-person pronoun	(+)	0.006 ± 0.001	0.005 ± 0.000	0.005 ± 0.001
"It" pronoun	(+)	0.004 ± 0.002	0.003 ± 0.001	·
Conditional subordination	(+)	·	0.002 ± 0.002	0.004 ± 0.002
Attributive adjective	(+)	·	0.006 ± 0.001	0.001 ± 0.001
Emphatic	(+)	·	·	0.007 ± 0.004
Preposition	(-)	-0.003 ± 0.000	-0.001 ± 0.001	-0.002 ± 0.001
Infinitive	(-)	-0.002 ± 0.001	-0.012 ± 0.001	·
Nominalization	(-)	·	-0.003 ± 0.001	-0.003 ± 0.002
Prediction modal	(-)	0.008 ± 0.002	0.008 ± 0.002	0.014 ± 0.002
Adverbial–conjuncts	(-)	·	0.004 ± 0.004	·

that frequent occurrences of these features increase the likelihood of modified answers being chosen. This outcome suggests that participants are more likely to prefer answers in which these features are less frequent. The same conclusion applies to *"it" pronouns*, *conditional subordinations*, *attributive adjectives*, and *emphatics*, but these features are not relevant for all constructs. *"It" pronouns* did not show up as relevant for user experience, *conditional subordinations* and *attributive adjectives* did not influence appropriateness, and *emphatics* show up as relevant only for user experience.

Modifications have significantly more *prepositions* than the original answers. This feature has a negative coefficient for all three constructs, which indicates that frequent occurrences of *prepositions* increase the likelihood of original answers being chosen. This out-

come suggests that participants are more likely to prefer answers in which these features are less frequent. The same inference applies to *infinitives* and *nominalizations*, although *infinitives* did not show up as a relevant feature for user experience, and *nominalizations* did not predict the appropriateness.

Modifications also have a larger number of *Prediction modals*. This feature has a positive coefficient for all three constructs, which indicates that increasing the occurrences of *prediction modals* increased the likelihood of modified answers being chosen. This outcome highlights the participant's preferences for answers in which these features are more consistently present. The same conclusion applies to *adverbial–conjuncts*, although this feature shows up as relevant only for credibility. Noticeably, when we aggregate the estimate to the standard deviation for this particular feature, it sums up to zero, suggesting that this outcome may be noise.

In summary, the outcomes confirm the association between the use of register-specific language and the user perceptions of appropriateness, credibility, and user experience. In accordance with the user study 1's outcomes, linguistic features are stronger predictors than variables that indicate individual characteristics of either participants or assistants, suggesting that adopting the expected register influence positively the user perceptions of the interaction.

8.4.3 Discussion

In this section, we discuss the interpretation of the linguistic features that were selected in this study and compare the results to the observations from user study 1.

The replication with a completely new modified corpus supported the insights from user study 1. Firstly, it confirmed the importance of using register-specific language as a

means of improving the human-chatbot interactions. The study strengthens the conclusion that the chatbot's language–characterized in this research under the lens of conversational register–can impact the user perceptions and behaviors toward the agent, which, ultimately, may result in increased quality, acceptance, and adoption Jakic *et al.* (2017); Morrissey and Kirakowski (2013); Gnewuch *et al.* (2017).

Secondly, the analysis of individual linguistic features supports the previous inferences for at least four linguistic features, namely *preposition*, *causative subordination*, *third-person pronoun*, and *conditional subordination*. In user study 1, *prepositions* showed inconsistent results across constructs, indicating that participants preferred answers with less frequent occurrences of this feature for appropriateness and credibility, and more frequent occurrences for user experience. We suggested that the effect on user experience could be noise, since the aggregation of coefficient means and standard deviation sum up to zero. In this new study, *preposition* was selected once again for all the constructs, and participants indeed preferred answers in which this feature is less frequent. As for *conditional subordination*, this feature is negatively associated with credibility in both studies, which supports the inference that when the chatbot gives options, it sounds as it is not confident about the information provided. *Causative subordination* (e.g., *because*) was selected as a predictor of user experience in the previous study and as a predictor of appropriateness in the current one. In both cases, users are more likely to prefer answers in which this feature occurs. Noticeably, *causative subordination* is a fairly uncommon feature (mean of 1 per 1,000 words Biber (1988)) when compared to others such as *nouns* (mean of 180 per 1,000 words Biber (1988)). The user preferences for causative subordination indicate that, when this feature occurs, it is preferred over receiving the corresponding information without the subordination, likely due to the natural connection between related sentences.

Participants preferred less frequent occurrences of *third-person pronouns* in both studies. In the first study, this effect was significant for user experience only, but all three constructs were influenced by this feature in the current study. *Third-person pronouns* are used in FLG to refer to business or attractions, usually to add details about them (e.g., "*they* have public restrooms"). As we discussed in Chapter 6, efficiency is crucial in information search scenarios, especially in the context of en-route travel information search, which makes detailed descriptions unnecessary in many cases. Additionally, by using *third-person pronouns*, the assistant provides the information from an external, narrative standpoint (e.g., "Tonight **they** *[the museum community]* have musical performances" vs. "Tonight **we** *[the assistant as part of the museum community]* have musical performances"), which results in a more impersonal tone Biber *et al.* (1999). Hence, these results suggest that reinforcing the tourist assistant chatbot as a representative of its social category by using (plural) first-person pronouns would be preferable over the impersonal third-person pronoun.

On the other hand, this study shows conflicting outcomes regarding the levels of *prediction modals* and *contractions*. In the user study 1, participants preferred lower levels of these features, whereas the current study indicates a preference for higher levels of these features in all the three constructs. This dissonance can be explained by the differences in how $Frames$ and $DailyDialog$ use prediction modals. In $DailyDialog$, *would* is the most common *prediction modal*, which mostly co-occurred with first-person pronouns (i.e., *"I would"* and the contracted form *"I'd"*). As we discussed in Chapters 6, the co-occurrences with first-person pronouns may have influenced the outcomes for prediction modals and contractions in that study, as these co-occurrences imply an excessive personification, which may cause uncanny effects on the user perceptions. In contrast, $Frames$

has more frequent occurrences of other forms of prediction modals, such as *shall* and *will*. *Will* is the most frequent modal and it often co-occurs with *nouns* (e.g., *"campgrounds will open"* or *"downtown will have vegetarian options"*) instead of personal pronouns. As a consequence, the negative effect was flattened and the frequent occurrences of *prediction modals* and *contractions* resulted in a positive effect.

Given the register differences between $Frames$ and $DailyDialog$, the modified corpora (FLG_{mod} and FLG_{mod_2}) present a varying set of manipulated features. Thus, this study allowed us to observe inferences about features that were not manipulated in the previous study. *Split auxiliaries* are likely influenced by the preferences for frequent occurrences of *adverbs*, as split auxiliaries occur when adverbs are placed between auxiliaries and their main verb Biber (1988) (e.g., *"will* ***obviously*** *limit"*). Many adverbs are used in FLG to indicate the tourist assistant's stance (e.g., *"Absolutely!"* and *"there are definitely some,"* which indicate assurance, or *"you'd probably be fine,"* which indicates uncertainty). The presence of these features emphasizes the level of confidence that the assistant has about the information, resulting in a positive effect on the user perceptions of all the evaluated constructs. The same conclusions apply to *predicative adjectives*, which is also frequently used for marking stance Biber (1988). This feature was selected as a predictor of the tourist assistant's credibility, which clearly relates to the ability to express opinion (e.g., *"That would* ***be difficult****." "their sandwiches* ***are delicious****," "the restaurant* ***is good****"*).

In contrast, results show that participants are more likely to prefer answers in which the occurrences of *agentless passives*, *infinitives*, *nominalizations* are less frequent. These features are, in general, uncommon in conversations Biber *et al.* (1999); Biber (1988), which may justify the user's negative impressions about higher frequencies of these features. *Emphatics* and *"it" pronouns* are rather frequent in conversations. However, *emphatics*

are characteristic of informal, colloquial discourse, whereas “it” pronouns are generalized pronouns that indicate limited informational content Biber (1988). The tourist assistant chatbots evaluated in this research are representatives of a professional category that specialize in providing information, which may increase the user’s expectations for a more formal and specialized discourse.

Chapter 9

CONCLUSIONS

The adoption of chatbots in a variety of domains is continuously rising, heading to a global market that is expected to reach over one billion US dollars in a few years. In order to convey competence and be recognized by human-interlocutors for the role they stand in, chatbots must cohere with the social role they aim to represent, particularly in information search contexts where users are sensitive to the high negative consequences of incorrect advice. To this date, there is no formal techniques to guide the design of a chatbot's linguistic choices, which is often based on the designer personal linguistic habits or an ad-hoc analyses of user characteristics. In this research, we analyze the effect of linguistic choices on user perceptions, and take first steps toward establishing register theory as a theoretical foundation for tailoring chatbot linguistic choices to a particular interactional situations.

Register is an established theory in the sociolinguistics domain (see e.g., (Biber, 1988; Argamon, 2019; Biber, 2012; Biber *et al.*, 1999)), and has been shown as a reliable predictor of language variation across conversational contexts. We investigated the applicability of register theory to human-chatbot interactions, developed a rationale for accounting for register in chatbot design, and provided a concrete mechanism for implementing theory into design practice. In this chapter, we summarize the conclusions by bringing all of the results developed in previous chapters to bear on the main research question. Then, we discuss the implications for design of future chatbot conversational engines, as well as limitations of this research, which in turn lead to a discussion of directions for future

work.

9.1 How does a chatbot's use of register-appropriate language affect the user perceptions and experiences with chatbots?

The user studies clearly show that register characteristics influence user perceptions of appropriateness and credibility. The results presented in this research demonstrate that: (i) users perceive the chatbot's language as more appropriate when they use register-specific language; (ii) register characteristics are more relevant than individual preferences or personal habits; and (iii) there is an association between user perceptions of a chatbot's appropriateness of language and perceived credibility.

User perceptions of conversational skills are important in chatbot design because chatbots are targeted to fluidly interact using natural language. Chatbots are often deployed to perform social roles traditionally associated with humans, particularly in contexts where there may be consequences for a human if they choose to act on the chatbot's information. This means that user perceptions of chatbot competence and credibility are crucial for a chatbot's success. Some previous studies have found that appropriate language style is not relevant for determining user satisfaction as long as the user can understand the chatbot's answer, only advising that the chatbot's language style should be "mildly appropriate to the service the chatbot provides" (Balaji, 2019). Chapters 6 and 8 suggested, however, that the narrow focus on traditional usability metrics, such as effectiveness and efficiency, fail to adequately capture the broader context of user experience, which goes beyond merely comprehending a chatbot's utterance to, more importantly, whether the user trusts and ultimately uses the chatbot's advice. Specifically, user experience also involves the "user perceptions and responses that result from the use of the system" (ISO

9241-11, 2018), including emotions, beliefs, preferences, and perceptions, among others. Results suggest that user perceptions are also shaped by *how* information is conveyed– as characterized by the conversational register–rather than by the explicit information content of conversational exchanges.

User study 2 reinforces the complexity of the user experience concept. The outcomes did not support the association of perceived appropriateness of language and user experience, neither directly nor indirectly. However, it brought to light a crucial aspect of human-chatbot interaction, namely the need for balancing the chatbot's anthropomorphic clues. The study confirmed that using register-appropriate language increases the user perceptions of social presence, which leads to higher levels of anthropomorphism. However, the literature has widely demonstrated that overly humanized agents provoke uncanny feelings (Chaves and Gerosa, 2020; Appel *et al.*, 2020; Ciechanowski *et al.*, 2018), and create a higher expectation regarding performance and quality on the part of users, which eventually leads to more frustration when the chatbot fails to live up to these increased expectations (Gnewuch *et al.*, 2017). Since the language in the baseline corpus (FLG) was human-written and not tailored to represent the identity of a chatbot, participants likely perceived some of the chatbot's answers as overly human-like, which potentially justifies the negative effect of anthropomorphism and user experience. As register theory suggests (Biber and Conrad, 2019), the interlocutors' identities and the relationship among them are influential parameters when defining the interactional context. Therefore, tailoring the chatbot's language to the appropriate register includes not only adapting to the language of the professional category it represents, but also revealing the chatbot's social identity as an artificial agent. These observations strengthen the relevance of conversational register as a theoretical foundation for the design of chatbot

utterances.

In conclusion, the results support earlier behavioral observations that found that language fit can impact the user perceptions of the interactions as well as their behaviors toward chatbots (Jakic *et al.*, 2017). Importantly, the work presented here refines these observations by demonstrating that register theory can provide a sound theoretical framework for concretely characterizing the concept of "conversational style" and analytically exploring how variations in the patterns of language impact user perceptions of chatbot quality, including critical factors, like appropriateness and credibility.

9.2 Implications for chatbot design

Results demonstrate how register analysis can be used as an effective technique for exposing the set of linguistic features that have the greatest impact on user perceptions of the conversation. This obviously has important implications for designers of the next generation of chatbots. For chatbots that find and retrieve knowledge snippets from external sources, utterances should be adapted to the conversational situation in which the chatbot is embedded. This is not generally done in the current generation of chatbots; it is common to find chatbots that extract and present information directly from websites, books, or manuals without any adaptation. For example, Golem [1] is a chatbot designed to guide tourists through Prague (Czech Republic); its utterances are extracted from an online travel magazine [2] without any adaptation to the new interactional situation (which differs in production, channel, and setting). Moreover, new generations of chatbots will be expected to generate their own custom-constructed utterances dynamically, which will

[1] Available in Facebook Messenger at http://m.me/praguevisitor. Last accessed: June, 2020

[2] https://www.praguevisitor.eu

require sophisticated natural language engines that are able to adapt dynamically their conversational register to changing situational parameters. In this context, corpora such as $DailyDialog$ or $Frames$ are likely to become a baseline for training the conversation models (Galitsky *et al.*, 2019) at the heart of such natural language engines; this study emphasizes that designers should carefully ensure that the register found in any corpus used to train such models matches the optimal register implied by the situational parameters, or that the learning algorithms can adapt the language accordingly.

This research indicates a list of linguistic features that conform with user expectations about a chatbot's language use in the context of tourism information searches, which can be directly applied to the design of chatbots for this domain. Researchers could leverage these outcomes to evaluate the application of these results to similar interactional situations in other domains. Additionally, the methodology presented in this research can be applied to other domains that are more distant in terms of interactional situations from FLG, aiming to identify the associations between new interactional situations and core linguistic features used in the domain. Additional effort is required to develop tools to automate these methodological steps and dynamically generate register-specific language for chatbot design.

Register theory aims to link the occurrences of certain linguistic features in utterances to the situational parameters of the conversation Biber and Conrad (2019). Although characterizations of situational parameters and their detailed impacts on the selection of conversational register may continue to evolve and be refined over time, the patterns of language should be similar in domains that share similar situational parameters. For instance, information search, the core interactional purpose of the interactions studied in this research, is also a common interactional purpose in customer service interactions, i.e.,

two participants working to fulfill an information request. The two domains share other situational parameters as well, such as channel, production, and setting. Abu Shawar and Atwell (2004) observed that conversations from a corpus of Spoken Professional American English portray more *coordinating conjunctions*, *subordinating conjunctions*, and *plural personal pronouns* than transcripts from the chatbot ALICE. These same linguistic features were selected in the analysis as predictors of appropriateness, credibility, and user experience (see Chapter 6). In contrast, we expect that a sales chatbot would require more persuasion (features in Dimension 4, see Appendix C for the full list of linguistic features), and recommendation-based chatbots would require more personalization. In any case, a register analysis, as presented in this research, could be used as a tool to analyze the conversation register used by expert humans in such conversational scenarios, and to identify specific linguistic features of that register that are relevant to producing the desired impact on user perceptions.

Using register analysis to characterize different situations and how they map to the most appropriate register profiles provides a way forward in the design of the next generation of chatbots; one could imagine a chatbot language engine that, given a particular situational profile for planned conversations, could automatically configure its language to present information in the most appropriate register. More generally, this research demonstrates that the theoretical foundation of register analysis can be an effective tool for characterizing the conversational register used in other target domains, and can systematically expose the specific linguistic features within conversational utterances that most strongly impact user perceptions and experiences. Finally, the parallel corpora FLG, FLG_{mod}, and FLG_{mod_2} are available for researchers and practitioners who are interested in (i) developing chatbots for tourism information searches, as they comprise a set of fre-

quently asked tourism questions with register-specific answers; or (ii) research on natural language generation that requires parallel data. The corpora and associated materials are available online (Chaves, 2020a).

9.3 Limitations

The register characterization relies on the multidimensional approach proposed by Biber (1988, 1995), which is the main theoretically-motivated approach taken within register analysis (Argamon, 2019). Other approaches, such as register classification (Argamon, 2019), could be explored in the context of human-chatbot interactions. Additionally, the register analysis also relies on Biber's grammatical tagger (Biber, 2017) to automatically tag the linguistic features. The tagger has been used for many previous large-scale corpus investigations, including multidimensional studies of register variation (e.g., (Biber, 1995, 1988; Conrad and Biber, 2014)), The Longman Grammar of Spoken and Written English (Biber *et al.*, 1999), and a study of register variation on the Internet (Biber and Egbert, 2016, 2018). Although this tagger achieves accuracy levels comparable to other existing taggers (Biber, 1988), mis-taggings are possible. To mitigate this effect, we manually inspected a subset of tagged files to search for mis-tags that could potentially impact the outcomes and fixed them across the texts in the corpus.

We performed manual linguistic modifications to produce the FLG_{mod} and FLG_{mod_2} corpora, which inherently introduced a subjective element in the changes we applied to shift the register. We mitigated this threat by manually inspecting $DailyDialog$ and $Frames$ corpora for every feature we modified, to understand the function of the feature in the corpora, and produce modifications using similar patterns. For the FLG_mod, we also performed a validation with human participants for content preservation and quality

of modifications (see Section 5.2).

We included in the cross-validation model only features that vary between FLG and the comparison corpus (either $DailyDialog$ or $Frames$) and, therefore, were manipulated during the text modification. We considered that the linguistic features that do not significantly vary across the corpora are the standard in those particular contexts and are unlikely to signal user preferences. We claim that $DailyDialog$ and $Frames$ are appropriate datasets to be used in this study since they have been widely used in natural language and dialogue generation research (see, e.g., (Zhao *et al.*, 2018a; Shen *et al.*, 2018; Gu *et al.*, 2018; Tran *et al.*, 2018; Mensio *et al.*, 2018)), and might eventually become a baseline for learning conversation models (Galitsky *et al.*, 2019). However, we did not manipulate every possible feature, as we focused on the features that vary in the comparison corpora. It is important to compare FLG against corpora produced in other interactional situations to evaluate varying sets of features until reaching saturation.

The register analysis presented in this research was based on counts of the occurrences of features, normalized per 1,000 words. However, it does not consider sentence structure, i.e., where the features occur in the sentence. Additionally, because linguistic features are best understood in terms of co-occurrence patterns, it is important to extend this study to consider the effect of the linguistic features individually and the effect of features that typically co-occur with them.

This research is performed in the context of conversations in American English. The core linguistic features and their usage change from one language to the other; thus, these results may not apply to other languages and further investigation is necessary.

Three tourist assistants, all female, answered tourist questions in the FLG corpus, and tourists were recruited in Flagstaff, AZ, USA. To increase the diversity of tourists' ques-

tions, we mined questions from websites, as discussed in Chapter 3. Concerning tourist assistants, an interesting extension of this study would include hiring tourist assistants with a more diverse profile and including questions about other touristic cities to reduce the influence of stylistic choices and regionalisms in the register characterization. It is important to note that, even with only three tourist assistants, we were able to identify the impact of register on the evaluated metrics; this suggests that a very large corpus is not necessary to identify the core linguistic features of chatbot dialogues that influence user perceptions.

We collected limited qualitative observations from the in-lab sessions to support the user study 1's quantitative findings (see Chapter 6). Therefore, the ability to draw conclusions based on participant statements is incomplete. The purpose of adding the qualitative element was to augment and clarify the quantitative findings by identifying participant impressions aligned with the quantitative analysis and comparing the impressions of participants to the interpretations of linguistic features we find in the register literature, such as in the Longman Grammar (Biber *et al.*, 1999). A deeper qualitative investigation is needed to draw stronger conclusions about motivations behind a participant's choices and to understand the user's impressions about a chatbot's language use in more natural interactions, such as the scenario presented in the user study 2.

The small sample size used to evaluate the user study 2's hypothesis may have resulted in reduced power to detect associations between the perceived appropriateness of language and user experience. Additionally, we measured the overall user experience with a single item (i.e., "On a scale of 1 to 10, please rate your overall experience with the Tourist Assistant Chatbot #A/B"). This item might be too general to capture all the nuances of what the user can understand by their experience. Defining a model that specifies the

most relevant aspects of user experiences with chatbots is still an open challenge.

9.4 Future work

The work presented here is the initial step toward formalizing register as a theoretical basis for chatbot's language generation. Future developments are needed to developing chatbots that dynamically adapt their language to conform with the expected register. In the following, we point to some important open challenges that can be addressed in future work:

Text modification automation: manual text modification is the most effort intensive step in the presented methodology. Research on style transfer has made progress on reproducing the patterns of language in a new corpus (see e.g., (Zhao *et al.*, 2018b; Syed, 2020; Tikhonov and Yamshchikov, 2018); however, most algorithms assume similar content distributions. To make it possible to automate the text modifications as proposed in this research, we would need to consider the frequencies of words in a given corpus and reproduce these patterns without assuming similar content distribution. This is, until this point, an open challenge.

Identification of relevant linguistic features: in the current research, we selected two corpora that represent different registers ($DailyDialog$ and $Frames$) and evaluated only the features that were significantly different from FLG. As a consequence, some linguistic features were not modified in either user studies 1 or 2, such as hedges and perfect aspects. A large scale study that includes several different corpora with a larger variety of linguistic features could increase the power to measure the strength of particular linguistic features in a given register.

Extensibility of register characteristics to similar domains: we suggested that the tourism information search is similar to other domains in terms of situational characteristics, such as information search in customer service. One interesting extension of this research would include performing situational analysis of several domains where chatbots are commonly used and find the intersections in the situation parameters. Then, we can empirically evaluate the extent to which the generalization to similar domains would apply.

Influence of sentence structure: as we discussed in the previous section, we analyzed the influence of linguistic features individually without considering the position where they occur in the sentence or the frequent co-occurrences with other features. We are not aware of a methodology that automatically weighs the features according to their place in the sentence. Developing such methodologies would potentially improve the ability to detect the factors that influence the user perceptions, which likely include sentence structure and flow.

Ranking sentences according to the register: to allow machine generation of register-specific language, it is crucial that we develop statistical models that, given a set of candidate sentences for a particular situation, rank the sentences according to their coherence with the expected register. This model would select the sentence that has the highest frequency of preferred features and lowest frequency of unfavored features as the top rank position. Such a challenge is still unexplored in the current literature.

Corpus selection automation: the research results show the importance of the reference corpus. Even considering corpora within the tourism domain, differences in

the interlocutor's social role and communication purpose resulted in varying patterns of language. As a consequence, it is important to create a tool that allows practitioners to find the appropriate reference corpus when designing chatbot utterances. We suggest a search platform in which researchers and practitioners could publish corpora of conversations that could be used for particular situations as well as search for a reference corpus for designing their own chatbot. The search would consider the situational parameters in which the conversation takes place to suggest a corpus for training that complies with the expected register. Such platform would be relevant not only for the chatbot industry but also for researchers in several domains such as Natural Language Processing and Generation, Corpus Linguistics, and Computational Linguistics.

Considering alternative forms of conversational agents: this research focused on chatbots that interact through text-based environments. This research could be extended to consider the numerous multi-mode conversational technologies, such as the popular personal/home assistants. Voice-based chatbots present attributes such as tone, accent, and prosody that must be taken into account in addition to the word choices. Future research could consider the influence of these attributes in the user experiences with conversational technologies. Additionally, future research could explore the effect of register for conversational agents that interact in environments other than chat tools, for example, Twitter or GitHub. Currently, there are thousands of social bots that produce content on Twitter or interact with contributors and maintainers on GitHub Wessel *et al.* (2018). We believe that the language used by these agents may also influence the success of their interactions with humans, and the methodology presented in this dissertation may be a starting point for im-

proving the acceptance and credibility of these conversational agents.

Fairness of language: this research suggests that a chatbot's language can be designed by training the chatbot using a corpus of register-appropriate conversations. Corpus-based language generation must, however, consider the ethical concerns, such as the fairness of language. Scholars Schlesinger *et al.* (2018); Marino (2014, 2006) have pointed out the risks of using biased contents of databases to generate language. Future research could extend the methodology presented in this dissertation to include the evaluation of the fairness of the produced language, in order to avoid negative implications of using biased language.

Accomplishing educational goals: Although highly inter-related, Computer Science and Applied Linguistics programs are very often disconnected. In future steps, we want to foster the adoption of strategies to strengthen the connection between students and faculties from the two areas beyond the scope of our research groups. These strategies include devising an open challenge comprising interdisciplinary tasks, which would involve both linguistic and computational expertise, ultimately fostering collaboration among students from these areas. Moreover, this challenge could be based on our annotated corpus, which can bring to the project innovative ideas and usages for the data. We also want to develop educational materials about analyzing register for chatbot design that can be used in undergraduate courses as well as in workshops and conferences in both Linguistics and Computer Science domains.

In conclusion, we expect that the results presented in this dissertation open new research avenues that bring together researchers in computer science, machine learning,

and linguistics to design chatbots that use appropriate language and, consequently, provides an enriched user experience.

Abu Shawar, B. and E. Atwell, "Evaluation of chatbot information system", in "Proceedings of the Eighth Maghrebian Conference on Software Engineering and Artificial Intelligence", (2004).

Adam, M., M. Wessel and A. Benlian, "Ai-based chatbots in customer service and their effects on user compliance", Electronic Markets pp. 1–19 (2020).

Alexis, P., "R-tourism: Introducing the potential impact of robotics and service automation in tourism", Ovidius University Annals, Series Economic Sciences **17**, 1, 211–216 (2017).

Anthony, L., "Antconc: design and development of a freeware corpus analysis toolkit for the technical writing classroom", in "IPCC 2005. Proceedings. International Professional Communication Conference, 2005.", pp. 729–737 (IEEE, Limerick, Ireland, 2005).

Appel, M., D. Izydorczyk, S. Weber, M. Mara and T. Lischetzke, "The uncanny of mind in a machine: Humanoid robots as tools, agents, and experiencers", Computers in Human Behavior **102**, 274–286 (2020).

Araujo, T., "Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions", Computers in Human Behavior **85**, 183–189 (2018).

Argamon, S., "Register in computational language research", Register Studies **1**, 1, 100–135 (2019).

Argamon, S., M. Koppel and G. Avneri, "Routing documents according to style", in "First International workshop on innovative information systems", pp. 85–92 (Citeseer, Pisa, Italy, 1998).

Ashktorab, Z., M. Jain, Q. V. Liao and J. D. Weisz, "Resilient chatbots: Repair strategy preferences for conversational breakdowns", in "Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)", pp. 254:1–254:12 (ACM, New York, NY, USA, 2019).

Asri, L. E., H. Schulz, S. Sharma, J. Zumer, J. Harris, E. Fine, R. Mehrotra and K. Suleman, "Frames: A corpus for adding memory to goal-oriented dialogue systems", in "Proceedings of the SIGDIAL 2017 Conference", pp. 207–219 (Association for Computational Linguistics, Saarbrucken, Germany, 2017).

Avula, S., G. Chadwick, J. Arguello and R. Capra, "Searchbots: User engagement with chatbots during collaborative search", in "Proceedings of the 2018 Conference on Human Information Interaction & Retrieval", pp. 52–61 (ACM, New York, NY, USA, 2018).

Ayedoun, E., Y. Hayashi and K. Seta, "Communication strategies and affective backchannels for conversational agents to enhance learners' willingness to communicate in a second language", in "Artificial Intelligence in Education", edited by E. André, R. Baker, X. Hu, M. M. T. Rodrigo and B. du Boulay, pp. 459–462 (Springer International Publishing, Cham, 2017).

Bakhtin, M. M., *Speech genres and other late essays* (University of Texas Press, Austin, TX, 2010).

Balaji, D., *Assessing user satisfaction with information chatbots: a preliminary investigation*, Master's thesis, University of Twente (2019).

Banks, J., "Perceived moral agency scale: Development and validation of a metric for humans and social machines", Computers in Human Behavior **90**, 363–371 (2018).

Baron, N. S., "Computer mediated communication as a force in language change", Visible language **18**, 2, 118 (1984).

Bates, D., M. Mächler, B. Bolker and S. Walker, "Fitting linear mixed-effects models using lme4", Journal of Statistical Software **67**, 1, 1–48 (2015).

Bi, W., J. Gao, X. Liu and S. Shi, "Fine-grained sentence functions for short-text conversation", in "Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics", pp. 3984–3993 (Association for Computational Linguistics, Florence, Italy, 2019).

Biber, D., *Variation across speech and writing* (Cambridge University Press, Cambridge, UK, 1988).

Biber, D., *Dimensions of register variation: A cross-linguistic comparison* (Cambridge University Press, New York, NY, USA, 1995).

Biber, D., "Register as a predictor of linguistic variation", Corpus Linguistics and Linguistic Theory **8**, 1, 9–37 (2012).

Biber, D., "Mat–multidimensional analysis tagger. available at: https://goo.gl/u7h9gb", (2017).

Biber, D. and S. Conrad, *Register, genre, and style*, Cambridge Textbooks in Linguistics (Cambridge University Press, Cambridge, 2019), 2 edn.

Biber, D., S. Conrad and V. Cortes, "If you look at…: Lexical bundles in university teaching and textbooks", Applied linguistics **25**, 3, 371–405 (2004).

Biber, D. and J. Egbert, "Using multi-dimensional analysis to study register variation on the searchable web", Corpus Linguistics Research **2**, 1–23 (2016).

Biber, D. and J. Egbert, *Register variation online* (Cambridge University Academic press, Cambridge, 2018).

Biber, D. and B. Gray, "Challenging stereotypes about academic writing: Complexity, elaboration, explicitness", Journal of English for Academic Purposes **9**, 1, 2–20 (2010).

Biber, D., B. Gray and K. Poonpon, "Should we use characteristics of conversation to measure grammatical complexity in l2 writing development?", Tesol Quarterly **45**, 1, 5–35 (2011).

Biber, D., S. Johansson, G. Leech, S. Conrad, E. Finegan and R. Quirk, *Longman grammar of spoken and written English*, vol. 2 (Pearson Longman, London, UK, 1999).

Biswas, M., *IBM Watson Chatbots*, pp. 101–137 (Apress, Berkeley, CA, 2018), URL https://doi.org/10.1007/978-1-4842-3754-0$_$4.

Björkqvist, K., K. Österman and A. Kaukiainen, "Social intelligence- empathy= aggression?", Aggression and violent behavior **5**, 2, 191–200 (2000).

Boiteux, M., "Messenger at f8 2018. retrieved october 18, 2019 from https://blog.messengerdevelopers.com/messenger-at-f8-2018-44010dc9d2ea", Messenger Developer Blog (2019).

Bosher, S. and M. Bowles, "The effects of linguistic modification on esl students' comprehension of nursing course test items-a collaborative process is used to modify multiple-choice questions for comprehensibility without damaging the integrity of the item.", Nursing Education Perspectives **29**, 4, 174 (2008).

Brahnam, S. and A. De Angeli, "Gender affordances of conversational agents", Interacting with Computers **24**, 3, 139–153 (2012).

Brandtzaeg, P. B. and A. Følstad, "Why people use chatbots", in "4th International Conference on Internet Science", pp. 377–392 (Springer International Publishing, Cham, 2017).

Brandtzaeg, P. B. and A. Følstad, "Chatbots: changing user needs and motivations", Interactions **25**, 5, 38–43 (2018).

Buhalis, D. and S. H. Jun, "E-tourism", in "Contemporary tourism reviews", edited by C. Cooper, pp. 1–38 (Goodfellow Publishers Limited, Woodeaton, Oxford, 2011).

Candello, H., C. Pinhanez and F. Figueiredo, "Typefaces and the perception of humanness in natural language chatbots", in "Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems", pp. 3476–3487 (ACM, New York, NY, USA, 2017).

Cassell, J., "Social practice: Becoming enculturated in human-computer interaction", in "International Conference on Universal Access in Human-Computer Interaction", pp. 303–313 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2009).

Cerezo, J., J. Kubelka, R. Robbes and A. Bergel, "Building an expert recommender chatbot", in "2019 IEEE/ACM 1st International Workshop on Bots in Software Engineering (BotSE)", pp. 59–63 (IEEE, New York, 2019).

Chaves, A. P., "Github repository. available at: https://github.com/chavesana/chatbots-register", (2020a).

Chaves, A. P., "Should my chatbot be register-specific? designing appropriate utterances for tourism", in "Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems", CHI EA '20, p. 1–11 (Association for Computing Machinery, New York, NY, USA, 2020b), URL https://doi-org.libproxy.nau.edu/10.1145/3334480.3375033.

Chaves, A. P., E. Doerry, J. Egbert and M. Gerosa, "It's how you say it: Identifying appropriate register for chatbot language design", in "Proceedings of the 7th International Conference on Human-Agent Interaction (HAI '19)", p. 8 (ACM, New York, NY, USA, 2019a).

Chaves, A. P., J. Egbert and M. A. Gerosa, "Chatting like a robot: the relationship between linguistic choices and users' experiences", in "ACM CHI 2019 Workshop on Conversational Agents: Acting on the Wave of Research and Development", p. 8 (https://convagents.org/, Glasgow, UK, 2019b).

Chaves, A. P. and M. A. Gerosa, "Single or multiple conversational agents? an interactional coherence comparison", in "ACM SIGCHI Conference on Human Factors in Computing Systems", pp. 191:1–191:13 (ACM, New York, NY, USA, 2018).

Chaves, A. P. and M. A. Gerosa, "Survey on social characteristics of human-chatbot interaction", URL https://doi.org/10.5281/zenodo.3473358 (2019).

Chaves, A. P. and M. A. Gerosa, "How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design", International Journal of Human–Computer Interaction **0**, 0, 1–30, URL https://doi.org/10.1080/10447318.2020.1841438 (2020).

Chaves, A. P., T. Hocking, J. Egbert, E. Doerry and M. A. Gerosa, "Chatbots language design: the influence of language use on user experience", ACM Transactions on Computer-Human Interaction (TOCHI) **(under review)** (2021).

Chen, T. and C. Guestrin, "Xgboost: A scalable tree boosting system", in "Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining", KDD '16, p. 785–794 (Association for Computing Machinery, New York, NY, USA, 2016), URL https://doi.org/10.1145/2939672.2939785.

Christensen, R. H. B., "ordinal—regression models for ordinal data", R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal (2019).

Ciechanowski, L., A. Przegalinska, M. Magnuski and P. Gloor, "In the shades of the uncanny valley: An experimental study of human–chatbot interaction", Future Generation Computer Systems **92**, 539–548 (2018).

Coniam, D., "Evaluating the language resources of chatbots for their potential in English as a second language learning", ReCALL **20**, 1, 99–117 (2008).

Conrad, S. and D. Biber, *Multi-dimensional Studies of Register Variation in English* (Routledge, New York, NY, USA, 2014).

Corritore, C. L., R. P. Marble, S. Wiedenbeck, B. Kracher and A. Chandran, "Measuring online trust of websites: Credibility, perceived ease of use, and risk", in "AMCIS 2005 Proceedings", p. 370 (Association for Information Systems, Atlanta, GA, 2005).

Corti, K. and A. Gillespie, "Co-constructing intersubjectivity with artificial conversational agents: people are more likely to initiate repairs of misunderstandings with agents represented as human", Computers in Human Behavior **58**, 431–442 (2016).

Curry, A. C. and V. Rieser, "# metoo alexa: How conversational systems respond to sexual harassment", in "Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing", pp. 7–14 (Association for Computational Linguistics, New Orleans, Louisiana, USA, 2018).

Dahlbäck, N., A. Jönsson and L. Ahrenberg, "Wizard of oz studies–why and how", Knowledge-based systems **6**, 4, 258–266 (1993).

Dale, R., "The return of the chatbots", Natural Language Engineering **22**, 5, 811–817 (2016).

David Garson, G., "Partial least squares: Regression & structural equation models", (2016).

De Angeli, A., "To the rescue of a lost identity: Social perception in human-chatterbot interaction", in "Proceedings of the Joint Symposium on Virtual Social Agents", pp. 7–14 (The Society for the Study of Artificial Intelligence and the Simulation of Behaviour, Hatfield, UK, 2005).

De Angeli, A. and S. Brahnam, "Sex stereotypes and conversational agents", in "Proceedings of the AVI 2006 Workshop on Gender and Interaction: real and virtual women in a male world", (Online, Venice, Italy, 2006).

De Angeli, A., G. I. Johnson and L. Coventry, "The unfriendly user: exploring social reactions to chatterbots", in "Proceedings of The International Conference on Affective Human Factors Design, London", pp. 467–474 (Asean Academic Press, London, UK, 2001).

Dohsaka, K., R. Asai, R. Higashinaka, Y. Minami and E. Maeda, "Effects of conversational agents on activation of communication in thought-evoking multi-party dialogues", IEICE TRANSACTIONS on Information and Systems **97**, 8, 2147–2156 (2014).

Duijst, D., *Can we Improve the User Experience of Chatbots with Personalisation*, Master's thesis, University of Amsterdam (2017).

Duijvelshoff, W., "Use-cases and ethics of chatbots on plek: a social intranet for organizations", Workshop On Chatbots And Artificial Intelligence (2017).

Dyke, G., I. Howley, D. Adamson, R. Kumar and C. P. Rosé, "Towards academically productive talk supported by conversational agents", in "Intelligent Tutoring Systems", edited by S. A. Cerri, W. J. Clancey, G. Papadourakis and K. Panourgia, pp. 531–540 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013).

Egbert, J. and D. Biber, "Do all roads lead to rome?: Modeling register variation with factor analysis and discriminant analysis", Corpus Linguistics and Linguistic Theory **14**, 2, 233–273 (2016).

Elsholz, E., J. Chamberlain and U. Kruschwitz, "Exploring language style in chatbots to increase perceived product value and user engagement", in "Proceedings of the 2019 Conference on Human Information Interaction and Retrieval", pp. 301–305 (ACM, New York, NY, USA, 2019).

Facebook, "Tech for tourism–research page. accessible at: https://www.facebook.com/visitflagstaff/", (2018).

Feine, J., U. Gnewuch, S. Morana and A. Maedche, "A taxonomy of social cues for conversational agents", Int. J. Hum.-Comput. Stud. **132**, 138–161 (2019).

Ferrara, E., O. Varol, C. Davis, F. Menczer and A. Flammini, "The rise of social bots", Communications of the ACM **59**, 7, 96–104 (2016).

Finstad, K., "The usability metric for user experience", Interacting with Computers **22**, 5, 323–327 (2010).

Fitzpatrick, K. K., A. Darcy and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial", JMIR mental health **4**, 2, online (2017).

Flagstaff, "The city of flagstaff–arizona. available at: http://www.flagstaff.az.gov/2/community-profile", (2019).

Fodness, D. and B. Murray, "Tourist information search", Annals of tourism research **24**, 3, 503–523 (1997).

Fogg, B., "Computers as persuasive social actors", in "Persuasive Technology", edited by B. Fogg, Interactive Technologies, chap. 5, pp. 89 – 120 (Morgan Kaufmann, San Francisco, 2003).

Følstad, A. and P. B. Brandtzæg, "Chatbots and the new world of hci", interactions **24**, 4, 38–42 (2017).

Forlizzi, J., J. Zimmerman, V. Mancuso and S. Kwak, "How interface agents affect interaction between humans and computers", in "Proceedings of the 2007 conference on Designing pleasurable products and interfaces", pp. 209–221 (ACM, New York, NY, 2007).

Friedman, J., T. Hastie and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent", Journal of Statistical Software, Articles **33**, 1, 1–22, URL https://www.jstatsoft.org/v033/i01 (2010).

Galitsky, B., D. Ilvovsky and E. Goncharova, "On a chatbot providing virtual dialogues", in "Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)", pp. 382–387 (INCOMA Ltd., Varna, Bulgaria, 2019).

Garrido, P., J. Barrachina, F. J. Martinez and F. J. Seron, "Smart tourist information points by combining agents, semantics and ai techniques", Computer Science and Information Systems **14**, 1, 1–23 (2017).

Gavalas, D. and M. Kenteris, "A web-based pervasive recommendation system for mobile tourist guides", Personal and Ubiquitous Computing **15**, 7, 759–770 (2011).

Gnewuch, U., S. Morana and A. Maedche, "Towards designing cooperative and social conversational agents for customer service", in "International Conference on Information Systems 2017, Proceedings 1", (Association for Information Systems, South Korea, 2017).

Go, E. and S. S. Sundar, "Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions", Computers in Human Behavior **97**, 304–316 (2019).

Goertzel, B., "Artificial general intelligence: concept, state of the art, and future prospects", Journal of Artificial General Intelligence **5**, 1, 1–48 (2014).

Gong, L., "How social is social responses to computers? the function of the degree of anthropomorphism in computer representations", Computers in Human Behavior **24**, 4, 1494–1509 (2008).

Grand View Research, "Chatbot market size to reach $1.25 billion by 2025. cagr: 24.3%", Retrieved from: https://www.grandviewresearch.com/press-release/global-chatbot-market (2017).

Gu, X., K. Cho, J.-W. Ha and S. Kim, "Dialogwae: Multimodal response generation with conditional wasserstein auto-encoder", CoRR **abs/1805.12352**, URL http://arxiv.org/abs/1805.12352, published as a conference paper at ICLR 2019 (2018).

Gunawardena, C. N. and F. J. Zittle, "Social presence as a predictor of satisfaction within a computer-mediated conferencing environment", American journal of distance education **11**, 3, 8–26 (1997).

Hair, J. F., J. J. Risher, M. Sarstedt and C. M. Ringle, "When to use and how to report the results of pls-sem", European Business Review (2019).

Hayashi, Y., "Social facilitation effects by pedagogical conversational agent: Lexical network analysis in an online explanation task", in "International Conference on Educational Data Mining (EDM)", (International Educational Data Mining Society, Japan, 2015).

Hayashi, Y., "The effect of" mood": Group-based collaborative problem solving by taking different perspectives", in "Proceedings of the 38th Annual Conference of the Cognitive Science Society(CogSci2016)", pp. 818–823 (The Cognitive Science Society, Philadelphia, USA, 2016).

Heyselaar, E. and T. Bosse, "Using theory of mind to assess users' sense of agency in social chatbots", in "Conversations 2019: 3rd International Workshop on Chatbot Research", pp. 1–13 (Cham: Springer, Cham, 2019).

Hill, J., W. R. Ford and I. G. Farreras, "Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations", Computers in Human Behavior **49**, 245–250 (2015).

Hirschberg, J. and C. D. Manning, "Advances in natural language processing", Science **349**, 6245, 261–266 (2015).

Ho, A., J. Hancock and A. S. Miner, "Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot", Journal of Communication **68**, 4, 712–733 (2018).

Hoegen, R., D. Aneja, D. McDuff and M. Czerwinski, "An end-to-end conversational style matching agent", in "Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents", IVA '19, p. 111–118 (Association for Computing Machinery, New York, NY, USA, 2019).

Hothorn, T., P. Buehlmann, S. Dudoit, A. Molinaro and M. Van Der Laan, "Survival ensembles", Biostatistics **7**, 3, 355–373 (2006).

Hunston, S. *et al.*, *Corpus approaches to evaluation: Phraseology and evaluative language* (Routledge, New York, NY, US, 2010).

Hyland, K., "Bundles in academic discourse", Annual Review of Applied Linguistics **32**, 150–169 (2012).

ISO 9241-11, *Ergonomics of human-system interaction: Part 11: Usability: Definitions and concepts* (International Organization for Standardization, ISO, 2018).

Ivanov, S. and C. Webster, "Adoption of robots, artificial intelligence and service automation by travel, tourism and hospitality companies–a cost-benefit analysis", in "INVTUR Conference", pp. 19–21 (Prepared for the International Scientific Conference "Contemporary tourism–traditions and innovations, Sofia University, 2017).

Jabri, M., A. D. Adrian and D. Boje, "Reconsidering the role of conversations in change communication: A contribution based on bakhtin", Journal of Organizational Change Management **21**, 6, 667–685 (2008).

Jain, M., R. Kota, P. Kumar and S. N. Patel, "Convey: Exploring the Use of a Context View for Chatbots", in "Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems", p. 468 (ACM, New York, NY, USA, 2018a).

Jain, M., P. Kumar, R. Kota and S. N. Patel, "Evaluating and Informing the Design of Chatbots", in "Proceedings of the 2018 on Designing Interactive Systems Conference 2018", pp. 895–906 (ACM, New York, NY, USA, 2018b).

Jakic, A., M. O. Wagner and A. Meyer, "The impact of language style accommodation during social media interactions on brand trust", Journal of Service Management **28**, 3, 418–441 (2017).

Jenkins, M.-C., R. Churchill, S. Cox and D. Smith, "Analysis of user interaction with service oriented chatbot systems", in "Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments", edited by J. A. Jacko, pp. 76–83 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2007).

Jiang, R. and R. E Banchs, "Towards improving the performance of chat oriented dialogue system", in "2017 International Conference on Asian Language Processing (IALP)", pp. 23–26 (IEEE, New York, NY, USA, 2017).

Kamberelis, G., "Genre as institutionally informed social practice", J. Contemp. Legal Issues **6**, 115 (1995).

Katkute, K. *et al.*, *Designing user engagement with text-based chatbots*, Master's thesis, Psychologie Faculteit der Sociale Wetenschappen–Universiteit Leiden (2017).

Keijsers, M. and C. Bartneck, "Mindless robots get bullied", in "Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction", pp. 205–214 (ACM, New York, NY, USA, 2018).

Kessler, B., "Computational dialectology in irish gaelic", in "Proceedings of the Seventh Conference on European Chapter of the Association for Computational Linguistics", EACL '95, p. 60–66 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995), URL https://doi.org/10.3115/976973.976983.

Kilgarriff, A., "Language is never, ever, ever, random", Corpus linguistics and linguistic theory **1**, 2, 263–276 (2005).

Kirakowski, J., A. Yiu *et al.*, "Establishing the hallmarks of a convincing chatbot-human dialogue", in "Human-Computer Interaction", (InTech, London, UK, 2009).

Kiseleva, J., K. Williams, A. Hassan Awadallah, A. C. Crook, I. Zitouni and T. Anastasakos, "Predicting user satisfaction with intelligent assistants", in "Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval", pp. 45–54 (ACM, New York, NY, USA, 2016).

Komatsu, T., R. Kurosawa and S. Yamada, "How does the difference between users' expectations and perceptions about a robotic agent affect their behavior?", International Journal of Social Robotics **4**, 2, 109–116 (2012).

Komiak, S. Y. and I. Benbasat, "The effects of personalization and familiarity on trust and adoption of recommendation agents", MIS quarterly **30**, 4, 941–960 (2006).

Krauss, R. M. and C.-Y. Chiu, "Language and social behavior", in "The handbook of social psychology", edited by D. T. Gilbert, S. T. Fiske and G. Lindzey, pp. 41–88 (McGraw-Hill, New York, NY, US, 1998).

Kumar, R., H. Ai, J. L. Beuth and C. P. Rosé, "Socially capable conversational tutors can be effective in collaborative learning situations", in "International Conference on Intelligent Tutoring Systems", edited by V. Aleven, J. Kay and J. Mostow, pp. 156–164 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2010).

Laban, G. and T. Araujo, "Working together with conversational agents: The relationship of perceived cooperation with service performance evaluations", in "Chatbot Research and Design", edited by A. Følstad, T. Araujo, S. Papadopoulos, E. L.-C. Law, O.-C. Granmo, E. Luger and P. B. Brandtzaeg, pp. 215–228 (Springer International Publishing, Cham, 2020).

Labov, W., S. Ash and C. Boberg, *The atlas of North American English: Phonetics, phonology and sound change* (Walter de Gruyter, Boston, MA, USA, 2005).

Lang, T. C., "The effect of the internet on travel consumer purchasing behaviour and implications for travel agencies", Journal of vacation marketing **6**, 4, 368–385 (2000).

Lasek, M. and S. Jessa, "Chatbots for Customer Service on Hotels' Websites", Information Systems in Management **2**, 2, 146–158 (2013).

Lee, M., G. Lucas, J. Mell, E. Johnson and J. Gratch, "What's on your virtual mind?: Mind perception in human-agent negotiations", in "Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents", IVA '19, pp. 38–45 (ACM, New York, NY, USA, 2019), URL http://doi.acm.org/10.1145/3308532.3329465.

Lee, M. K., S. Kiesler and J. Forlizzi, "Receptionist or information kiosk: How do people talk with a robot?", in "Proceedings of the 2010 ACM CSCW", pp. 31–40 (ACM, New York, NY, USA, 2010).

Lee, S. and J. Choi, "Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity", International Journal of Human-Computer Studies **103**, 95–105 (2017).

Leech, G. N. and M. Short, *Style in fiction: A linguistic introduction to English fictional prose*, no. 13 in English language series (Pearson Education, London, UK, 2007).

Li, Y., H. Su, X. Shen, W. Li, Z. Cao and S. Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset", in "International Joint Conference on Natural Language Processing (IJCNLP)", pp. 986–995 (Asian Federation of Natural Language Processing, Taipei, Taiwan, 2017).

Liao, Q. V., M. Mas-ud Hussain, P. Chandar, M. Davis, Y. Khazaeni, M. P. Crasso, D. Wang, M. Muller, N. S. Shami and W. Geyer, "All work and no play? conversations with aquestion-and-answer chatbot in the wild", in "Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems", pp. 3:1–3:13 (ACM, New York, NY, USA, 2018).

Liao, V. Q., M. Davis, W. Geyer, M. Muller and N. S. Shami, "What can you do?: Studying social-agent orientation and agent proactive interactions with an agent for employees", in "Proceedings of the 2016 ACM Conference on Designing Interactive Systems", pp. 264–275 (ACM, New York, NY, USA, 2016).

Lin, G. I. and M. A. Walker, "Stylistic variation in television dialogue for natural language generation", in "Proceedings of the Workshop on Stylistic Variation", pp. 85–93 (Association for Computational Linguistics, Copenhagen, Denmark, 2017).

Linden, G., S. Hanks and N. Lesh, "Interactive assessment of user preference models: The automated travel assistant", in "User Modeling", pp. 67–78 (Springer, Vienna, Vienna, 1997).

Long, M. H. and S. Ross, "Modifications that preserve language and content.", Tech. rep., ERIC (1993).

Loo, J., "The future of travel: New consumer behavior and the technology giving it flight", Google/Phocuswright Travel Study 2017 (2017).

Luger, E. and A. Sellen, "Like having a really bad pa: The gulf between user expectation and experience of conversational agents", in "Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems", pp. 5286–5297 (ACM, New York, NY, USA, 2016).

Mack, R. W., J. E. Blose and B. Pan, "Believe it or not: Credibility of blogs in tourism", Journal of Vacation marketing **14**, 2, 133–144 (2008).

Mairesse, F. and M. A. Walker, "Can conversational agents express big five personality traits through language?: Evaluating a psychologically-informed language generator", Cambridge & Sheffield, United Kingdom: Cambridge University Engineering Department & Department of Computer Science, University of Sheffield (2009).

Marino, M. C., *I, chatbot: the gender and race performativity of conversational agents*, Ph.D. thesis, University of California, Riverside (2006).

Marino, M. C., "The racial formation of chatbots", CLCWeb: Comparative Literature and Culture **16**, 5, 13 (2014).

Maslowski, I., D. Lagarde and C. Clavel, "In-the-wild chatbot corpus: from opinion analysis to interaction problem detection", in "International Conference on Natural Language, Signal and Speech Processing", pp. 115–120 (International Science and General Applications, Marocco, 2017).

Massaro, D. W., M. M. Cohen, S. Daniel and R. A. Cole, "Developing and evaluating conversational agents", in "Human performance and ergonomics", edited by P. Hancock, chap. 7, pp. 173–194 (Elsevier, 1999), second edn.

Mauldin, M. L., "Chatterbots, tinymuds, and the turing test entering the loebner prize competition", in "Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence", AAAI'94, pp. 16–21 (AAAI Press, Seattle, Washington, 1994).

Mäurer, D. and K. Weihe, "Benjamin Franklin's decision method is acceptable and helpful with a conversational agent", in "Intelligent Interactive Multimedia Systems and Services", pp. 109–120 (Springer International Publishing, Cham, 2015).

McGrath, J. E., "Dilemmatics: The study of research choices and dilemmas", American Behavioral Scientist **25**, 2, 179–210 (1981).

McNamara, N. and J. Kirakowski, "Functionality, usability, and user experience", interactions **13**, 6, 26–28 (2006).

Meany, M. M. and T. Clark, "Humour theory and conversational agents: An application in the development of computer-based agents", International Journal of the Humanities **8**, 5, 129–140 (2010).

Mensio, M., G. Rizzo and M. Morisio, "Multi-turn qa: A rnn contextual approach to intent classification for goal-oriented systems", in "Companion Proceedings of the The Web Conference 2018", pp. 1075–1080 (2018).

Miner, A., A. Chow, S. Adler, I. Zaitsev, P. Tero, A. Darcy and A. Paepcke, "Conversational agents and mental health: Theory-informed assessment of language and affect", in "Proceedings of the Fourth International Conference on Human Agent Interaction", pp. 123–130 (ACM, New York, NY, USA, 2016).

Montenegro, J. L. Z., C. A. da Costa and R. da Rosa Righi, "Survey of conversational agents in health", Expert Systems with Applications **129**, 56–67 (2019).

Morana, S., U. Gnewuch, D. Jung and C. Granig, "The effect of anthropomorphism on investment decision-making with robo-advisor chatbots.", in "ECIS", (2020).

Morris, T. W., "Conversational agents for game-like virtual environments", in "Artificial Intelligence and Interactive Entertainment.", pp. 82–86 (AAAI Press, Palo Alto, CA, USA, 2002).

Morrissey, K. and J. Kirakowski, "'realness' in chatbots: Establishing quantifiable criteria", in "International Conference on Human-Computer Interaction", pp. 87–96 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013).

Mou, Y. and K. Xu, "The media inequality: Comparing the initial human-human and human-ai social interactions", Computers in Human Behavior **72**, 432–440 (2017).

Nass, C. and Y. Moon, "Machines and mindlessness: Social responses to computers", Journal of social issues **56**, 1, 81–103 (2000).

Nass, C., J. Steuer and E. R. Tauber, "Computers are social actors", in "Proceedings of the SIGCHI conference on Human factors in computing systems", pp. 72–78 (ACM, New York, NY, USA, 1994).

Neururer, M., S. Schlögl, L. Brinkschulte and A. Groth, "Perceptions on authenticity in chat bots", Multimodal Technologies and Interaction **2**, 3, 60 (2018).

Nevill, C. and T. Bell, "Compression of parallel texts", Information Processing & Management **28**, 6, 781–793 (1992).

Niculescu, A. I., K. H. Yeo, L. F. D'Haro, S. Kim, R. Jiang and R. E. Banchs, "Design and evaluation of a conversational agent for the touristic domain", in "Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)", pp. 1–10 (IEEE, Siem Reap, Cambodia, 2014).

Niederhoffer, K. G. and J. W. Pennebaker, "Linguistic style matching in social interaction", Journal of Language and Social Psychology **21**, 4, 337–360 (2002).

Noor, A. K., "Potential of cognitive computing and cognitive systems", Open Engineering **5**, 1, 75–88 (2015).

Nordheim, C. B., A. Følstad and C. A. Bjørkli, "An initial model of trust in chatbots for customer service—findings from a questionnaire study", Interacting with Computers **31**, 3, 317–335 (2019).

Palan, S. and C. Schitter, "Prolific. ac—a subject pool for online experiments", Journal of Behavioral and Experimental Finance **17**, 22–27 (2018).

Paltridge, B., "Genre analysis and the identification of textual boundaries", Applied linguistics **15**, 3, 288–299 (1994).

Pawłowska, A., "Tourists and social media: Already inseparable marriage or still a long-distance relationship? analysis of focus group study results conducted among tourists using social media", World Scientific News **57**, 106–115 (2016).

Petta, P. and R. Trappl, *Why to create personalities for synthetic actors*, chap. Creating Personalities for Synthetic Actors: Towards Autonomous Personality Agents, pp. 1–8 (Springer, Berlin, Heidelberg, 1997).

Pillai, R. and B. Sivathanu, "Adoption of ai-based chatbots for hospitality and tourism", International Journal of Contemporary Hospitality Management (2020).

Portela, M. and C. Granell-Canut, "A new friend in our smartphone? observing interactions with chatbots in the search of emotional engagement", in "Interaccion", pp. 48:1–48:7 (ACM, New York, NY, USA, 2017).

Prabhumoye, S., Y. Tsvetkov, R. Salakhutdinov and A. W. Black, "Style transfer through back-translation", in "56th Annual Meeting of the Association for Computational Linguistics", (Association for Computational Linguistics, Philadelphia, USA, 2018).

Przegalinska, A., L. Ciechanowski, A. Stroz, P. Gloor and G. Mazurek, "In bot we trust: A new methodology of chatbot performance measures", Business Horizons (2019).

Ptaszynski, M., P. Dybala, S. Higuhi, W. Shi, R. Rzepka and K. Araki, "Towards socialized machines: Emotions and sense of humour in conversational agents", in "Web Intelligence and Intelligent Agents", edited by Z. ul-hassan Usmani (InTech, Rijeka, Croatia, 2010).

Radlinski, F. and N. Craswell, "A theoretical framework for conversational search", (2017).

Radziwill, N. M. and M. C. Benton, "Evaluating quality of chatbots and intelligent conversational agents", Software Quality Professional **19**, 3, 25–36 (2017).

Raij, A. B., K. Johnsen, R. F. Dickerson, B. C. Lok, M. S. Cohen, M. Duerson, R. R. Pauly, A. O. Stevens, P. Wagner and D. S. Lind, "Comparing interpersonal interactions with a virtual human to those with a real human", IEEE transactions on visualization and computer graphics **13**, 3, 443–457 (2007).

Reiter, E. and R. Dale, *Building natural language generation systems* (Cambridge university press, New York, NY, USA, 2000).

Ricci, F., L. Rokach and B. Shapira, "Introduction to recommender systems handbook", in "Recommender systems handbook", pp. 1–35 (Springer, Boston, MA, USA, 2011).

Ringle, C. M., S. Wende and J.-M. Becker, "Smartpls 3", http://www.smartpls.com (2015).

Salovey, P. and J. D. Mayer, "Emotional intelligence", Imagination, cognition and personality **9**, 3, 185–211 (1990).

Sano, A. V. D., T. D. Imanuel, M. I. Calista, H. Nindito and A. R. Condrobimo, "The application of agnes algorithm to optimize knowledge base for tourism chatbot", in "2018 International Conference on Information Management and Technology (ICIMTech)", pp. 1–9 (IEEE, Jakarta, Indonesia, 2018).

Sato, E., "A guide to linguistic modification: Strategies for increasing english language learner access to academic content", (2007).

Schanke, S., G. Burtch and G. Ray, "Estimating the impact of 'humanizing'customer service chatbots", (2020).

Schlesinger, A., K. P. O'Hara and A. S. Taylor, "Let's talk about race: Identity, chatbots, and ai", in "Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems", pp. 315:1–315:14 (ACM, New York, NY, USA, 2018).

Schuetzler, R. M., G. M. Grimes and J. S. Giboney, "An investigation of conversational agent relevance, presence, and engagement", in "Americas Conference on Information Systems 2018 Proceedings", (Association for Information Systems, New Orleans, 2018).

Shawar, B. A. and E. Atwell, "Chatbots: are they really useful?", in "LDV Forum", vol. 22:1, pp. 29–49 (GSCL German Society for Computational Linguistics, German, 2007).

Shechtman, N. and L. M. Horowitz, "Media inequality in conversation: how people behave differently when interacting with computers and people", in "Proceedings of the SIGCHI conference on Human factors in computing systems", pp. 281–288 (ACM, New York, NY, USA, 2003).

Shen, X., H. Su, S. Niu and V. Demberg, "Improving variational encoder-decoders in dialogue generation", in "Thirty-Second AAAI Conference on Artificial Intelligence", pp. 5456–5463 (AAAI Press, Palo Alto, California USA, 2018).

Short, J., E. Williams and B. Christie, *The social psychology of telecommunications* (John Wiley and Sons Ltd, London, UK, 1976).

Shum, H.-y., X.-d. He and D. Li, "From Eliza to XiaoIce: challenges and opportunities with social chatbots", Frontiers of Information Technology & Electronic Engineering **19**, 1, 10–26 (2018).

Silvervarg, A. and A. Jönsson, "Iterative development and evaluation of a social conversational agent", in "6th International Joint Conference on Natural Language Processing (IJCNLP 2013)", pp. 1223–1229 (Asian Federation of Natural Language Processing, Nagoya, Japan, 2013).

Silvervarg, A., K. Raukola, M. Haake and A. Gulz, "The effect of visual gender on abuse in conversation with ecas", in "International Conference on Intelligent Virtual Agents", edited by Y. Nakano, M. Neff, A. Paiva and M. Walker, pp. 153–160 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2012).

Sjödén, B., A. Silvervarg, M. Haake and A. Gulz, "Extending an educational math game with a pedagogical conversational agent: Facing design challenges", in "Interdisciplinary Approaches to Adaptive Learning. A Look at the Neighbours", pp. 116–130 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011).

Stets, J. E. and P. J. Burke, "Identity theory and social identity theory", Social psychology quarterly **63**, 3, 224–237 (2000).

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin and A. Zeileis, "Conditional variable importance for random forests", BMC Bioinformatics **9**, 307, URL http://www.biomedcentral.com/1471-2105/9/307 (2008).

Strobl, C., A.-L. Boulesteix, A. Zeileis and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution", BMC Bioinformatics **8**, 25, URL http://www.biomedcentral.com/1471-2105/8/25 (2007).

Sundar, S. S., S. Bellur, J. Oh, H. Jia and H.-S. Kim, "Theoretical importance of contingency in human-computer interaction: effects of message interactivity on user engagement", Communication Research **43**, 5, 595–625 (2016).

Sweeney, J. and J. Swait, "The effects of brand credibility on customer loyalty", Journal of retailing and consumer services **15**, 3, 179–193 (2008).

Syed, B. H., *Adapting Language Models for Style Transfer*, Master's thesis, International Institute of Information Technology Hyderabad (2020).

Szmrecsanyi, B., "Corpus-based dialectometry: a methodological sketch", Corpora **6**, 1, 45–76 (2011).

Tallyn, E., H. Fried, R. Gianni, A. Isard and C. Speed, "The Ethnobot: Gathering Ethnographies in the Age of IoT", in "Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems", p. 604 (ACM, New York, NY, USA, 2018).

Tamayo-Moreno, S. and D. Pérez-Marín, "Designing and evaluating pedagogic conversational agents to teach children", World Academy of Science, Engineering and Technology, International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering **11**, 3, 491–496 (2017).

Tariverdiyeva, G., *Chatbots' Perceived Usability in Information Retrieval Tasks: An Exploratory Analysis.*, Master's thesis, University of Twente (2019).

Tegos, S., S. Demetriadis, P. M. Papadopoulos and A. Weinberger, "Conversational agents for academically productive talk: a comparison of directed and undirected agent interventions", International Journal of Computer-Supported Collaborative Learning **11**, 4, 417–440 (2016a).

Tegos, S., S. Demetriadis and T. Tsiatsos, "An investigation of conversational agent interventions supporting historical reasoning in primary education", in "International Conference on Intelligent Tutoring Systems", edited by A. Micarelli, J. Stamper and K. Panourgia, pp. 260–266 (Springer International Publishing, Cham, 2016b).

Thies, I. M., N. Menon, S. Magapu, M. Subramony and J. O'neill, "How do you want your chatbot? an exploratory wizard-of-oz study with young, urban indians", in "Human-Computer Interaction–INTERACT", edited by B. R., D. G., J. A., K. B. D., O. J. and W. M., vol. 10513 of *Lecture Notes in Computer Science*, pp. 441–459 (Springer, Cham, Switzerland, 2017).

Think with Google, "How micro-moments are reshaping the travel customer journey. available at: https://www.thinkwithgoogle.com/marketing-resources/micro-moments/micro-moments-travel-customer-journey/", online (2016).

Thomas, P., M. Czerwinski, D. McDuff, N. Craswell and G. Mark, "Style and alignment in information-seeking conversation", in "Proceedings of the 2018 Conference on Human Information Interaction&Retrieval", pp. 42–51 (ACM, New York, NY, USA, 2018).

Thomas Combrink, R. R., Melinda Bradford, "2017-2018 flagstaff visitor survey", Tech. rep., Alliance Bank Eonomic Policy Institute, The W.A. Franke College of Business, Northern Arizona University. Prepared for the Flagstaff Convention and Visitors Bureau, Arizona Office of Tourism (2018).

Tikhonov, A. and I. P. Yamshchikov, "What is wrong with style transfer for texts?", CoRR **abs/1808.04365**, URL http://arxiv.org/abs/1808.04365 (2018).

Toxtli, C., J. Cranshaw *et al.*, "Understanding chatbot-mediated task management", in "Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems", p. 58 (ACM, New York, NY, USA, 2018).

Tran, Q. H., T. Lai, G. Haffari, I. Zukerman, T. Bui and H. Bui, "The context-dependent additive recurrent neural net", in "Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)", pp. 1274–1283 (2018).

Tullis, T. and W. Albert, *Measuring the User Experience, Second Edition: Collecting, Analyzing, and Presenting Usability Metrics* (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2013), 2nd edn.

Turing, A. M., "Computing machinery and intelligence", Mind **59**, 236, 433–460 (1950).

Tussyadiah, I., "A review of research into automation in tourism: Launching the annals of tourism research curated collection on artificial intelligence and robotics in tourism", Annals of Tourism Research **81**, 102883 (2020).

UNWTO, "Unwto tourism highlights, 2017 edition", World Tourism Organization, Madrid (2017).

Valério, F. A., T. G. Guimarães, R. O. Prates and H. Candello, "Here's what i can do: Chatbots' strategies to convey their features to users", in "Proceedings of the XVI Brazilian Symposium on Human Factors in Computing Systems", p. 28 (ACM, New York, NY, USA, 2017).

Vinciarelli, A., A. Esposito, E. André, F. Bonin, M. Chetouani, J. F. Cohn, M. Cristani, F. Fuhrmann, E. Gilmartin, Z. Hammal *et al.*, "Open challenges in modelling, analysis and synthesis of human behaviour in human–human and human–machine interactions", Cognitive Computation **7**, 4, 397–413 (2015).

Wallace, R. S., "The anatomy of a.l.i.c.e.", in "Parsing the Turing Test", pp. 181–210 (Springer Netherlands, Dordrecht, 2009).

Wallis, P. and E. Norling, "The trouble with chatbots: social skills in a social world", in "Proceedings of the Joint Symposium on Virtual Social Agents", pp. 29–38 (The Society for the Study of Artificial Intelligence and the Simulation of Behaviour, Hatfield, UK, 2005).

Walther, J. B., "Selective self-presentation in computer-mediated communication: Hyperpersonal dimensions of technology, language, and cognition", Computers in Human Behavior **23**, 5, 2538–2557 (2007).

Wang, D., Z. Xiang and D. R. Fesenmaier, "Smartphone use in everyday life and travel", Journal of Travel Research **55**, 1, 52–63 (2016).

Weizenbaum, J., "Eliza—a computer program for the study of natural language communication between man and machine", Communications of the ACM **9**, 1, 36–45 (1966).

Wessel, M., B. M. de Souza, I. Steinmacher, I. S. Wiese, I. Polato, A. P. Chaves and M. A. Gerosa, "The power of bots: Characterizing and understanding bots in oss projects", Proc. ACM Hum.-Comput. Interact. **2**, CSCW, URL https://doi-org.libproxy.nau.edu/10.1145/3274451 (2018).

Wikibooks, "Algorithm implementation/strings/levenshtein distance — wikibooks, the free textbook project", [Online; accessed 8-May-2020] (2020).

Winograd, T., "Procedures as a representation for data in a computer program for understanding natural language", Tech. rep., DTIC Document (1971).

Zamora, J., "I'm sorry, dave, i'm afraid i can't do that: Chatbot perception and expectations", in "Proceedings of the 5th International Conference on Human Agent Interaction", pp. 253–260 (ACM, New York, NY, USA, 2017).

Zdravkova, K., "Conceptual framework for an intelligent chatterbot", in "Information Technology Interfaces, 2000. ITI 2000. Proceedings of the 22nd International Conference on", pp. 189–194 (IEEE, New York, NY, USA, 2000).

Zhang, W.-N., Q. Zhu, Y. Wang, Y. Zhao and T. Liu, "Neural personalized response generation as domain adaptation", World Wide Web **22**, 4, 1427–1446 (2017).

Zhao, T., K. Lee and M. Eskenazi, "Unsupervised discrete sentence representation learning for interpretable neural dialog generation", in "Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics", vol. 1: Long Papers, pp. 1098–1107 (Association for Computational Linguistics, Melbourne, Australia, 2018a), URL https://www.aclweb.org/anthology/P18-1101.

Zhao, Y., V. W. Bi, D. Cai, X. Liu, K. Tu and S. Shi, "Language style transfer from non-parallel text with arbitrary styles", CLR 2018 Conference Withdrawn Submission (2018b).

Zhu, P., Z. Zhang, J. Li, Y. Huang and H. Zhao, "Lingke: A fine-grained multi-turn chatbot for customer service", in "Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations", pp. 108–112 (Association for Computational Linguistics, Santa Fe, New Mexico, 2018).

Zue, V. W. and J. R. Glass, "Conversational interfaces: Advances and challenges", Proceedings of the IEEE **88**, 8, 1166–1180 (2000).

Zumstein, D. and S. Hundertmark, "Chatbots–an interactive technology for personalized communication, transactions and services", IADIS International Journal on WWW/Internet **15**, 1, 96–109 (2017).

APPENDIX A

ACRONYMS AND ABBREVIATIONS

A.L.I.C.E.: Artificial Linguistic Internet Computer Entity

ACM: Association for Computing Machinery

AIML: Artificial Intelligence Markup Language

ANOVA: Analysis of the Variance

ANTHR: Anthropomorphism

APP: Perceived Appropriateness of Language

AUC: Area Under the Curve

AVE: Average Variance Extracted

AZ: Arizona

CHI: Conference on Human Factors in Computing Systems

CLMM: Cumulative Link Mixed Model

CMC: Computer-Mediated Communication

CR: Composite Reliability

CRED: Credibility

DSTC: Dialog State-Tracking Challenge

ECA: Embodied Conversational Agent

FDR: False Discovery Rate

HAI: [International Conference on] Human-Agent Interaction

HCI: Human-Computer Interaction

IBM: International Business Machines

IRB: Institutional Review Board

ISO: International Organization for Standardization

MANOVA : Multivariate Analysis of the Variance

MAT: Multidimensional Analysis Tagger

NAU: Northern Arizona University

PLS-SEM: Partial Least Square–Structural Equation Modeling

ROC: Receiver Operator Characteristic

RQ: Research Question

SOC_PR: Social Presence

USA: United States of America

UX: User Experience

WoZ: Wizard of Oz

APPENDIX B

OVERVIEW OF THE SURVEYED LITERATURE

The literature presents no coherent definition of chatbots; thus, to find relevant studies we used a search string that includes synonyms *chatbots, chatterbots, conversational agents, conversational interfaces, conversational systems, conversation systems, dialogue systems, digital assistants, intelligent assistants, conversational user interfaces,* and *conversational UI*. We explicitly left out studies that relate to embodiment (e.g., *ECA, multimodal, robots, eye-gaze, gesture*), and speech input mode (e.g., *speech-based, speech-recognition, voice-based*). The search string did not include the term social bots, because it refers to chatbots that produce content for social networks such as Twitter (Ferrara *et al.*, 2016). We did not include personal assistants either, since this term consistently refers to commercially available, voice-based assistants such as Google Assistant, Amazon Alexa, Apple Siri, and Microsoft Cortana. We decided not to include terms that relate to social characteristics/traits, because most studies do not explicitly label their results as such. Additionally, as we expected to find relevant studies in a variety of domains (not only computing-related venues), we used Google Scholar as the search engine. Even selecting to search in title and abstract only, the search resulted in a massive number of papers to be analyzed. We started with a set of about one thousand papers. In the first round, we reduced to half of this amount by reading the titles only. We also removed the papers in which full text was not available or written in a language other than English. We started the second round with 464 studies. In this round, we excluded short/position papers, master/doctoral thesis, and book chapters, sticking with papers published in scientific venues. We also read the abstracts and removed all the paper that focused on technical aspects, rather than social. The third and last round started with 104 papers. After reading the full text, we removed the papers that either do not highlight social characteristics or are previous, ongoing research of another complete studied from the same authors. The datasets with all analyzed studies (original list of results and exclusions per round) are available in (Chaves and Gerosa, 2019). After filtering the search results, we had 56 remaining studies.

We analyzed the papers by searching for chatbot behavior or attributed characteristics that influence the way users perceive and behave toward chatbots. Noticeably, the characteristics and categories are seldom explicitly pointed to in the literature, so the conceptual model was derived using a qualitative open coding process. For each study (*documents*), we selected relevant statements from the paper (*quotes*) and labeled them as a characteristic (*code*). After coding all the studies, a second researcher reviewed the produced set of characteristics, and discussion sessions were performed to identify characteristics that could be merged, renamed, or removed. It the end, the characteristics were grouped into the categories, depending on whether the characteristic relates to the chatbot's virtual representation, conversational behavior, or social protocols. Finally, the quotes for each characteristic were labeled as references to benefits, challenges, or strategies. We derived

a total of 11 social characteristics, and grouped them into three categories: **conversational intelligence**, **social intelligence**, and **personification**.

Table B.1 describes the list of topics the chatbots could handle along with the number of papers. Most of the selected studies are recent publications (less than 10 years). The publication venues include the domains of human-computer interactions (25 papers), learning and education (8 papers), information and interactive systems (8 papers), virtual agents (5 papers), artificial intelligence (3 papers), and natural language processing (2 papers). We also found papers from health, literature & culture, computer systems, communication, and humanities (1 paper each). Most papers (59%) focus on task-oriented chatbots. General purpose chatbots reflect 33% of the surveyed studies. Most general purpose chatbots (16 out of 19) are designed to handle topic-unrestricted conversations. The most representative specific-domain is education, with 9 papers, followed by customer services, with 5 papers.

Table B.1: Conversational topics for chatbots in the surveyed studies

# of papers	Topics handled by the chatbots
16	Open domain (unrestricted topics)
9	Education
5	Customer services
2	Financial services, game, health, information search, race, task management, virtual assistants
1	Business, credibility assessment interviews, e-commerce, decision-making coach, ethnography, human resources, humor, movie recommendation, news, tourism

Most surveyed studies adopted real chatbots (35 out of 56); among them, 18 studies analyze logs of conversations or users' perceptions of third-party chatbots such as Cleverbot (e.g., (Corti and Gillespie, 2016)), Talkbot (e.g., (Brahnam and De Angeli, 2012)), and Woebot (Fitzpatrick *et al.*, 2017). Nine studies introduce a self-developed architecture and/or dialogue management (e.g. (Dohsaka *et al.*, 2014; Kumar *et al.*, 2010)). In other nine studies, the chatbots were designed for research purpose using third-party platforms for chatbots development and deployment, such as IBM Watson service (e.g., (Liao *et al.*, 2018)) and Microsoft Bot Framework (Toxtli *et al.*, 2018) as well as pattern-matching packages such as Artificial Intelligence Markup Language (AIML) (Silvervarg and Jönsson, 2013). When a chatbot is simulated (11 studies), Wizard of Oz (WoZ) is the most used technique (9 studies). In a WoZ study, participants believe to be interacting with a chatbot when, in fact, a person (or wizard) pretends to be the automated system (Dahlbäck *et al.*, 1993). Eight studies do not address a particular chatbot. Tables B.2 and B.3 reports whether the studies investigates real or simulated chatbots and, in case of real chatbots, the platforms used to develop them.

Table B.2: Chatbots introduced in the reviewed literature

Chatbot investigated	**#papers**	**Surveyed studies**
None	8	(Brandtzaeg and Følstad, 2017) (Brandtzaeg and Følstad, 2018) (Gnewuch *et al.*, 2017) (Mairesse and Walker, 2009) (Meany and Clark, 2010) (Morris, 2002) (Neururer *et al.*, 2018) (Schlesinger *et al.*, 2018)
Real chatbot	35	(Araujo, 2018) (Ayedoun *et al.*, 2017) (Brahnam and De Angeli, 2012) (Ciechanowski *et al.*, 2018) (Coniam, 2008) (Corti and Gillespie, 2016) (Curry and Rieser, 2018) (De Angeli, 2005) (De Angeli and Brahnam, 2006) (De Angeli *et al.*, 2001) (Dohsaka *et al.*, 2014) (Duijst, 2017) (Fitzpatrick *et al.*, 2017) (Hayashi, 2015) (Hill *et al.*, 2015) (Jain *et al.*, 2018a) (Jain *et al.*, 2018b) (Kirakowski *et al.*, 2009) (Kumar *et al.*, 2010) (Lasek and Jessa, 2013) (Liao *et al.*, 2016) (Liao *et al.*, 2018) (Marino, 2014) (Mäurer and Weihe, 2015) (Miner *et al.*, 2016) (Morrissey and Kirakowski, 2013) (Ptaszynski *et al.*, 2010) (Schuetzler *et al.*, 2018) (Shum *et al.*, 2018) (Silvervarg and Jönsson, 2013) (Tallyn *et al.*, 2018) (Tegos *et al.*, 2016b) (Toxtli *et al.*, 2018) (Valério *et al.*, 2017) (Zamora, 2017)
Simulated chatbot	11	Wizard of Oz ((Avula *et al.*, 2018) (Chaves and Gerosa, 2018) (Duijvelshoff, 2017) (Dyke *et al.*, 2013) (Ho *et al.*, 2018) (Lee and Choi, 2017) (Sjödén *et al.*, 2011) (Thies *et al.*, 2017) (Wallis and Norling, 2005)), Video chatbot (Banks, 2018), Pre-recorded conversations (Candello *et al.*, 2017)
Real chatbot and WoZ	2	(Jenkins *et al.*, 2007) (Portela and Granell-Canut, 2017)

Table B.3: Real chatbots' technologies

Not available	(Ciechanowski *et al.*, 2018) (Schuetzler *et al.*, 2018)
Third-party/commercial chatbot	(Brahnam and De Angeli, 2012) (Coniam, 2008) (Corti and Gillespie, 2016) (Curry and Rieser, 2018) (De Angeli, 2005) (De Angeli and Brahnam, 2006) (De Angeli *et al.*, 2001) (Fitzpatrick *et al.*, 2017) (Hill *et al.*, 2015) (Jain *et al.*, 2018b) (Lasek and Jessa, 2013) (Marino, 2014) (Miner *et al.*, 2016) (Shum *et al.*, 2018) (Valério *et al.*, 2017) (Zamora, 2017)
Self-developed architecture/ dialogue management	(Ayedoun *et al.*, 2017) (Dohsaka *et al.*, 2014) (Hayashi, 2015) (Kumar *et al.*, 2010) (Mäurer and Weihe, 2015) (Ptaszynski *et al.*, 2010) (Shum *et al.*, 2018) (Tallyn *et al.*, 2018) (Tegos *et al.*, 2016b)
Pattern-matching	AIML (Silvervarg and Jönsson, 2013), adapted ELIZA ((Kirakowski *et al.*, 2009), (Morrissey and Kirakowski, 2013))
Third-party platforms	Facebook Messenger (Araujo, 2018), Chatfuel platform (Duijst, 2017), IBM/Watson (Jain *et al.*, 2018a; Liao *et al.*, 2016, 2018), Microsoft Bot Framework (Toxtli *et al.*, 2018)

APPENDIX C

GLOSSARY

The following sections present a glossary of the register dimensions and the associated linguistic features used in the register analysis. Definitions and examples are based on Biber (1988) and Biber *et al.* (1999).

C.1 Dimension 1–Involvement

Dimension 1 is associated with the oral vs literate opposition, where high, positive scores indicate personal involvement, interactional and generalized content, while low, negative scores indicate informational density and exact informational content (Biber, 1988). The linguistic features with positive weights in this factor are:

Private verb: express intellectual states or nonobservable intellectual acts (Biber, 1988). Examples: *think, believe, discover*

That-deletion: omission of the *that*-complementizer (complementizer is a type of subordinator that begins a complement clause) (Biber *et al.*, 1999). Example (Biber *et al.*, 1999): *"I **hope** you **realized** they said a few words on there." vs. "I **hope** [that] you **realized** [that] they said a few words on there."*

Contraction: a form of phonological or orthographic reduction (Biber, 1988). Examples: *isn't* (contracted form of "is not"), *I'm* (contracted form of "I am")

Present verb: verb form that describes a state that exists or an action that is happening at the present time; or a habitual action (Biber *et al.*, 1999). Example (from (Biber *et al.*, 1999)): *"I **want** a packet of crisps."*

Second-person pronoun: pronoun that refers to the addressee (Biber *et al.*, 1999). Examples: *he, she, they*

Do as a pro-verb: the verb *do* substituting for a lexical verb or a complete predicate (Biber *et al.*, 1999). Example (from (Biber *et al.*, 1999)): *A: "He doesn't even know you." B: "He **does**!"*

Demonstrative pronoun: a demonstrative form (*this, that, these, those*) functioning as a pronoun (Biber *et al.*, 1999). Example (from (Biber *et al.*, 1999)): *"You will need **those**."*

Emphatic: a word that mark the presence (versus absence) of certainty (Biber, 1988). Examples: *really, for sure, just*

First-person pronoun: pronoun that refers to the speaker/writter (Biber *et al.*, 1999). Examples: *I, we, mine*

Pronoun it: pronoun that has non-personal references. It can also work as a dummy pronoun, which does not have a specific reference (Biber *et al.*, 1999). Examples (from (Biber *et al.*, 1999)): *"**It**'s cold."*

Be as a main verb: when the verb **be** (*am, is, was, etc.*) is used as the main verb of a clause (Biber *et al.*, 1999). Example (Biber *et al.*, 1999): *"Radio waves **are** useful."*

Causative subordination: linking word that introduce a clause that express a justification for actions or beliefs (Biber, 1988). Example: *because.*

Discourse particle: a type of insert used to maintain conversational coherence (Biber *et al.*, 1999). It signals interactively how the speaker plans to steer the dialogue (Biber, 1988). Examples: *now, anyway, okay, well*

Indefinite pronoun: a pronoun with indefinite meaning (Biber *et al.*, 1999). Examples: *anybody, everyone, one, some*

General hedge: a word that conveys imprecision or uncertainty often used to lessen the force of what is said (Biber *et al.*, 1999). Example (from (Biber *et al.*, 1999)): *"It seems **sort of** a betrayal."*

Amplifier: an extent/degree adverb or adverbial that intensifies meaning (Biber *et al.*, 1999). Examples: *"very interesting," "absolutely correct"*

WH- question: a type of interrogative clause with an initial *wh*-word (Biber *et al.*, 1999). Examples: *where, who, why*

Possibility modal: a type of auxiliary verb used to express permission, possibility, or ability (Biber, 1988). Examples: *can, may, might, could*

Coordinating conjunction–clausal connector: string independent clauses together (Biber, 1988). Example (from (Biber *et al.*, 1999)): *"He had been called her parents **and** they didn't know where she was.*

WH- clause: a type of clause introduced by a *wh*-word as complementizer (Biber *et al.*, 1999). Example (Biber *et al.*, 1999): *"She didn't ask **what my plans were.**"*

Final (stranded) preposition: a preposition that is not followed by its prepositional complement (Biber *et al.*, 1999). Example: *"what can I help you **with**?"*)

Linguistic features with negative weights in this dimension are:

Noun: refer to concrete entities or substances, and abstract qualities or states (Biber *et al.*, 1999). Examples (from (Biber *et al.*, 1999)): *pencil, bread, friendship, joy*

Preposition: a word that introduces a prepositional phrase, linking the following noun phrase to other elements of the sentence (Biber *et al.*, 1999). Examples: *at, across*

Attributive adjective: an adjective functioning as a pre-modifier before a noun (occurring before the head noun in a noun phrase) (Biber *et al.*, 1999). Examples: *"parking lot," "playground equipment"*

C.2 Dimension 2–Narrative flow

Dimension 2 distinguishes narrative from non-narrative discourses, where high/positive scores indicate narrative and reconstruction of events while low/negative scores indicate descriptive or expository discourse (Biber, 1988). All the linguistic features in this dimension have positive weight in the dimension scores. The features are the following:

Past tense verb: verb form that describes a state that existed or an action that has happened in the past; or to show hypothetical, unreal conditions. Example: *"He **walked** away."*

Third-person pronoun: pronoun that refers to other person or entities, which are neither the speaker/writer nor the addressee (Biber *et al.*, 1999). Examples: *"they," "theirs"*

Perfect aspect verb: a verb construction that describes events or states taking place in the past, but linked to a subsequent time, especially the present (Biber *et al.*, 1999). Example: *have/has/had visited*

Public verb: involve actions that can be observed publicly and they are commonly used to introduce indirect statements (Biber, 1988). Example *say, assert, claim, promise*

Present participial clause: participle clause is a non-finite clause with a participle as the main verb (Biber *et al.*, 1999). Present participial clauses have an *-ing* participial clause where the subject is shared with the main clause. Example: *"**visiting the city**, you see..."*

C.3 Dimension 3–Contextual reference

Dimension 3 is associated with the explicit vs situation-dependent reference opposition, where high/positive scores indicate a discourse that presents highly explicit and elaborated, endophoric reference, where utterances use precise references to previous ones, and common ground among interlocutors is not assumed (Biber, 1988). For example, the sentence "*The Grand Canyon, which is one of the seven natural wonders, ...*," implies that the tourist does not necessarily have previous knowledge about the Grand Canyon. Low/negative scores indicate exophoric, situation-dependent reference, which implies common ground among interlocutors (Biber, 1988). For example, the sentence "*The Grand Canyon is a must-see*" implies that the interlocutors share knowledge about the Grand Canyon. The linguistic features with positive weight are the following:

WH relative clause on object position: *wh-* word that introduces a relative clause–a type of finite dependent clause used to modify a noun phrase–on the object position (Biber *et al.*, 1999). Example (Biber, 1988): *"the man **who** Sally likes"*

Pied-piping construction: *wh-* word that introduces a relative clause with prepositional fronting. Example (from (Biber, 1988)): *"the manner in which he was told"*

WH relative clause on subject position: *wh-* word that introduces a relative clause on the subject position. Example (from (Biber, 1988)): *"the man **who** likes popcorn"*

Coordinating conjunction – phrasal connector: word that unites ideas and integrates information (Biber, 1988). Example: *"John **and** Peter are roommates."*

Nominalization: use of words that are not nouns as nouns (Biber, 1988). Example: *"brewery," "equipment"*

The features with negative weights are:

Adverb: one of the four lexical word classes in English–its most common uses are as an adverbial an as a modifier of an adjective (Biber *et al.*, 1999). Examples: *about, probably*

Time adverbial: adverb that express position in time, frequency, duration, and relationship (Biber *et al.*, 1999). Examples: *afterwards, again*

Place adverbial: adverb that express distance, direction, or position (Biber *et al.*, 1999). Examples: *abroad, far*

C.4 Dimension 4–Persuasiveness

Dimension 4 focuses on the overt expression of persuasion, where positive scores indicate that persuasion is overtly marked, either for expressing the speaker's point of view or assessing advisability, while negative scores indicate discourses with no opinions or arguments (Biber, 1988). The linguistic features associated with this dimension all have positive weight and are listed below:

Infinitive: the verb in its basic form. Example: *to build, to go*

Prediction modal: a type of auxiliary verb used to express volition or prediction (Biber, 1988). Examples: *will, would, shall*

Suasive verb: verb that implies an intention to bring about some change in the future (Biber, 1988). Examples: *command, insist, propose*

Conditional subordination: linking word that introduce a clause that express a condition for actions or beliefs (Biber, 1988). Examples: *if, unless*

Necessity modal: a type of auxiliary verb used to express obligation or necessity (Biber, 1988). Examples: *ought, should, must*

Split auxiliary: occurs when adverbs are placed between auxiliaries and their main verb (Biber, 1988). Example: *"They are* ***easily*** *deceived."*

C.5 Dimension 5–Formality

Dimension 5 distinguishes abstract from non-abstract information, where high/positive scores indicate informational discourse that is technical and formal while low/negative scores indicate non-technical, informal discourse (Biber, 1988). The linguistic features in this dimension have positive weight and are listed below:

Conjuncts: a connector that explicitly mark logical relations between clauses (Biber, 1988). Examples: *furthermore, therefore, however.*

Agentless passive: passive voice when the agent does not have a salient role in the discourse (Biber, 1988). Example: *"The city* ***was founded*** *in 1881."*

Past participial: e.g., *"have seen, " "has visited"*

By-passive: passive voice when the patient is more closely related to the discourse theme than the patient (Biber, 1988). Example: *"The dam was built by Native Americans."*

Other adverbial subordinators: linking word that introduce a clause that has multiple functions (other than conditional or causal) (Biber, 1988). Examples: *since, while, such that, as long as*

Predicative adjective: adjective that occurs inthe subject predicative position, following a copular verb (Biber *et al.*, 1999). Example (Biber *et al.*, 1999): "He seems **tired**."

APPENDIX D

REGISTER CHARACTERIZATION: STATISTICAL RESULTS

Table D.1 presents the ANOVA results for all the individual features comparison between $DailyDialog$ and both original (left side of the table) and modified corpora (right side of the table), including the non-significant features. The left side of the table presents the estimates and standard deviations for each independent variable ($DailyDialog$, $TA1$, $TA2$, $TA3$), the F-values, and the corresponding p-values. The right side of the table shows the estimates, standard deviations, F-values, and p-values for the three experimental groups ($TA1_{mod}$, $TA2_{mod}$, $TA3_{mod}$) after modifications being performed ($DailyDialog$ column was omitted in the right side to avoid repetition). All the statistics are calculated with $df = 3, 1139$.

Table D.1: ANOVA results for individual features comparison

Features	$D.Dialog$	Estimates±Std.Dev. (FLG) TA1	TA2	TA3	F	P	Estimates±Std.Dev. (FLG_{mod}) $TA1_{mod}$	$TA2_{mod}$	$TA3_{mod}$	F	P
Dimension 1: personal involvement											
Private verb	14.70 ± 0.75	6.71 ± 3.46	5.87 ± 3.49	5.71 ± 3.32	5.56	0.00	14.50 ± 3.53	16.82 ± 3.49	14.72 ± 3.23	0.11	0.95
That-deletion	4.79 ± 0.39	0.65 ± 1.79	0.40 ± 1.81	0.56 ± 1.72	5.04	0.00	4.12 ± 1.92	4.27 ± 1.90	7.34 ± 1.76	0.75	0.52
Contraction	26.68 ± 1.09	11.66 ± 5.05	11.88 ± 5.10	1.55 ± 4.84	13.00	0.00	30.74 ± 5.13	26.53 ± 5.07	22.35 ± 4.70	0.49	0.69
Present verb	159.2 ± 1.7	126.0 ± 8.0	119.1 ± 7.7	120.8 ± 7.3	21.63	0.00	157.1 ± 7.6	153.2 ± 7.6	151.4 ± 7.0	0.58	0.63
2nd person pronoun	81.76 ± 1.52	41.38 ± 7.02	23.36 ± 7.09	44.52 ± 6.74	38.53	0.00	68.49 ± 7.10	59.36 ± 7.02	66.44 ± 6.50	5.60	0.00
Do as a pro-verb	5.72 ± 0.47	2.06 ± 2.18	1.90 ± 2.21	2.13 ± 2.10	2.55	0.05	2.85 ± 2.17	2.76 ± 2.15	1.85 ± 1.99	2.16	0.09
Dem. pronoun	6.22 ± 0.49	1.28 ± 2.27	4.39 ± 2.30	2.76 ± 2.18	2.34	0.07	3.63 ± 2.28	4.89 ± 2.25	4.08 ± 2.09	0.79	0.50
Emphatic	7.23 ± 0.55	9.29 ± 2.52	5.29 ± 2.54	5.10 ± 2.42	0.66	0.58	9.34 ± 2.55	7.57 ± 2.52	6.69 ± 2.33	0.25	0.86
1st person pronoun	55.13 ± 1.29	18.18 ± 5.94	16.77 ± 6.01	11.01 ± 5.71	40.59	0.00	37.55 ± 6.02	39.55 ± 5.96	41.94 ± 5.52	6.12	0.00
Pronoun it	17.45 ± 0.92	11.94 ± 4.24	15.05 ± 4.28	9.55 ± 4.07	1.72	0.16	12.44 ± 4.28	16.36 ± 4.23	12.27 ± 3.92	0.95	0.42
Be as a main verb	4.35 ± 0.39	3.59 ± 1.80	2.43 ± 1.82	6.52 ± 1.73	0.95	0.42	1.68 ± 1.77	1.79 ± 1.75	3.96 ± 1.62	1.35	0.26
Causative subord.	0.20 ± 0.06	1.19 ± 0.29	0.00 ± 0.29	0.33 ± 0.28	3.93	0.01	0.39 ± 0.28	0.00 ± 0.28	0.32 ± 0.26	0.39	0.76
Discourse particle	6.29 ± 0.49	3.70 ± 2.25	0.44 ± 2.27	0.00 ± 2.16	4.85	0.00	4.26 ± 2.27	3.46 ± 2.25	3.50 ± 2.08	1.21	0.30
Indefinite pronoun	7.57 ± 0.53	4.79 ± 2.43	1.60 ± 2.46	2.23 ± 2.33	3.68	0.01	5.08 ± 2.47	3.46 ± 2.45	3.36 ± 2.27	2.13	0.09
General hedge	0.55 ± 0.13	0.00 ± 0.60	0.20 ± 0.61	0.00 ± 0.58	0.60	0.61	0.00 ± 0.60	0.21 ± 0.59	0.00 ± 0.55	0.62	0.60
Amplifier	2.15 ± 0.25	3.05 ± 1.16	2.96 ± 1.17	0.24 ± 1.11	1.35	0.26	3.56 ± 1.18	2.35 ± 1.16	2.63 ± 1.08	0.51	0.68
WH- question	5.83 ± 0.48	4.46 ± 2.21	6.25 ± 2.23	0.58 ± 2.12	2.06	0.10	4.94 ± 2.23	5.33 ± 2.20	5.72 ± 2.04	0.06	0.98
Possibility modal	17.72 ± 0.75	15.77 ± 3.45	15.97 ± 3.49	12.08 ± 3.31	1.05	0.37	19.40 ± 3.49	19.82 ± 3.45	19.75 ± 3.19	0.29	0.83
Coord conj (clause)	6.55 ± 0.42	18.97 ± 1.93	8.75 ± 1.95	3.07 ± 1.85	14.99	0.00	6.43 ± 1.93	6.67 ± 1.90	6.23 ± 1.76	0.01	1.00
WH- clause	0.46 ± 0.14	0.91 ± 0.66	0.00 ± 0.66	0.00 ± 0.63	0.48	0.69	1.09 ± 0.67	0.00 ± 0.66	0.00 ± 0.61	0.65	0.59
Final preposition	1.50 ± 0.21	4.87 ± 0.98	0.27 ± 0.99	2.01 ± 0.94	4.49	0.00	1.44 ± 0.93	0.25 ± 0.92	0.83 ± 0.85	0.76	0.52
Nouns	225.6 ± 2.5	277.6 ± 11.6	328.0 ± 11.8	302.1 ± 11.2	41.93	0.00	258.1 ± 11.6	280.5 ± 11.4	275.6 ± 10.6	15.48	0.00
Prepositions	70.60 ± 1.44	100.1 ± 6.64	112.1 ± 6.71	91.30 ± 6.37	20.15	0.00	80.04 ± 6.57	77.15 ± 6.50	69.86 ± 6.02	0.96	0.41
Attrib. adjective	19.49 ± 0.91	43.67 ± 4.18	45.06 ± 4.22	55.09 ± 4.01	43.63	0.00	26.48 ± 4.04	22.96 ± 4.00	26.28 ± 3.70	2.08	0.10
Dimension 2: narrative flow											
Past tense verb	8.57 ± 0.58	9.29 ± 2.69	7.68 ± 2.72	7.01 ± 2.58	0.17	0.91	8.51 ± 2.72	7.76 ± 2.70	8.65 ± 2.50	0.03	0.99
3rd person pronoun	4.00 ± 0.45	12.73 ± 2.08	9.47 ± 2.10	9.84 ± 2.00	9.67	0.00	5.03 ± 1.99	5.07 ± 1.96	3.81 ± 1.82	0.18	0.91
Perfect aspect verb	3.48 ± 0.34	2.10 ± 1.55	1.72 ± 1.57	0.50 ± 1.49	1.79	0.15	2.64 ± 1.57	1.81 ± 1.55	2.03 ± 1.44	0.72	0.54
Public verb	2.71 ± 0.28	0.00 ± 1.31	2.22 ± 1.32	0.95 ± 1.25	1.93	0.12	2.73 ± 1.34	3.77 ± 1.33	3.24 ± 1.23	0.25	0.86
Dimension 3: contextual reference											
WH-rel. cl. (object)	0.12 ± 0.07	0.19 ± 0.34	0.00 ± 0.35	0.00 ± 0.33	0.09	0.96	0.22 ± 0.35	0.29 ± 0.35	0.00 ± 0.32	0.15	0.93
WH-rel. cl. (subject)	0.21 ± 0.07	2.19 ± 0.33	0.98 ± 0.33	0.55 ± 0.31	13.28	0.00	0.80 ± 0.28	0.00 ± 0.28	0.40 ± 0.26	1.75	0.16
WH-rel. pied piping	0.19 ± 0.09	0.00 ± 0.42	0.00 ± 0.42	0.26 ± 0.40	0.13	0.94	0.00 ± 0.42	0.00 ± 0.41	0.24 ± 0.38	0.13	0.94
Coord conj (phrasal)	1.11 ± 0.22	2.36 ± 0.99	0.17 ± 1.00	0.24 ± 0.95	1.09	0.35	1.73 ± 0.99	0.17 ± 0.98	0.95 ± 0.91	0.44	0.72
Nominalization	20.98 ± 0.90	27.41 ± 4.16	35.41 ± 4.21	23.30 ± 3.99	4.41	0.00	22.94 ± 4.10	26.19 ± 4.06	21.26 ± 3.76	0.58	0.63
Time adverbial	7.47 ± 0.49	3.50 ± 2.24	1.57 ± 2.26	2.01 ± 2.15	4.80	0.00	4.52 ± 2.30	4.70 ± 2.27	9.12 ± 2.10	1.22	0.30
Place adverbial	17.92 ± 0.84	15.73 ± 3.88	11.16 ± 3.92	13.34 ± 3.72	1.43	0.23	18.35 ± 3.89	13.86 ± 3.85	14.44 ± 3.57	0.64	0.59
Adverb	36.33 ± 1.15	46.80 ± 5.30	35.23 ± 5.36	23.67 ± 5.09	3.37	0.02	40.75 ± 5.29	40.88 ± 5.24	33.33 ± 4.85	0.60	0.62
Dimension 4: persuasiveness											
Infinitive	9.23 ± 0.50	6.85 ± 2.32	5.18 ± 2.35	8.77 ± 2.23	1.24	0.30	6.33 ± 2.32	9.25 ± 2.29	7.95 ± 2.12	0.60	0.62
Prediction modal	21.84 ± 0.81	10.45 ± 3.73	8.41 ± 3.77	13.97 ± 3.58	7.87	0.00	22.85 ± 3.79	18.05 ± 3.75	18.99 ± 3.47	0.55	0.65
Suasive verb	0.56 ± 0.14	4.51 ± 0.65	1.10 ± 0.65	3.00 ± 0.62	16.17	0.00	1.06 ± 0.56	0.80 ± 0.55	0.32 ± 0.51	0.39	0.76
Conditional subord.	2.47 ± 0.27	4.00 ± 1.24	6.48 ± 1.25	5.94 ± 1.19	5.98	0.00	1.73 ± 1.21	2.86 ± 1.20	2.77 ± 1.11	0.18	0.91
Necessity modal	4.18 ± 0.38	1.33 ± 1.74	1.44 ± 1.75	2.05 ± 1.67	1.99	0.11	3.27 ± 1.76	4.53 ± 1.74	4.32 ± 1.62	0.10	0.96
Split auxiliary	1.36 ± 0.20	5.49 ± 0.93	1.49 ± 0.94	1.40 ± 0.89	6.30	0.00	1.89 ± 0.90	1.41 ± 0.89	1.87 ± 0.83	0.23	0.88
Dimension 5: formality											
Conjuncts	5.22 ± 0.40	2.48 ± 1.83	1.16 ± 1.85	0.74 ± 1.75	3.99	0.01	3.05 ± 1.86	4.56 ± 1.84	4.36 ± 1.71	0.52	0.67
Agentless passive	3.05 ± 0.30	5.91 ± 1.40	4.32 ± 1.41	5.26 ± 1.34	2.26	0.08	4.76 ± 1.37	3.54 ± 1.35	2.99 ± 1.25	0.53	0.66
By-passive	0.06 ± 0.05	0.33 ± 0.21	0.28 ± 0.21	0.00 ± 0.20	0.89	0.44	0.24 ± 0.20	0.00 ± 0.20	0.00 ± 0.18	0.32	0.81
Past participal	0.56 ± 0.13	0.35 ± 0.60	1.03 ± 0.60	0.86 ± 0.57	0.33	0.81	0.83 ± 0.59	0.48 ± 0.58	0.27 ± 0.54	0.17	0.92
Other adv. subord.	3.40 ± 0.31	2.41 ± 1.44	1.91 ± 1.45	2.21 ± 1.38	0.67	0.57	1.77 ± 1.43	1.77 ± 1.41	2.08 ± 1.31	1.06	0.36
Predic. adjective	7.21 ± 0.56	13.33 ± 2.57	6.43 ± 2.60	9.87 ± 2.46	2.17	0.09	9.88 ± 2.50	4.97 ± 2.48	7.00 ± 2.29	0.66	0.58

APPENDIX E

TEXT MODIFICATION

This appendix brings additional details about the text modification of FLG corpus. The text modification process resulted in a corpus called FLG_{mod}, which preserves the content of FLG while mimics the linguistic form of $DailyDialog$.

E.1 Validation of modifications: statistical results

This appendix presents the statistical results of the validation of modifications. Figure E.1 shows the content preservation distribution for every question.

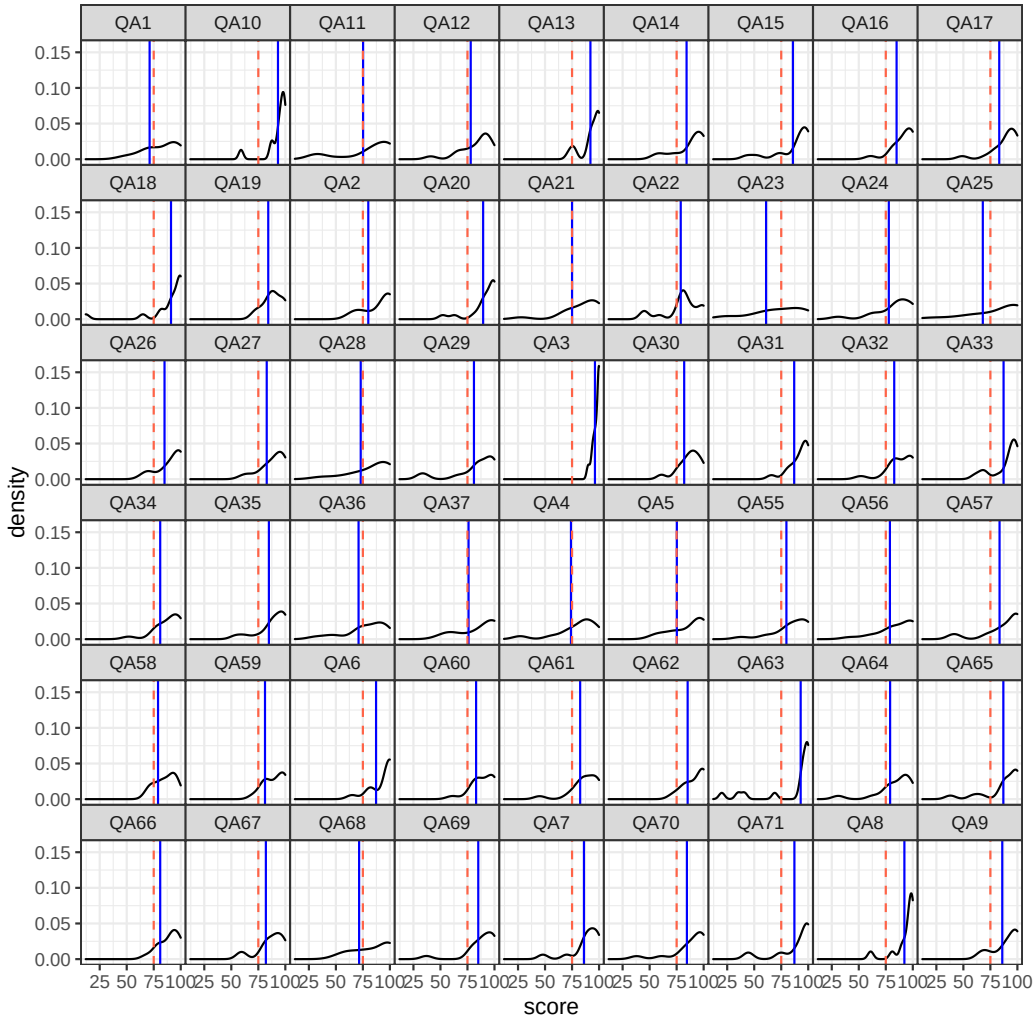


Figure E.1: Content preservation distribution for each question. The dashed line marks the third quantile (75) and the solid line marks the model coefficient for that question.

We fitted a model with the score as response variable and no fixed effect, we only care about the intercept. To support $H_0 : \mu = 75$, the intercept must be above 75, and 75 must not be within the confidence interval. The model also has two random effects: the individual questions and the participants who rated the content preservation for each question. The estimated mean for the intercept is 86.77 and the estimated error is 1.38 (df $= 99.04$, t $= 62.65$, p-value < 0.001). Table E.1 presents the outcomes for the random effects.

Table E.1: Random effects for linear mixed model fit, with a total of 880 observations and two groups: Participants $= 88$ and Questions $= 54$

Groups Name	**Variance**	**Std.Dev.**
Participants (Intercept)	143.67	11.99
Questions (Intercept)	9.11	3.02
Residual	102.49	10.12

E.2 Selected Question-Answer Pairs

As described in the paper, we ordered the question-answer pairs using the Levenshtein distance (Kessler, 1995; Wikibooks, 2020) between the pairs of original and modified answers, and selected the 54 question-answer pairs with the highest values of distance. Table E.2 presents the selected question-answer pairs as well as the modification score, which corresponds to the Levenshtein distance value.

Table E.2: Original and Modified sentences evaluated in the user study 1. The Levenshtein distance (Kessler, 1995) between the answers for each question was calculated using a Python algorithm, available in Wikibooks (Wikibooks, 2020).

Begin of Table E.2

ID	Question	FLG	FLG_{mod}	**Score**
QA1	Hello, we're wondering what are some notable buildings along route 66 to visit while going through the city?	There is a self-guided Rte 66 tour that starts in the Historic Train Center on 1 E. Rte. 66. In the visitor's center there is a self-guided map that shows the original alignment through the redeveloped Southside Historic District and passes by classic drive-in motels and Flagstaff landmarks of old. Let me know if you have further questions.	We have a self-guided Rte 66 tour that starts in the Train Center on 1 E. Rte. 66. The visitor center has a self-guided map that shows you how the redeveloped Southside Historic District aligns through the original and passes by classic motels and landmarks of old. Is there anything else I can do for you today?	136

Continuation of Table E.2

ID	Question	FLG	FLG_{mod}	Score
QA2	Where can I park for more than 15 minutes?	There is two-hour free parking on Rte 66 on the side streets east of San Francisco on the Southside and there is free parking in the neighborhoods west of City Hall. Paid parking downtown is $1 per hour, pay with a credit card at the kiosks located throughout downtown by entering your license plate #. The parking system tracks your license plate and knows that you paid (no receipt given).	Okay, you park free for two hours on Rte 66 on the streets east of San Francisco on the Southside and the neighborhoods west of City Hall have free parking, you park and walk to downtown. Paid parking downtown is $1/hour, you pay using your credit card at the kiosks located throughout downtown when you enter your license plate #. The parking system tracks your license plate and immediately knows that you paid (no receipt given).	131
QA3	What is parking like in downtown area?	Parking is $1 an hour most places you can park. There are kiosks on the street where you can pay by credit card by entering your license plate #. there is no receipt and add more time if needed at any kiosk downtown.	Most places you can park are $1 an hour. You'll find kiosks on the street where you pay using your credit card when you enter your license plate #. You don't need a receipt and you can use any kiosk downtown to add more time as needed.	124
QA4	Hello, is the Upper Antelope Canyon ok for one that is slightly claustrophobic?	Hello. Antelope Canyon is a narrow slot canyon so it isn't recommended for anyone with claustrophobia. A good alternative would be to do to Bryce Canyon in Utah which has a similar feel without being so narrow or the Sedona area. They both have beautiful red rock formations.	Hello. Antelope Canyon is a narrow slot canyon so I would not recommend. I think you could do the Bryce Canyon in Utah as it has a similar feel and this one isn't so narrow or visit Sedona. Both have rock formations I find beautiful.	121
QA5	What are the best toddler-friendly things to do in Flagstaff?	There is good playground equipment at Thorpe Park and Bushmaster Park. A short easy hike at Buffalo Park also has activity centers. There are kids activities at Heritage Square on Fridays and Saturdays starting at 4:30 (movies on Saturdays after the sun sets)	I would suggest the Thorpe Park and Bushmaster Park's playgrounds. You'd like to go on a hike at Buffalo Park where you'll also find activity centers. We have activities that kids would enjoy at Heritage Square on Fridays and Saturdays starting at 4:30 (we have movies on Saturdays when the sun sets)	117

Continuation of Table E.2

ID	Question	FLG	FLG_{mod}	Score
QA6	Is there a AAA office near here?	No there is not a AAA office in Flagstaff, the closest one is located in Prescott which is about an hour and a half away. If you are looking for maps of Arizona, Flagstaff, or our surrounding states, the Visitor Center can provide these to you for no charge.	No, we don't have a AAA office in Flagstaff, and you'll find the closest one in Prescott, which is roughly an hour and a half away. The Visitor Center can give you maps of Arizona, Flagstaff, or our surrounding states for no charge.	108
QA7	Which one is the best way to visit the Grand Canyon, take a tour or rent a car?	A tour will obviously limit your time but you'll learn a lot more about the Grand Canyon and its history, plus you have the bonus of not needing to drive yourself, so you'll be able to focus more on your surroundings. If you rent a car, you'll be able to set your own schedule and if there is a particular location you want to visit, you won't be limited by the tour.	A tour will obviously limit your time but you'll learn more about the Grand Canyon and its history, and you don't need to drive yourself, so you'll focus more on your surroundings. When you rent a car, you'll set your own schedule to reach places you want to visit, the tour would limit you.	108
QA8	Hi, is there a AAA office near here?	There is not a AAA office in Flagstaff (the closest is Prescott). There are a lot of maps and information within the visitor's center that would be able to help with trip planning.	We don't have an AAA office here (the closest one you'll find is Prescott). The visitor center has maps and information that would be able to help you plan your trip.	102
QA9	We are looking for a place to have fun and have a beer tonight	I would recommend getting the Ale Trail sheet from the Visitor's Center. It shows the 9 craft breweries most in walking distance of downtown. If you get stamps from all (no purchase necessary), you get a free souvenir glass.	I'd suggest that you get the Ale Trail sheet we have at the Visitor Center that shows you the 9 breweries most nearby downtown. When you get stamps from all (you don't need to purchase anything), you get a souvenir glass.	101
QA10	Can I leave my car parked at the visitor center?	You cannot leave it in 15-minute parking for an extended period of time. On the Amtrak side of the building, there is a paid parking lot. $1 per hour, the kiosk is by the green dumpster. Pay by credit card and enter your license plate #	You cannot leave it in 15-minute parking for more than that. However, the Amtrak side of the building has a paid parking lot you could use. You'll pay $1/hour the kiosk is by the dumpster that is green. You pay using your credit card and you have to enter your license plate #.	99

Continuation of Table E.2

ID	Question	FLG	FLG_{mod}	Score
QA11	Hi, what is the best breakfast in Flagstaff?	Hello. If you want a Southwest-style breakfast (on the spicy side), go to Martanne's (112 Historic Rte 66). For frou-frou breakfast with lots of options, Toasted Owl is good (12 S Mikes Pike St). And tradition breakfast with big portions, go to Mike & Ronda's The Place (21 S. Milton).	Hello. If you want a Southwest-style breakfast (on the spicy side), I guess you should visit Martanne's (112 Historic Rte 66). I think ToastedOwl is good to have a frou-frou breakfast with many options (12 S Mikes Pike St). And tradition breakfast with big portions, I'd prefer visiting Mike & Ronda's The Place (21 S. Milton).	98
QA12	Do you know if it is safe to drive through the Bearizona with a rental car?	Yes, it is, but if you are concerned about the car, they do have a Wild Ride Bus Tour option scheduled a couple times a day. What's really nice about the bus tour is that they will talk to you about the animals that you see in the park.	Yes, it is. If that concerns you, you should ride the buses that are scheduled a couple times a day. What I like when you take the bus is that you will hear about the animals that you see.	98
QA13	Hello, I have 1 hour only. What should I do?	Hello. With only one hour I would recommend visiting downtown Flagstaff. You can easily spend an hour wandering around. The majority of the downtown has kept its original historic buildings and there are also dozens of beautiful murals by local artists to see and most everything downtown is locally owned and operated.	Hello. With only one hour I'd visit downtown. You can spend an hour wandering around. The majority of the downtown has kept its historic buildings and we also have murals by local artists you should see. Most everything we have in downtown is locally owned and operated.	92
QA14	Any recommendations for post-hike places to eat or grab a beer?	To do both, I would recommend Historic Brewing on San Francisco or Dark Sky Brewing on Beaver St. For a large menu selection with beer, Lumberyard is another good choice. Check out the Ale Trail map in the visitor's center.	To do both, I'd suggest Historic Brewing on San Francisco or Dark Sky Brewing on Beaver St. To find a menu with beer, I believe Lumberyard is another choice. I'd suggest you visit the visitor center and grab the Ale Trail map.	92
QA15	We'd enjoy a nice long hike – any difficulty is fine. Preferably something with some good views! Do you have any specific suggestion?	Humphreys will be the most challenging with a rewarding view followed by Mt. Elden. Specific elevation gain, mileage etc. is within the visitor's center. Make sure to check weather prior as we frequently have thunderstorms and lightning this time of year.	I think Humphreys will be the most challenging with a rewarding view Mt. Elden comes next. The visitor center offers you details on elevation, mileage etc. Make sure you check weather prior as we frequently have thunderstorms and lightning this season.	92

Continuation of Table E.2

ID	Question	FLG	FLG_{mod}	Score
QA16	Where are the nearest used bookstores to the visitor center?	There is a used bookstore downtown just half a block from the Visitor Center. It is called Starrlight Books and it is just north on Leroux Street. There is also a local independent bookstore called Brightside Bookshop on N San Francisco St.	We have a bookstore that sells used books downtown just half a block from the Visitor Center. The name is Starrlight Books, it's just north on Leroux Street. You can visit Brightside Bookshop, a local bookstore on N San Francisco St.	92
QA17	What are the best toddler-friendly things to do in Flagstaff?	There are some great National Monuments that are toddler-friendly called Sunset Crater and Wupatki. Sunset Crater is an extinct volcano and Wuptaki is ancient ruins with a blowhole that is very popular with children. Pioneer Museum is another good attraction for toddlers as they have a room called playthings of the past and a historic train caboose.	We have some National Monuments that are toddler-friendly: Sunset Crater is a volcano that is extinct and Wuptaki is ruins with a blowhole that's very popular with children. Pioneer Museum is another attraction that toddlers enjoy as there's a room called playthings of the past and a train caboose.	90
QA18	What is parking like in downtown area?	There is a company called American Valet that runs a number of the parking lots downtown but they charge $2-3/hour. The Flagstaff Visitor Center has free 30-minute parking available 8-5 PM and it is free to park in that lot after hours.	You can try the American Valet, a company that runs the parking lots downtown but you'll pay $2-3/hour. Our Visitor Center has free 30-minute parking available 8-5 PM. You are free to park there after those hours.	90
QA19	Do the breweries do tours? Do I have to make a reservation for a tour?	I would recommend giving them a call to see. It depends on if they have someone available. I just did a tour of Mother Road's new brewery facilities last week and it was fascinating because they have grown significantly since they moved their brewing to the second location on Butler Ave.	I'd suggest you call them to see. It depends on whether they have someone available. I just did a tour of Mother Road's new brewery. It fascinates me, they have grown since they moved the brewing to the second location on Butler Ave.	90
QA20	Can you advise on a nearby camping ground. Dog-friendly and with drinking water. Tents only. Thanks	Many of the tent-only campgrounds just opened today in the Coconino National Forest (closest). Not all of them are opened just yet. Give Coconino National Forest a call at (928) 527-3600 to ask which campgrounds and if pets are allowed. Private campgrounds in the Flagstaff area will have a mix of RV and tent sites. Hope that helps.	Alright I know that many tent-only campgrounds just opened today in the Coconino National Forest (the closest one). Not all of them are opened just yet. I guess I'd call Coconino National Forest (928) 527-3600 and ask which campgrounds are open and whether pets are allowed. Private campgrounds in Flagstaff will have a mix of RV and tent sites. I hope that helps you.	86

Continuation of Table E.2

ID	Question	FLG	FLG_{mod}	Score
QA21	Should I go to the Grand Canyon or Sedona? I have National Parks pass	Okay perfect! In that case, I would definitely recommend going to the Grand Canyon. It is included in your national park pass and you can easily spend a full day there. Once you park you can take the free shuttle buses around the canyon and are able to access most of the park that way.	Okay perfect! I'd visit the Grand Canyon. It's included in your national park pass, you can spend a full day there. Once you park, you can take the shuttle around the canyon. It's free and you'd visit most of the park that way.	86
QA22	Hi, does the tourist center have a city map?	Hello. Yes, the Flagstaff Visitor Center has two city maps. There is the Discovery Map which is more geared toward visitors and highlights the important attractions and there is also a map produced by the Chamber of Commerce that has all the streets on it that is more for relocation.	Hello. Yes, our Visitor Center has two maps you can pick up. The Discovery Map is more geared toward visitors and it highlights the attractions. You can check out the map produced by the Chamber of Commerce that shows you all the streets which is more for relocation.	85
QA23	One thing that was mentioned was the underground tunnels dug by Chinese residents and later on the steam power plant, are there any areas to see these tunnels?	There are not any areas that are open to the general public but the Pioneer Museum has a map of the tunnels. You could ask individual business owners if you could see them.	Unfortunately, I believe we don't have any areas that are open to the public, but the Pioneer Museum has a map that delimits the tunnels. You could ask property owners to see them.	81
QA24	Are the shuttles to the nearby attractions expensive?	Arizona Shuttle is $32 each way plus park entry for Grand Canyon, and $45 each way for Sedona. Sedona Inspire is $60 each way. Not sure what the pricing is for National Park Express	Arizona Shuttle is $32 each way for Grand Canyon, you must also pay the park entry, and Sedona is $45 each way. Sedona Inspire is $60 each way. I'm not sure how much you'd pay for the National Park Express	81
QA25	Are there any vegetarian-friendly restaurants in or near downtown?	Red Curry Kitchen is a good vegan restaurant 10 N San Francisco St as well as Whyld Ass Restaurant 121 E. Birch Ave (just east of San Francisco). A lot of restaurants offer vegetarian options including the locally-made Tepa vegetarian burgers.	Well, Red Curry Kitchen is vegan 10 N San Francisco St, so does Whyld Ass 121 E. Birch Ave (which is just east of San Francisco). Many restaurants would offer you vegetarian options, these include the locally-made Tepa vegetarian burgers.	80

Continuation of Table E.2

ID	Question	FLG	FLG_{mod}	Score
QA26	Are there any thrift shops near here?	There are a few thrift shops in Flagstaff but the closest one to the Visitor Center is right across the train tracks on Beaver St called The Garden Thrift.	We have a few thrift shops, the closest one to the Visitor Center is the Garden Thrift. You'll locate it right across the train tracks on Beaver St.	80
QA27	Hi, where is the best view spot of the Grand Canyon?	Hello. The Grand Canyon has beautiful views everywhere in the park. Some of my favorite viewpoints are at the Desertview Watchtower on the East side of the canyon and Mohave Point. There are shuttle buses that can take you to most of the viewpoints or you are able to hike along the rim as well.	Hello. The Grand Canyon has views you'd enjoy everywhere. The viewpoints I prefer are the Desertview Watchtower on the East side of the canyon and Mohave Point. There are shuttle buses that include most viewpoints or you might hike along the rim as well.	79
QA28	What is the best jeep tour of the vortex in Sedona? Do I need to book in advance?	I've heard good things about the Pink Jeep tours but they do not have a specific Vortex package. Red Rock Western Jeep Tours is another option. You need to book a day or so in advance.	Well, I've heard Pink Jeep tours is good but I believe they do not have a specific Vortex package. Red Rock Western Jeep Tours is an alternative, you may want to check. You need to book a day or so beforehand.	76
QA29	I am a senior with disability for up-down steps. Do you think I can still access the Antelope Canyon?	Ahh that will be difficult. Upper Antelope Canyon has a number of ladders as part of the trail. You can look into Lower Antelope Canyon as an alternative. It isn't as narrow as the Upper Canyon and has fewer restrictions on visitation.	Oh that will be hard. Upper Antelope Canyon trail has ladders you need to climb. You can look into Lower Antelope Canyon. It isn't as narrow as the Upper Canyon and it has fewer restrictions.	76
QA30	How much does it cost the tour at the Lowell Observatory?	Lowell is $15 but you can also return at night to do stargazing with telescopes for same one day admission. There are $2 off coupons inside the visitor's center behind the desk.	Lowell is $15. You can return later to do stargazing with telescopes for same admission. The visitor center has $2 off coupons you can get.	74
QA31	Hi, my family and I like adventure sports. Is there any site in Flagstaff for this?	Hello. What kind of adventure sports are you interested in? There is the Flagstaff Extreme Adventure Course just south of town at Fort Tuthill. There are two different courses available one that is purely ziplining and another that involves a lot of climbing and obstacles through the trees as well as some ziplines.	Hello. What adventures attract you? We have the Flagstaff Extreme Adventure just south of town at Fort Tuthill. There are two courses available, one is ziplining and another that involves climbing and obstacles through the trees as well as some ziplines.	74

Continuation of Table E.2

ID	Question	FLG	FLG_{mod}	Score
QA32	I'm looking for some ideas of things to do. We enjoy being outside, mountain biking, hiking, etc.	There are 50 miles of trails within Flagstaff - get a Flagstaff Urban Trails Map within the visitor's center. There are also good mountain biking trails near Ft. Tuthill County Park and on some of the Forest Service trails.	We have 50 miles of trails within Flagstaff - visit the visitor center, you can get an Urban Trails Map. We also have mountain biking trails near Ft. Tuthill County Park and on the Forest Service trails.	74
QA33	I am afraid of heights. Should I go to Grand Canyon anyways?	There are some sections where the Canyon juts out further that might not be for you but there are many viewing points with good railing (like Mather Point) that would make it well your while and you will feel safe enough.	We have some sections where the Canyon juts out further that you might not like but we have many points that have a railing and you'd see the canyon (like Mather Point) that would make it well your while and you will feel safe enough.	74
QA34	Should I go to the Grand Canyon if it is likely to rain?	Yeah It is still worth a visit. Typically it doesn't rain all day, it will just be on and off. When the clouds are in the sky after a thunderstorm they make the best sunsets if you are able to stay at the Grand Canyon until Sunset.	Certainly, I would still visit. Typically it doesn't rain all day, it will just be on and off and the clouds after a thunderstorm make the best sunsets in my opinion if you're able to stay there until Sunset.	73
QA35	How much walking to see the Wupatki?	The main ruin does require a bit of walking but if walking is an issue there are also lookout points where you can sit on a bench and see the ruins. The trail makes a loop around the main ruin and it is paved and an easy grade.	The main ruin does require walking. Being you not able to walk, we also have lookout points where you can sit on a bench and see the ruins. The trail makes a loop around the main ruin, it's paved and I would rate the hike as easy.	73
QA36	What time of the day is the best for hiking?	This time of year you can do these hikes in the middle of the day and it's still lovely. It can get pretty cold in the mornings but that can be nice for a good midday view of the surrounding area.	This season you can do these hikes in the middle of the day. It's cool. It can get pretty cold in the mornings but I believe you might like to enjoy a midday view.	72
QA37	Is it worth to visit Sedona?	There are some very pretty red rocks there and has different scenery than Flagstaff. If you like hiking there are some good ones there but also just nice to go to some viewpoints. The shopping area can be a bit too much in my opinion.	Well, we have some very pretty red rocks there and it has a different view than Flagstaff. Do you like hiking? We have some ones I would suggest there, I think it's also just nice to visit some viewpoints though. The shopping area can be too much in my opinion.	72

Continuation of Table E.2

ID	Question	FLG	FLG_{mod}	Score
QA38	We like adventure sports. Is there any site in Flagstaff for this?	Flagstaff Extreme Adventure Course would be a good option or there are some rock climbing area such as the Pit and Priest Draw or many mountain biking trail such as Shultz Creek or the Upper and Lower Oldham trails.	I would suggest you visit Flagstaff Extreme Adventure or we have some climbing area such as the Pit and Priest Draw. We also have many mountain biking trail such as Shultz Creek or the Upper and Lower Oldham trails.	71
QA39	Is it easy to use public transportation to go to Sedona, Antelope Canyon, Grand Canyon, etc?	There are shuttles between the locations. Arizona Shuttle goes to Grand Canyon and Sedona and there is a shuttle called National Park Express that goes to Page where you can get a tour of Antelope Canyon or there are guided tours from Flagstaff by Grand Canyon Adventures. Also For Sedona there is Inspire Shuttle	We do have shuttles between these locations. Arizona Shuttle goes to Grand Canyon and Sedona and we have the National Park Express that goes to Page where you can get a tour of Antelope Canyon or we have guided tours from Flagstaff by Grand Canyon Adventures. I heard that Inspire Shuttle also does Sedona	70
QA40	Is there any zoos around here?	There are not any zoos but Bearizona is a drive through north american wildlife park that is 30 miles west of here in Williams and Out of Africa is a African wildlife park about one hour south in Camp Verde.	We don't have any zoos but you should try Bearizona, a drive through wildlife park that's 30 miles west in Williams, and Out of Africa is a wildlife park one hour south in Camp Verde.	69
QA41	How much is Antelope Canyon for adults?	They are typically about $60 for Upper Antelope. Inside the Visitor's Center they have a full listing of the Antelope Canyon tours behind the desk.	Upper Antelope will typically cost $60. Inside the Visitor Center we have a listing of the Antelope Canyon tours you may want to pick up.	68
QA42	What activities can I do with my family in Flagstaff?	The Lowell Observatory has good exhibits for all ages. You can go to exhibits during the day and come back at night for the same price. On weekends they also have music, family movies and children activities at Heritage Square starting at 4:30	The Lowell Observatory has exhibits all ages enjoy. You can visit the exhibits during the day and come back later for the same price. On weekends we also have music, family movies and activities that would entertain your children at Heritage Square starting at 4:30	68
QA43	We're wondering what are some notable buildings along route 66 to visit while going through the city?	There are definitely some good ones! Downtown has managed to preserve more of the original Route 66 buildings than further out in town. If you stop by the Visitor Center, there is a Route 66 walking tour brochure that you can pick up to see and read about them.	We do have some ones I would suggest! Downtown has managed to preserve more of the Route 66 buildings than further out in town. I would suggest you visit the Visitor Center, we have a walking tour brochure that you can pick up to see and read about them.	68

Continuation of Table E.2

ID	Question	FLG	FLG_{mod}	Score
QA44	Hi. I have 1 hour only. What should I do?	If you only have one hour in Flagstaff I would explore the historic downtown area. there are many shops local restaurants and art galleries in the area and this will give you a good feel for the community.	Well having you only one hour in Flagstaff, I'd explore downtown. We have many shops restaurants and art galleries and this will give you a feel for the community.	67
QA45	Do you have any 2-days schedule to visit Flagstaff and the area? I like outdoor and adventure activities	One day could be hiking the San Francisco peaks. Humphreys is the highest point in Arizona and you can hike to the top or for a more mellow trail. The Sunset Trail is nice too. Another day could be seeing our monuments Sunset Crater and Wuptaki are really neat to see.	Okay one day you could hike the San Francisco peaks. Humphreys is the highest point in Arizona and you can hike to the top or choose a trail that is more mellow. I believe the sunset trail is really nice too. Another day you could see our monuments I think Sunset Crater and Wuptaki are really neat to see.	67
QA46	What tourist-type places are open late?	Most close at 5pm but Lowell Observatory is open until 10 pm. Sunset Crater and Wupatki can also be driven through after 5 pm although their visitor centers are closed.	Okay I believe most close at 5pm. Lowell Observatory is open late until 10 pm. You can drive through Sunset Crater and Wupatki after 5 pm although their visitor centers are closed.	67
QA47	What restaurants in Flagstaff are good for kids?	Mama Burger is good for a more classic burger place with fries and milkshakes. Mix in Heritage Square is good for making own salads for the healthier route.	I think Mama Burger is good when you want a burger place that is classic with fries and milkshakes. Mix in Heritage Square is good to make your own salad when you want to take the healthier route.	66
QA48	Hi, where can I park for more than 15 minutes for free?	You can park along Route 66 for 2 hours or park outside of downtown and walk into downtown either west of Sitgreves Street or North of Cherry	You can park along Route 66 for 2 hours or you park either west of Sitgreves Street or North of Cherry and walk into downtown	66
QA49	What is the best, Lower or Upper Canyon with kids (9 and 4)? Can I do both in the same day?	You could but it's not entirely necessary. Lower Canyon is $10 cheaper but there are more stairs involved which might be tough for the 4-year-old. The back of the jeep tour to Upper Antelope might be fun for both of them.	You could but I don't think it's entirely necessary. Lower Canyon is $10 cheaper but it has more stairs involved which the 4-year-old might find tough. I guess the jeep tour to Upper Antelope might be fun for both.	66
QA50	What is nightlife like in Flagstaff?	Hello. Flagstaff is a university town so we have what I would describe as a semi-active nightlife scene. There aren't any dancing clubs but there are a number of breweries and local bars with music trivia nights and so on.	Hello. Flagstaff is a university town so we have a nightlife that I'd describe as semi-active. We don't have any dancing clubs. What we have are the breweries and bars with music trivia nights and so on.	66

Continuation of Table E.2

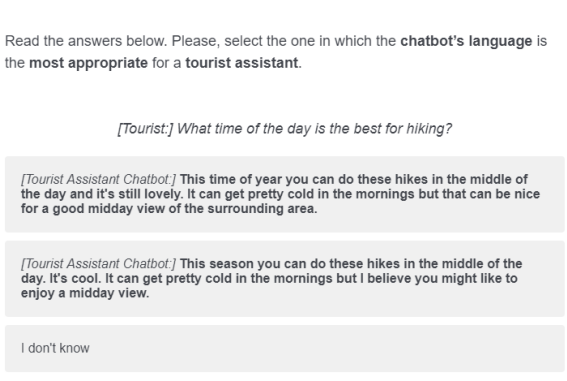
ID	Question	FLG	FLG_{mod}	Score
QA51	It is worth to go to the observatory when it is cloudy?	They do some great historic tours and talks but telescope viewing would not be possible.	There will be some historic tours and talks you could attend but unfortunately you wouldn't do telescope viewing.	63
QA52	Can you walk to the top of the crater and look down into the inside?	They do not have that since the cinders around it are incredibly loose. I would recommend Meteor Crater if you want to look into the inside of a large impact (although that was a huge meteor and not a volcano)	You aren't allowed to do that since the cinders around it are incredibly loose. I'd suggest you visit Meteor Crater instead where you can look the inside of a large impact (although that was a meteor and not a volcano)	63
QA53	Is there any zoos around here?	We don't have any zoos in Northern Arizona. The closest you'll find is Bearizona in Williams 30 minutes west of Flagstaff which is a drive through wildlife park.	We don't have any zoos in Northern Arizona. The closest you'll find is Bearizona, a drive-through wildlife park in Williams, 30 minutes west of Flagstaff.	63
QA54	What are some things that I should know when doing a day trip to the Grand Canyon?	Hello. This time of year the traffic isn't so bad going into the Grand Canyon. I would recommend leaving Flagstaff no later than 8am because that will ensure that you arrive before the Grand Canyon Railway visitors arrive on the train.	Hello. This season the traffic isn't so bad going into the grand canyon. I'd leave Flagstaff no later than 8am to ensure that you arrive early before the Grand Canyon Railway visitors arrive.	63

End of Table

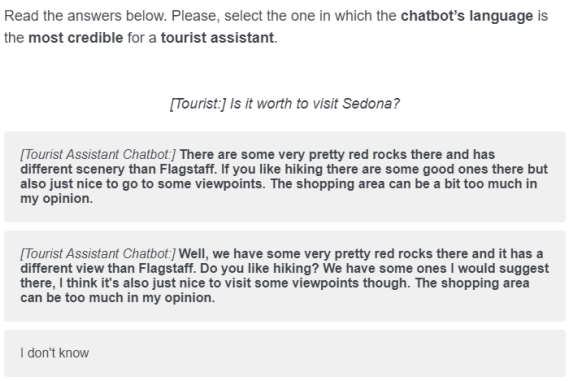
APPENDIX F

USER'S PERCEPTIONS STUDY

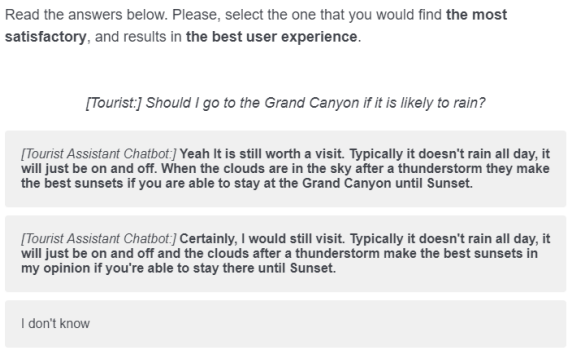
In the user's perceptions study, participants selected one of the three options presented in the screen: the original answer (from FLG), the modified answer (from FLG_{mod}), or "I don't know." Constructs were evaluated one at a time. For each construct, participants read the initial instruction, and then rated a sequence of ten questions for that construct. Figure F.1 depict examples of questions for each construct. The initial instruction to the constructs read as the following:



(a) Question to evaluate appropriateness.



(b) Question to evaluate credibility.



(c) Question to evaluate overall user experience.

Figure F.1: Examples of questions for each construct.

Appropriateness: We invite you to identify the conversation excerpts that present the best fit between the chatbot's language and its expected social role (tourist assistant).

Credibility: We invite you to identify the sentences in which the chatbot sounds more credible. We define credible as a chatbot that sounds the most competent, which makes them the most believable tourist assistant.

User Experience: We invite you to imagine that you are interacting with the Tourist Assistant Chatbot to find information about a destination. Identify the excerpts that suggest you the most satisfactory interaction with the chatbot, which would result in the best overall experience.

Participants answered 10 questions per construct, one of which was an attention check question. A single participant did not evaluate the same question-answer pair more than once, which means that each participant evaluated 27 out of 54 possible question-answer pairs in total (9 per construct). Table F.1 shows the counts of number of votes each option received per question and the total of votes each question received.

Table F.1: Number of votes each answer received per construct (FLG vs. FLG_{mod}).

Begin of Table F.1

	Appropriateness			Credibility			User experience			
QA ID	**#original**	**#mod.**	**#none**	**#original**	**#mod.**	**#none**	**#original**	**#mod.**	**#none**	**Total**
QA1	20	10	0	21	10	0	12	16	1	90
QA2	22	3	1	25	4	0	24	5	0	84
QA3	12	18	0	7	25	1	6	21	0	90
QA4	28	2	0	23	6	0	27	3	0	89
QA5	13	13	2	17	13	0	14	15	0	87
QA6	18	11	0	14	15	0	21	11	0	90
QA7	23	5	0	24	10	1	20	7	0	90
QA8	10	17	1	14	21	0	13	17	0	93
QA9	19	6	0	26	6	0	29	2	0	88
QA10	13	16	0	14	17	0	18	12	2	92
QA11	26	5	0	19	10	0	23	6	1	90
QA12	28	2	0	24	5	0	22	7	0	88
QA13	23	5	1	27	5	1	23	5	0	90
QA14	28	2	0	21	5	0	25	4	2	87
QA15	23	7	1	25	4	0	18	8	1	87
QA16	24	4	0	23	6	2	29	2	0	90
QA17	21	9	0	16	9	0	20	11	0	86
QA18	22	11	0	15	14	0	15	15	0	92
QA19	28	4	1	21	6	4	23	2	1	90
QA20	24	5	0	26	2	1	23	4	0	85
QA21	22	9	0	19	10	2	18	6	0	86
QA22	17	13	1	12	17	3	16	11	0	90
QA23	18	13	0	14	15	0	13	14	0	87
QA24	14	13	2	11	16	0	19	12	0	87
QA25	24	5	0	27	3	1	28	1	1	90
QA26	15	15	0	20	10	0	16	12	1	89
QA27	24	6	0	21	5	1	25	6	2	90
QA28	25	4	0	22	10	0	25	6	0	92
QA29	20	9	0	21	8	1	23	7	1	90
QA30	26	6	0	23	8	0	20	7	0	90
QA31	19	7	1	19	11	0	23	8	0	88
QA32	14	14	0	20	6	1	14	16	0	85
QA33	21	11	0	16	15	1	14	10	1	89
QA34	5	26	0	8	24	1	10	19	0	93
QA35	26	1	1	24	4	1	26	4	1	88
QA36	19	4	1	27	3	0	26	5	0	85
QA37	18	11	3	17	6	1	15	11	2	84
QA38	14	15	0	13	17	1	15	11	1	87
QA39	20	9	0	20	7	0	15	16	0	87
QA40	16	15	0	13	13	1	14	13	0	85
QA41	13	19	0	15	16	0	11	17	0	91
QA42	19	5	1	21	10	0	18	13	0	87
QA43	18	5	2	18	12	0	27	7	0	89
QA44	24	4	2	25	2	1	32	1	0	91

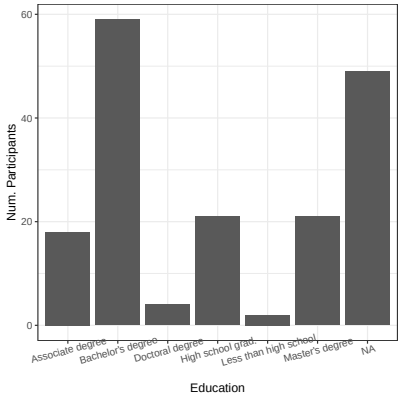
Continuation of Table F.1

QA ID	Appropriateness #original	Appropriateness #mod.	Appropriateness #none	Credibility #original	Credibility #mod.	Credibility #none	User experience #original	User experience #mod.	User experience #none	Total
QA45	21	12	0	22	9	0	16	11	2	93
QA46	28	2	0	26	0	0	31	1	0	88
QA47	16	14	1	18	9	0	21	8	0	87
QA48	18	11	0	17	12	1	19	10	0	88
QA49	29	1	0	26	3	1	25	4	0	89
QA50	27	6	0	22	6	1	21	6	3	92
QA51	18	14	0	18	10	1	16	16	0	93
QA52	10	19	0	8	21	0	5	24	0	87
QA53	6	21	1	4	24	0	9	24	0	89
QA54	14	16	1	15	10	0	28	7	0	91

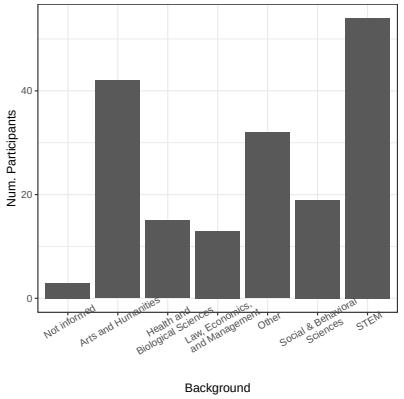
End of Table

F.1 Participants' distributions

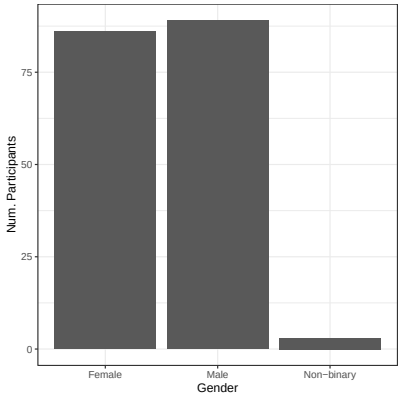
In the following, we present the participants' responses for the demographics (Figure F.2) and profile questions (Figure F.3).



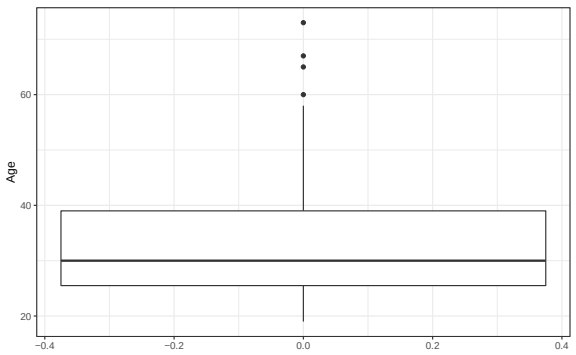
(a) Number of participants per education



(b) Number of participants per background



(c) Number of participants per gender



(d) Age distribution

Figure F.2: Demographics

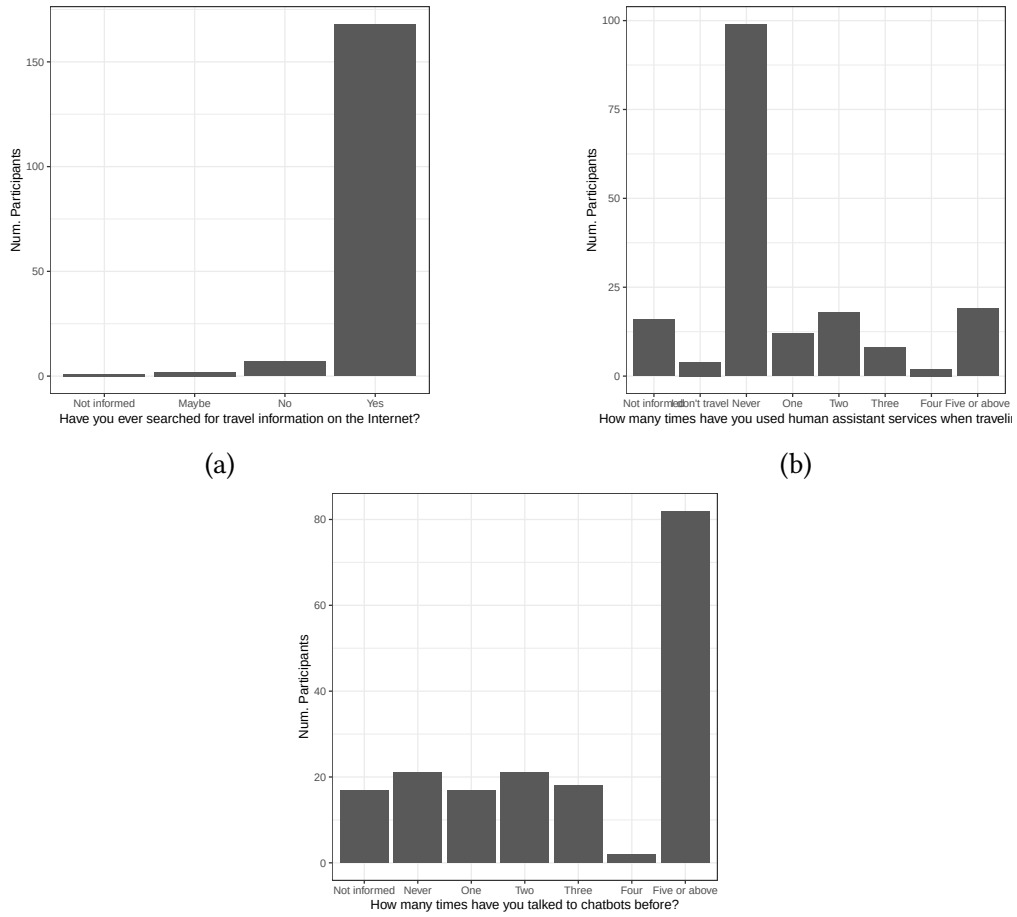


Figure F.3: Tourism information search profile

F.2 Analysis of the linguistic features

The Algorithm 1 is an overview of the cross-validation algorithm used to identify the linguistic features that contributes to predict the users preferences. The actual code is written in R and available on GitHub (Chaves, 2020a).

Figure F.4 depicts the ROC (Receiver Operator Characteristic) curve. The plot shows that our model predictions are consistently above the baseline, represented by the diagonal.

Finally, Figure F.5 shows the coefficients from glmnet model. We considered relevant only the variables that were selected in six or more folds.

ALGORITHM 1: Cross-Validation

attribute a fold number K=1..10 to each observation in the dataset
forall *fold K in folds* **do**
  forall *(construct in c=(appropriateness, credibility, ux)* **do**
    testset = data[fold]
    trainset = data[-fold]
    forall *model in m=(baseline, glmnet, random forest, xgboost)* **do**
      fit the model
      compute prediction
      compute accuracy
      compute ROC/AUC
    end
  end
end

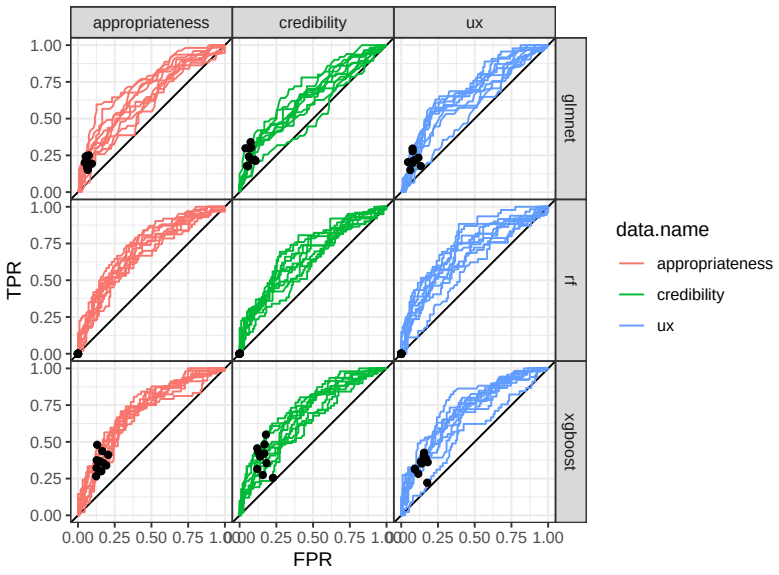
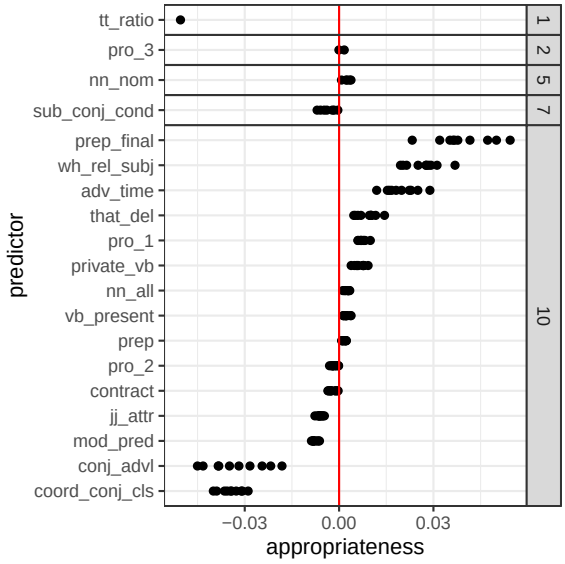
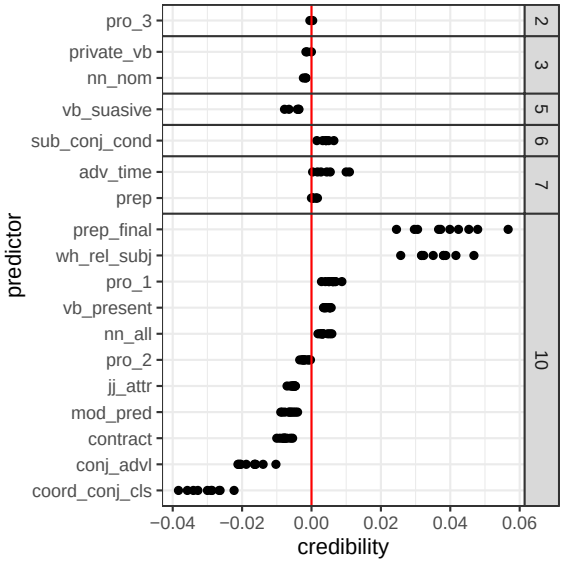


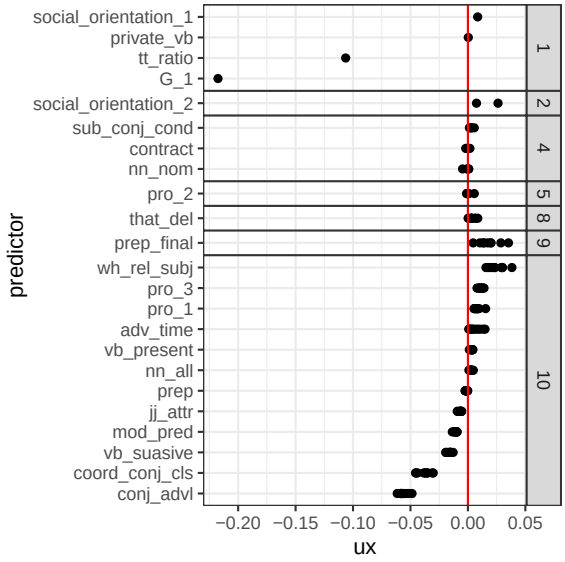
Figure F.4: ROC curve



(a) Coefficients for Appropriateness.



(b) Coefficients for Credibility.



(c) Coefficients for User Experience

Figure F.5: Variables selected for the glmnet model fit. Each dot represents one of the 10 folds. Variables are grouped by the number of folds in which they appear. the x-axis are the coefficients and the solid vertical line marks the zero. Negative coefficients indicates that the variable increases the likelihood of the model predicting the original class, while positive coefficients increases the likelihood of the model predicting the modified class. Variables selected in less than six folds were considered noise and discarded.

APPENDIX G

USER EXPERIENCE'S STUDY

In this study, participants interacted with two versions of a chatbot and evaluated their user experiences in terms of appropriateness, credibility, and overall user experiences as well as social presence and anthropomorphism. In the following section, we present the instruments we used to evaluate these constructs.

G.1 Instruments

This section introduces the instruments used to evaluate the hypothesis in Chapter 7. These questions were presented to the user in the chatbot's user interface.

1. Appropriateness (adapted from Jakic *et al.* (2017)): Considering the current interaction with the Tourist Assistant Chatbot #ID (A/B), please rate the following (1-strongly disagree; 2-disagree; 3-somewhat disagree; 4-neither agree or disagree; 5-somewhat agree; 6-agree; 7-strongly agree):

    - There is a good fit between the tourist assistant's language and its service category.
    - It is logical that this service provider uses such a language style.
    - It is appropriate that this service provider uses such a language style.
    - The chatbot answers with the right amount of formality (Balaji (2019)).

2. Credibility (adapted from Nordheim *et al.* (2019); Corritore *et al.* (2005)):

    - The chatbot appears knowledgeable.
    - The chatbot reflects expertise.
    - The chatbot reflects competency.
    - The chatbot is well equipped for the task it is set to do.
    - The chatbot appears truthful.
    - The information provided by the chatbot is believable.

3. Anthropomorphism (from Nordheim *et al.* (2019)):

    - The chatbot's language is natural.
    - The chatbot's language is human-like.
    - The chatbot's language is realistic.
    - The chatbot's language is present.

- The chatbot's language is authentic.

4. Social Presence (adapted from Katkute *et al.* (2017)): Considering the current interaction with the Tourist Assistant Chatbot #ID (A/B), please enter an answer on a scale of 1 to 10 (1-Not at all; 10-Very much).

    - How much did you feel as if you were interacting with an intelligent being?
    - How much did you feel as if you were accompanied with an intelligent being?
    - How much did you feel as if you were alone?
    - How much attention did you pay to the tourist assistant?
    - How much did you feel involved with the tourist assistant?
    - How much did you feel as if the tourist assistant was responding to you?
    - How much did you feel as if you and the tourist assistant were communicating to each other?

5. Overall user experience:

    - On a scale of 1 to 10, please rate your overall experience with the Tourist Assistant Chatbot #ID (A/B). (1-Extremely bad; 10-Extremely good).

After both interactions:

1. Overall user experience:

    - The tourist assistant that offered me the best experience was (please choose one): Tourist Assistant Chatbot A / Tourist Assistant Chatbot B

2. Chatbots use:

    - Before this study, how many times have you interacted with chatbots? (one, two, three, four, five or above, I have never interacted with chatbots before.)

3. Social agent orientation (from Liao *et al.* (2016)): Please rate the following (1-strongly disagree; 7-strongly agree):

    - I like chatting casually with a chatbot.
    - I think small talk with a chatbot is enjoyable.

APPENDIX H

USER'S PERCEPTIONS: REPLICATION

As in user study 1, participants answered 10 questions per construct, one of which was an attention check question. A single participant did not evaluate the same question-answer pair more than once, which means that each participant evaluated 27 out of 54 possible question-answer pairs in total (9 per construct). Table H.2 shows the counts of number of votes each option received per question and the total of votes each question received.

Table H.2: Number of votes each answer received per construct (FLG vs. FLG_{mod_2}).

Begin of Table H.2

	Appropriateness			Credibility			User experience			
QA ID	**#original**	**#mod.**	**#none**	**#original**	**#mod.**	**#none**	**#original**	**#mod.**	**#none**	**Total**
Q1	19	5	0	17	5	0	16	9	0	71
Q2	21	3	0	20	5	1	20	4	0	74
Q3	14	10	0	17	6	2	13	9	0	71
Q4	24	2	0	17	5	0	23	3	0	74
Q5	21	3	0	21	1	0	20	4	1	71
Q6	18	6	0	19	6	0	20	2	0	71
Q7	19	4	0	22	2	0	16	8	0	71
Q8	19	4	0	23	1	0	17	7	0	71
Q9	15	8	1	12	13	0	10	12	0	71
Q10	18	8	0	15	7	0	18	6	2	74
Q11	15	9	0	19	6	1	19	5	0	74
Q12	21	2	1	20	5	1	19	5	0	74
Q13	18	5	0	16	7	1	14	10	0	71
Q14	7	16	1	10	12	0	12	13	0	71
Q15	15	7	2	16	6	4	16	8	0	74
Q16	20	6	0	18	4	0	19	7	0	74
Q17	13	11	0	16	10	0	12	12	0	74
Q18	13	11	0	17	8	0	11	11	0	71
Q19	16	7	0	13	11	0	16	8	0	71
Q20	8	16	0	9	17	0	11	13	0	74
Q21	17	7	0	18	6	2	17	7	0	74
Q22	16	9	1	16	5	1	17	9	0	74
Q23	19	4	1	16	10	0	18	6	0	74
Q24	19	4	1	20	6	0	21	3	0	74
Q25	12	10	1	14	9	1	16	7	0	70
Q26	15	9	0	12	10	0	15	10	0	71
Q27	19	5	0	21	4	0	18	3	1	71
Q28	15	11	0	8	14	0	15	9	2	74
Q29	21	3	0	23	3	0	20	4	0	74
Q30	16	8	0	20	5	0	17	4	1	71
Q31	14	9	0	16	8	0	17	7	0	71
Q32	21	3	0	18	4	0	19	6	0	71
Q33	19	5	0	22	3	0	20	2	0	71
Q34	26	0	0	20	2	0	23	2	1	74
Q35	19	5	0	18	4	0	25	0	0	71
Q36	11	13	0	14	11	1	16	8	0	74
Q37	18	8	0	15	7	0	19	7	0	74
Q38	12	12	0	16	6	0	19	6	0	71
Q39	11	9	4	11	11	3	8	10	4	71
Q40	15	8	0	14	10	0	15	9	0	71
Q41	20	3	1	21	4	1	19	5	0	74
Q42	19	3	2	21	5	0	17	7	0	74
Q43	17	9	0	11	11	0	11	14	1	74

Continuation of Table H.2

QA ID	Appropriateness			Credibility			User experience			Total
	#original	**#mod.**	**#none**	**#original**	**#mod.**	**#none**	**#original**	**#mod.**	**#none**	
Q44	1	23	0	1	21	0	0	25	0	71
Q45	7	15	2	6	18	2	8	16	0	74
Q46	16	7	0	19	5	0	18	6	0	71
Q47	18	5	1	22	4	0	18	6	0	74
Q48	19	5	0	24	2	0	21	3	0	74
Q49	16	10	0	13	9	0	15	11	0	74
Q50	17	7	0	18	4	0	22	3	0	71
Q51	20	4	0	19	7	0	20	4	0	74
Q52	13	10	0	16	7	1	14	10	0	71
Q53	20	4	0	16	8	2	16	8	0	74
Q54	24	0	0	21	4	0	21	1	0	71

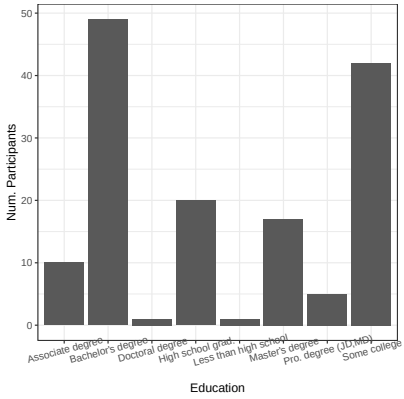
End of Table

H.1 Participants' distributions

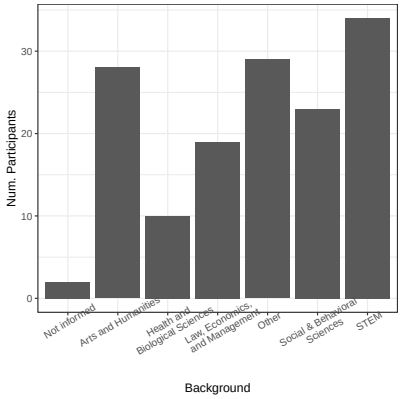
In the following, we present the participants' responses for the demographics (Figure F.2) and profile questions (Figure F.3).

Figure H.3 depicts the ROC (Receiver Operator Characteristic) curve. The plot shows that our model predictions are consistently above the baseline, represented by the diagonal.

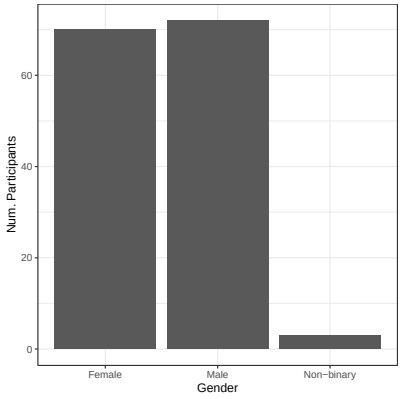
Finally, Figure H.4 shows the coefficients from glmnet model. We considered relevant only the variables that were selected in six or more folds.



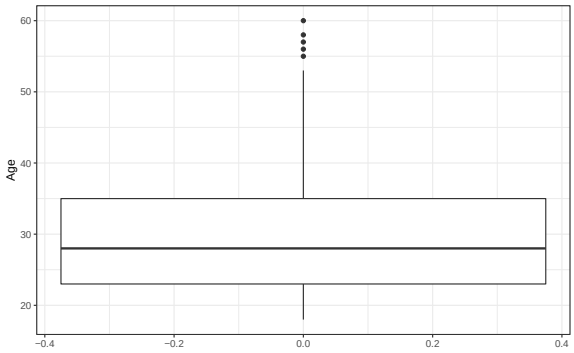
(a) User study 2: Number of participants per education



(b) Number of participants per background



(c) Number of participants per gender



(d) Age distribution

Figure H.1: Demographics

Table H.1: ANOVA results for individual features comparison between $Frames$ and both original and modified corpora, including the non-significant features. The left side of the table presents the estimates and standard deviations for each independent variable ($Frames$, $TA1$, $TA2$, $TA3$), the F-values, and the corresponding p-values. The right side of the table shows the estimates, standard deviations, F-values, and p-values for the three experimental groups ($TA1_{mod_2}$, $TA2_{mod_2}$, $TA3_{mod_2}$) after modifications being performed ($Frames$ column was omitted in the right side to avoid repetition). All the statistics are calculated with $df = 3, 1489$

	Estimates±Std.Dev. (FLG)						Estimates±Std.Dev. (FLG_{mod_2})				
Features	$Frames$	TA1	TA2	TA3	F	P	$TA1_{m_2}$	$TA2_{m_2}$	$TA3_{m_2}$	F	P
Dimension 1: personal involvement											
Private verb	4.37 ± 0.2	6.71 ± 1.3	5.87 ± 1.3	5.71 ± 1.3	1.70	0.16	3.02 ± 1.2	5.16 ± 1.3	4.74 ± 1.2	0.59	0.62
That-deletion	1.03 ± 0.1	0.65 ± 0.6	0.40 ± 0.6	0.56 ± 0.6	0.62	0.60	0.38 ± 0.6	1.71 ± 0.6	0.71 ± 0.6	0.89	0.44
Contraction	6.61 ± 0.4	11.66 ± 2.0	11.88 ± 2.0	1.55 ± 1.9	7.02	0.00	6.89 ± 1.8	6.53 ± 1.9	7.05 ± 1.8	0.03	0.99
Present verb	126.00 ± 1.0	125.93 ± 5.6	119.10 ± 5.6	120.76 ± 5.4	0.77	0.51	119.58 ± 5.3	125.78 ± 5.5	115.63 ± 5.2	1.72	0.16
2nd person pronoun	52.50 ± 0.8	41.38 ± 4.1	23.36 ± 4.2	44.52 ± 4.0	18.60	0.00	38.65 ± 3.9	38.03 ± 4.1	37.71 ± 3.8	12.1	0.00
Do as a pro-verb	2.45 ± 0.3	2.06 ± 1.5	1.90 ± 1.5	2.13 ± 1.5	0.07	0.97	2.10 ± 1.4	1.85 ± 1.5	2.11 ± 1.4	0.09	0.97
Dem. pronoun	4.02 ± 0.2	1.28 ± 1.2	4.39 ± 1.2	2.76 ± 1.1	2.16	0.09	3.48 ± 1.1	5.08 ± 1.2	3.74 ± 1.1	0.37	0.77
Emphatic	1.61 ± 0.2	9.29 ± 0.8	5.29 ± 0.8	5.10 ± 0.8	39.56	0.00	0.95 ± 0.7	1.63 ± 0.7	0.99 ± 0.7	0.58	0.63
1st person pronoun	37.13 ± 0.7	18.18 ± 3.9	16.77 ± 3.9	11.01 ± 3.7	30.68	0.00	38.49 ± 3.7	37.66 ± 3.9	37.72 ± 3.6	0.06	0.98
Pronoun it	8.67 ± 0.3	11.94 ± 1.7	15.05 ± 1.8	9.55 ± 1.7	5.26	0.00	10.13 ± 1.7	11.10 ± 1.8	10.67 ± 1.7	1.15	0.33
Be as a main verb	4.27 ± 0.2	3.59 ± 1.2	2.43 ± 1.2	6.52 ± 1.2	2.12	0.10	2.74 ± 1.1	4.75 ± 1.2	4.67 ± 1.1	0.73	0.53
Causative subord.	0.03 ± 0.0	1.19 ± 0.1	0.00 ± 0.1	0.33 ± 0.1	26.71	0.00	0.00 ± 0.1	0.00 ± 0.1	0.00 ± 0.1	0.11	0.95
Discourse particle	3.30 ± 0.2	3.70 ± 1.1	0.44 ± 1.1	0.00 ± 1.1	5.28	0.00	4.27 ± 1.1	1.87 ± 1.1	1.56 ± 1.0	1.76	0.15
Indefinite pronoun	5.43 ± 0.3	4.79 ± 1.8	1.60 ± 1.8	2.23 ± 1.7	2.60	0.05	5.19 ± 1.7	4.97 ± 1.8	2.78 ± 1.7	0.84	0.47
General hedge	0.07 ± 0.0	0.00 ± 0.1	0.20 ± 0.2	0.00 ± 0.1	0.44	0.72	0.00 ± 0.1	0.00 ± 0.1	0.00 ± 0.1	0.26	0.85
Amplifier	0.92 ± 0.1	3.05 ± 0.6	2.96 ± 0.6	0.24 ± 0.6	8.28	0.00	0.60 ± 0.5	0.29 ± 0.5	0.00 ± 0.5	1.83	0.14
WH- question	6.19 ± 0.3	4.46 ± 1.5	6.25 ± 1.5	0.58 ± 1.4	5.25	0.00	4.43 ± 1.4	5.99 ± 1.5	5.05 ± 1.4	0.67	0.57
Possibility modal	11.45 ± 0.4	15.77 ± 2.3	15.97 ± 2.3	12.08 ± 2.2	2.30	0.08	10.51 ± 2.1	14.47 ± 2.2	14.16 ± 2.1	1.19	0.31
Coord conj (clause)	8.20 ± 0.3	18.97 ± 1.5	8.75 ± 1.6	3.07 ± 1.5	20.16	0.00	8.85 ± 1.5	7.48 ± 1.6	6.89 ± 1.4	0.40	0.75
WH- clause	0.37 ± 0.1	0.91 ± 0.4	0.00 ± 0.4	0.00 ± 0.3	1.50	0.21	1.02 ± 0.3	0.00 ± 0.4	0.00 ± 0.3	2.06	0.10
Final preposition	3.95 ± 0.2	4.87 ± 1.2	0.27 ± 1.3	2.01 ± 1.2	3.82	0.01	5.35 ± 1.2	7.04 ± 1.3	4.46 ± 1.2	2.29	0.08
Nouns	266.43 ± 1.9	277.56 ± 10.2	328.02 ± 10.3	302.11 ± 9.8	15.64	0.00	300.61 ± 9.6	310.56 ± 10.1	307.87 ± 9.5	15.1	0.00
Prepositions	108.22 ± 1.0	100.08 ± 5.4	112.05 ± 5.5	91.30 ± 5.2	4.26	0.01	117.72 ± 5.1	117.29 ± 5.3	107.06 ± 5.0	2.02	0.11
Attrib. adjective	23.01 ± 0.6	43.67 ± 3.0	45.06 ± 3.1	55.09 ± 2.9	65.96	0.00	29.73 ± 2.8	25.82 ± 2.9	24.45 ± 2.7	2.19	0.09
Dimension 2: narrative flow											
Past tense verb	3.87 ± 0.3	9.29 ± 1.3	7.68 ± 1.3	7.01 ± 1.3	9.37	0.00	4.39 ± 1.4	4.60 ± 1.4	7.17 ± 1.3	2.07	0.10
3rd person pronoun	1.07 ± 0.2	12.73 ± 0.9	9.47 ± 0.9	9.84 ± 0.9	105.64	0.00	1.13 ± 0.6	1.83 ± 0.6	1.68 ± 0.6	0.83	0.48
Perfect aspect verb	2.17 ± 0.2	2.10 ± 0.8	1.72 ± 0.8	0.50 ± 0.8	1.54	0.20	2.37 ± 0.8	1.78 ± 0.8	2.66 ± 0.8	0.22	0.88
Public verb	1.09 ± 0.1	0.00 ± 0.6	2.22 ± 0.6	0.95 ± 0.6	2.02	0.11	0.03 ± 0.6	1.58 ± 0.6	0.67 ± 0.6	1.40	0.24
Dimension 3: contextual reference											
WH-rel. cl. (object)	0.16 ± 0.0	0.19 ± 0.2	0.00 ± 0.2	0.00 ± 0.2	0.37	0.78	0.00 ± 0.2	0.00 ± 0.2	0.52 ± 0.2	1.27	0.28
WH-rel. cl. (subject)	0.61 ± 0.1	2.19 ± 0.4	0.98 ± 0.4	0.55 ± 0.4	5.87	0.00	1.03 ± 0.4	1.44 ± 0.4	0.39 ± 0.4	2.15	0.09
WH-rel. pied piping	0.12 ± 0.0	0.00 ± 0.2	0.00 ± 0.2	0.26 ± 0.2	0.39	0.76	0.11 ± 0.2	0.00 ± 0.2	0.25 ± 0.2	0.27	0.85
Coord conj (phrasal)	0.31 ± 0.1	2.36 ± 0.3	0.17 ± 0.3	0.24 ± 0.3	15.14	0.00	0.05 ± 0.2	0.00 ± 0.3	0.00 ± 0.2	1.31	0.27
Nominalization	38.08 ± 0.7	27.41 ± 3.9	35.41 ± 4.0	23.30 ± 3.8	7.09	0.00	35.21 ± 3.7	36.49 ± 3.9	28.48 ± 3.7	2.37	0.07
Time adverbial	2.71 ± 0.2	3.50 ± 1.0	1.57 ± 1.0	2.01 ± 1.0	0.77	0.51	3.83 ± 1.0	2.08 ± 1.0	2.27 ± 1.0	0.64	0.59
Place adverbial	10.57 ± 0.4	15.73 ± 2.1	11.16 ± 2.2	13.34 ± 2.1	2.39	0.07	14.03 ± 2.1	12.19 ± 2.2	12.94 ± 2.0	1.45	0.23
Adverb	24.80 ± 0.7	46.80 ± 3.6	35.23 ± 3.6	23.67 ± 3.5	14.50	0.00	30.85 ± 3.3	27.66 ± 3.5	25.49 ± 3.3	1.24	0.29
Dimension 4: persuasiveness											
Infinitive	10.55 ± 0.3	6.85 ± 1.8	5.18 ± 1.8	8.77 ± 1.7	4.35	0.00	11.59 ± 1.7	7.66 ± 1.8	12.30 ± 1.7	1.32	0.27
Prediction modal	20.98 ± 0.5	10.45 ± 2.5	8.41 ± 2.5	13.97 ± 2.4	15.85	0.00	19.02 ± 2.4	19.04 ± 2.5	16.32 ± 2.3	1.64	0.18
Suasive verb	0.84 ± 0.1	4.51 ± 0.6	1.10 ± 0.6	3.00 ± 0.5	19.17	0.00	0.95 ± 0.5	1.77 ± 0.5	0.78 ± 0.5	1.15	0.33
Conditional subord.	1.46 ± 0.1	4.00 ± 0.7	6.48 ± 0.7	5.94 ± 0.7	30.59	0.00	1.05 ± 0.6	0.61 ± 0.6	1.85 ± 0.6	0.96	0.41
Necessity modal	0.58 ± 0.1	1.33 ± 0.5	1.44 ± 0.5	2.05 ± 0.4	5.54	0.00	1.35 ± 0.4	0.82 ± 0.4	0.63 ± 0.4	1.54	0.20
Split auxiliary	0.79 ± 0.1	5.49 ± 0.5	1.49 ± 0.5	1.40 ± 0.5	27.69	0.00	1.02 ± 0.4	0.70 ± 0.4	1.03 ± 0.4	0.22	0.88
Dimension 5: formality											
Conjuncts	3.46 ± 0.2	2.48 ± 1.1	1.16 ± 1.2	0.74 ± 1.1	3.31	0.02	4.00 ± 1.1	2.54 ± 1.2	1.65 ± 1.1	1.19	0.31
Agentless passive	2.51 ± 0.2	5.91 ± 0.9	4.32 ± 0.9	5.26 ± 0.8	9.54	0.00	2.74 ± 0.7	1.70 ± 0.8	2.14 ± 0.7	0.46	0.71
By-passive	0.08 ± 0.0	0.33 ± 0.2	0.28 ± 0.2	0.00 ± 0.1	1.65	0.18	0.33 ± 0.1	0.29 ± 0.2	0.00 ± 0.1	1.74	0.16
Past participal	1.67 ± 0.1	0.35 ± 0.6	1.03 ± 0.6	0.86 ± 0.6	2.24	0.08	0.95 ± 0.6	1.13 ± 0.6	0.57 ± 0.6	1.77	0.15
Other adv. subord.	1.60 ± 0.1	2.41 ± 0.8	1.91 ± 0.8	2.21 ± 0.7	0.63	0.60	1.79 ± 0.7	1.69 ± 0.7	2.05 ± 0.7	0.17	0.92
Predic. adjective	3.33 ± 0.3	13.33 ± 1.3	6.43 ± 1.3	9.87 ± 1.3	27.37	0.00	4.83 ± 1.2	4.35 ± 1.2	5.18 ± 1.1	1.50	0.21

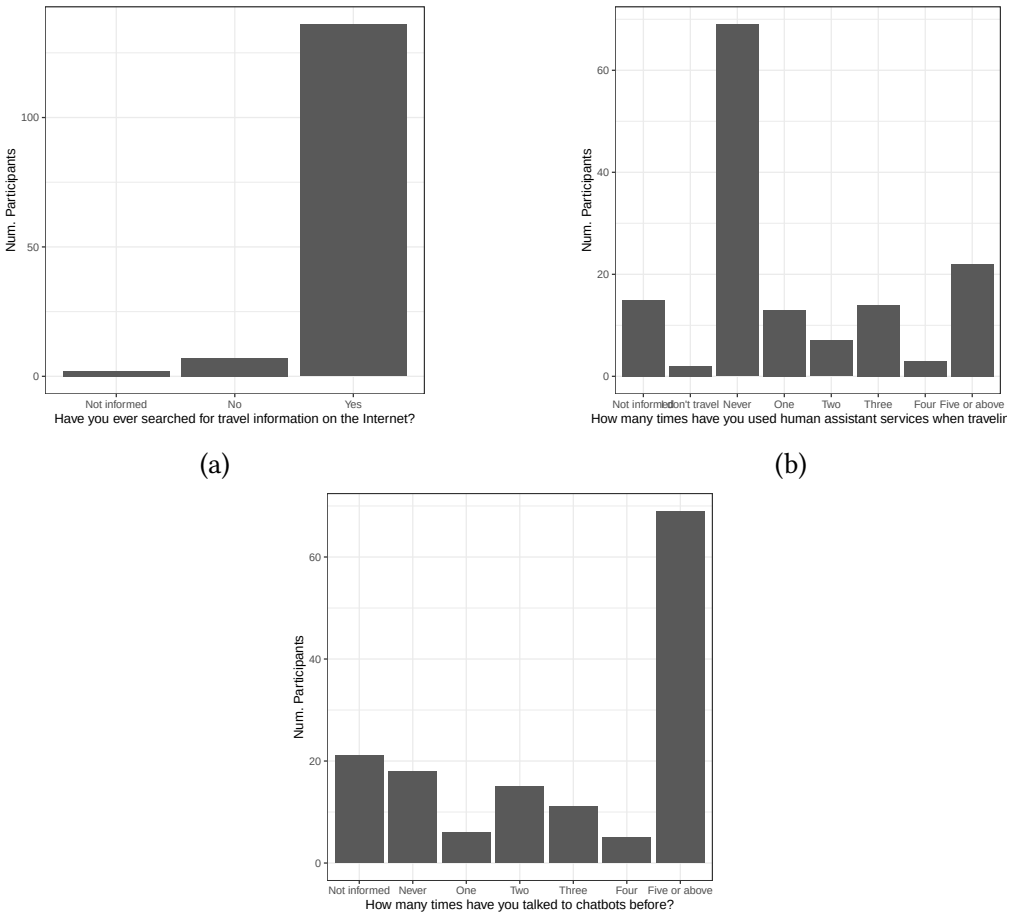


Figure H.2: Tourism information search profile

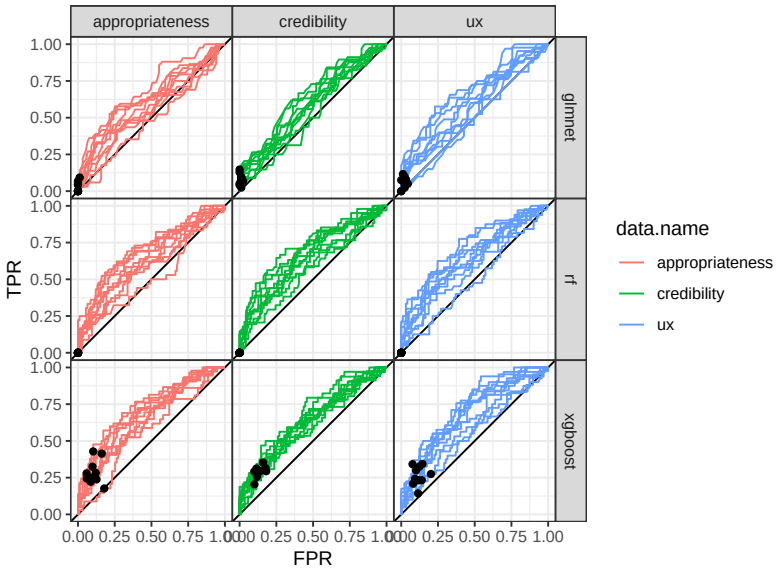
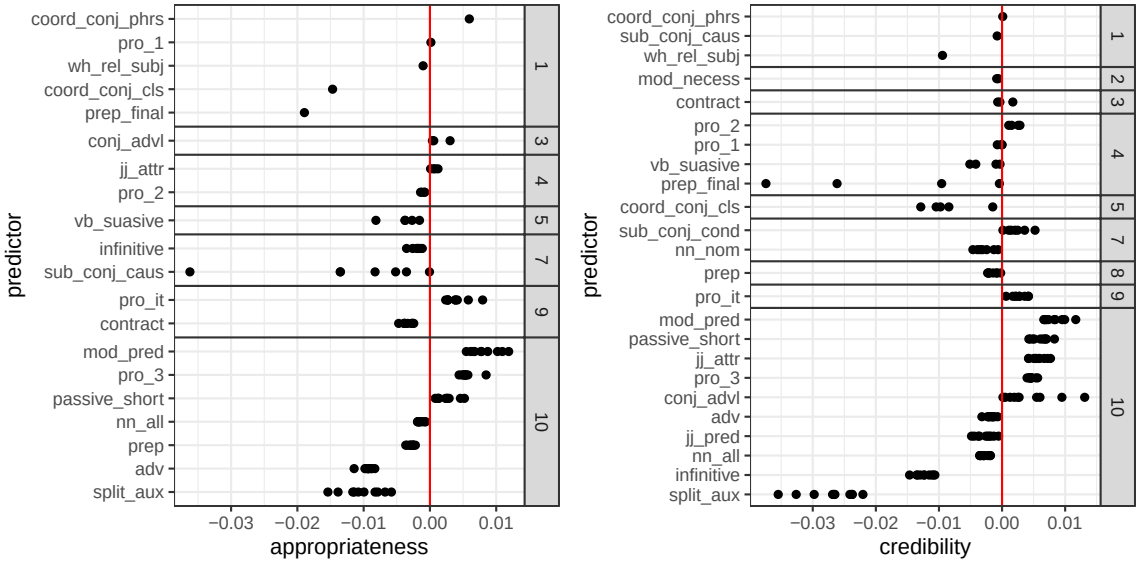
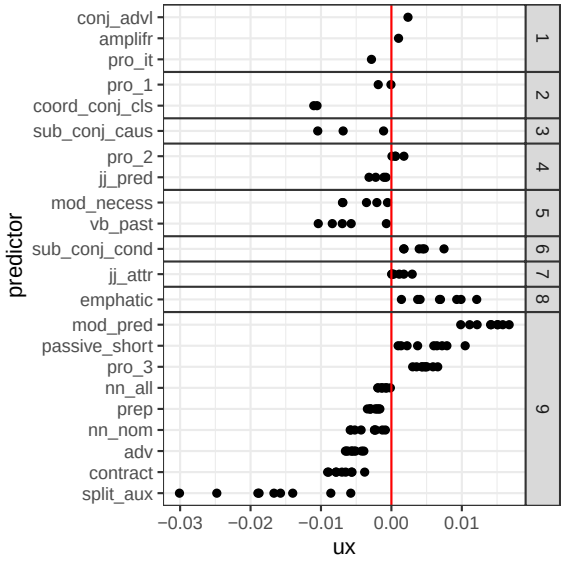


Figure H.3: ROC curve (FLG vs. FLG_{mod_2})



(a) Appropriateness (FLG vs. FLG_{mod_2})

(b) Credibility (FLG vs. FLG_{mod_2})



(c) User Experience (FLG vs. FLG_{mod_2})

Figure H.4: Variables selected for the glmnet model fit. Each dot represents one of the 10 folds. Variables are grouped by the number of folds in which they appear. the x-axis are the coefficients and the solid vertical line marks the zero. Negative coefficients indicates that the variable increases the likelihood of the model predicting the original class, while positive coefficients increases the likelihood of the model predicting the modified class. Variables selected in less than six folds were considered noise and discarded.